# Challenges of Real Life Few-Shot Image Classification

Etienne Bennequin

université
PARIS-SACLAY

# Challenges of Real-Life Few-Shot Image Classification

*Les Défis des Applications Concrètes de la Classification d'Images à partir de Peu d'Exemples*

**Thèse de doctorat de l'université Paris-Saclay**

École doctorale n°573 : interfaces : matériaux, systèmes, usages (INTERFACES)
Spécialité de doctorat : INFORMATIQUE
Graduate School : Sciences de l'ingénierie et des systèmes
Référent : CentraleSupélec

Thèse préparée dans l'unité de recherche MICS (Université Paris-Saclay, CentraleSupélec), sous la direction de Marc Aiguier, Professeur des Universités, Céline Hudelot, Professeure des Universités, et le co-encadrement de Myriam Tami, Maître de conférences.

**Thèse soutenue à Paris-Saclay, le 19 Septembre 2023, par**

**Etienne Bennequin**

**Composition du jury**

| | |
|---|---|
| **Anissa Mokraoui**<br>Professeure des Universités, Université Paris XIII | Rapporteur |
| **Vincent Gripon**<br>Professeur des Universités, IMT Atlantique | Rapporteur |
| **Hervé Le Borgne**<br>Ingénieur-Chercheur, CEA LIST | Examinateur |
| **Florian Yger**<br>Maître de conférences, Université Paris-Dauphine | Examinateur |
| **Stéphane Canu**<br>Professeur des Universités, INSA de Rouen | Président du jury |

THESE DE DOCTORAT
NNT : 2023UPAST122

université PARIS-SACLAY
ÉCOLE DOCTORALE
Interfaces:
matériaux, systèmes, usages

**Titre :** Les Défis des Applications Concrètes de la Classification d'Images à partir de Peu d'Exemples
**Mots clés :** vision automatique, apprentissage machine, apprentissage profond, apprentissage avec peu d'exemples

**Résumé :** En 2015, alors que les réseaux de neurones convolutionnels atteignaient des performances sur-humaines en reconnaissance d'image à grande échelle, la communauté a commencé à observer que ces performances peinaient à se reproduire avec de petits volumes de données. Les algorithmes d'apprentissage profond présentaient de faibles résultats lorsqu'on leur demandait de classifier des images parmi des classes pour lesquelles on ne leur fournissait qu'une poignée d'exemples. À l'inverse, la capacité à reconnaître de nouveaux concepts à partir de très peu d'exemples est considérée comme une capacité naturelle des humains. Ces observations ont donné lieu à l'apparition, dans le paysage de l'apprentissage machine, du *Few-Shot Learning*, ou apprentissage à partir de peu de données (peu de *shots*). Au sein de ce nouveau domaine, nous avons rapidement développé des algorithmes dédiés, construit des jeux de données et établi de nombreuses règles et configurations restrictives pour évaluer les modèles d'apprentissage à partir de peu de données.

Si ce procédé s'est montré très propice à des itérations rapides, et a mené à des découvertes intéressantes, il a également restreint la recherche en apprentissage à partir de peu d'exemples à la résolution d'un problème hypothétique. Nous observons que ce problème, créé artificiellement, est, par de nombreux aspects, non représentatif des problèmes industriels réels que nous avons rencontrés à Sicara. Dans cette thèse, nous mettons en évidence plusieurs divergences entre les hypothèses utilisées en recherche académique et les applications réelles de l'apprentissage à partir de peu d'exemples. Nous proposons des contre-mesures pour réduire cet écart.

Tout d'abord, la configuration standard de la classification d'images à partir de peu d'exemples suppose que les quelques exemples d'images disponibles (l'*ensemble support*) sont issues de la même distribution que les images à classifier (l'*ensemble des requêtes*). En réalité, cette hypothèse est souvent non vérifiée, par exemple lorsque l'ensemble support correspond à des images acquises dans un environnement contrôlé (*e.g.*, le catalogue d'un site d'e-commerce) tandis que les images requêtes sont plus chaotiques (*e.g.*, des photographies prises par des utilisateurs individuels). Nous formalisons ce problème sous la dénomination de *Few-Shot Learning under Support-Query Shift*, ou *apprentissage à partir de peu d'exemples avec changement de distribution entre le support et les requêtes*. Nous proposons des jeux de données et des procédés d'évaluation dédiés, ainsi qu'une première méthode pour faciliter les efforts de recherche consacrés à ce problème.

Par ailleurs, dans de nombreuses applications de l'apprentissage à partir de peu de données, nous ne pouvons pas assurer que les images requêtes appartiennent effectivement aux classes définies dans l'ensemble support. Ce problème, connu dans la littérature sous le nom de *Few-Shot Open-Set Recognition*, ou *reconnaissance à partir de peu d'exemples dans un ensemble ouvert*, était déjà abordé dans des précédents travaux. Cependant, les méthodes complexes développées pour ce problème ne montraient pas d'incrément notable par rapport à des méthodes naïves. Dans cette thèse, nous mettons à profit l'ensemble des requêtes par une approche simple et raisonnée pour atteindre des performances utilisables en reconnaissance à partir de peu d'exemples dans un ensemble ouvert.

Enfin, nous avons observé que les bancs de test les plus populaires dans la recherche académique présentaient un biais important. Ainsi, les modèles étaient évalués sur des tâches de classification à partir de peu d'exemples non représentatives d'applications réelles. En effet, dans ces bancs de test, nous avions tendances à demander aux modèles de classifier parmi des classes correspondant à des concepts très distants *e.g.*, distinguer une tarte d'un serpent. À l'inverse, la plupart des applications du monde réel impliquent une distinction entre des concepts très proches *e.g.*, des bactéries d'autres bactéries, des outils d'autres outils, ou des composants électroniques d'autres composants électroniques. Dans cette thèse, nous proposons une nouvelle méthode d'évaluation pour résoudre ces biais.

**Title :** Challenges of Real-Life Few-Shot Image Classification
**Keywords :** computer vision, machine learning, deep learning, few-shot learning

**Abstract :** In 2015, while deep neural networks achieved superhuman performance in large-scale image recognition, we started observing that this performance could not be reproduced with small volumes of data. Deep learning algorithms showed weak results when asked to classify images among classes for which there were given only a handful of examples. In contrast, the ability to recognize new concepts from very few examples was deemed to be a signature ability of human beings. As a result, a new field emerged in the Machine Learning landscape : *Few-Shot Learning i.e.*, the ability to learn from a few examples, or *shots*. In this new field, we rapidly developed specific algorithms, designed specific benchmarks, and drew many rules and restrictive settings to evaluate Few-Shot Learning methods.

While this abstraction process was very useful for easy comparison and rapid iterations and led to many interesting findings, it also restricted Few-Shot Learning research to the resolution of a toy problem. We find that this artificially created problem is, in many ways, not representative of the real industrial use cases that we encountered at Sicara. In this thesis, we highlight several divergences between academic research and the real use cases for Few-Shot Learning and propose counter-measures to bridge this gap.

Firstly, the standard Few-Shot Image Classification setting uses the assumption that the few available example images (the *support set*) are drawn from the same distribution as the images we intend to classify (the *query set*). In reality, this assumption often breaks, when the support set corresponds to images acquired in a controlled environment (*e.g.*, the catalog of an online marketplace) while query images are taken more chaotically (*e.g.*, photos uploaded by individual users). We formalize this problem as *Few-Shot Learning under Support-Query Shift* and provide specific benchmarks, evaluation processes, and a baseline to quickstart the efforts towards its solving.

Secondly, in many applications of Few-Shot Learning, we cannot enforce that query images do indeed belong to the classes defined in the support set. This problem, known in the literature as Few-Shot Open-Set Recognition, was already addressed by a handful of previous works. However, the convoluted methods that were designed for this specific problem fail to improve on naive baselines. In this thesis, we leverage the query set through a simple and principled solution to achieve usable performance in Few-Shot Open-Set Recognition.

Finally, we observed that the most popular academic benchmarks presented an important bias, resulting in models being evaluated on few-shot classification tasks that are not representative of real-life applications. Indeed, with those benchmarks, we tend to ask the model to classify between classes that correspond to very distant concepts *e.g.*, distinguishing a pie from a snake. In contrast, most applications involve a distinction between very similar concepts *e.g.*, bacteria from bacteria, tools from tools, food from food, or electronic parts from electronic parts. In this thesis, we propose a new benchmarking method to combat this bias in our evaluation process.

Thèse de doctorat de l'université Paris-Saclay

# Challenges of Real-Life Few-Shot Image Classification

*by*

Etienne Bennequin

Director: Prof. Dr. Céline Hudelot
Director: Prof. Dr. Marc Aiguier
Co-supervisor: Asst. Prof. Dr. Myriam Tami

École doctorale n°573 : interfaces : matériaux, systèmes, usages (INTERFACES)
Spécialité de doctorat : INFORMATIQUE
Graduate School : Sciences de l'ingénierie et des systèmes
Référent : CentraleSupélec

Thèse préparée dans l'unité de recherche MICS (Université Paris-Saclay, CentraleSupélec), sous la direction de Marc Aiguier, Professeur des Universités, Céline HUDELOT, Professeure des Universités, et le co-encadrement de Myriam TAMI, Maître de conférences.

Université Paris-Saclay, CentraleSupélec, MICS, 91190, Gif-sur-Yvette, France.

# Résumé

En 2015, alors que les réseaux de neurones convolutionnels atteignaient des performances surhumaines en reconnaissance d'image à grande échelle, la communauté a commencé à observer que ces performances peinaient à se reproduire avec de petits volumes de données. Les algorithmes d'apprentissage profond présentaient de faibles résultats lorsqu'on leur demandait de classifier des images parmi des classes pour lesquelles on ne leur fournissait qu'une poignée d'exemples. À l'inverse, la capacité à reconnaître de nouveaux concepts à partir de très peu d'exemples est considérée comme une capacité naturelle des humains. Ces observations ont donné lieu à l'apparition, dans le paysage de l'apprentissage machine, du *Few-Shot Learning*, ou apprentissage à partir de peu de données (peu de *shots*). Au sein de ce nouveau domaine, nous avons rapidement développé des algorithmes dédiés, construit des jeux de données et établi de nombreuses règles et configurations restrictives pour évaluer les modèles d'apprentissage à partir de peu de données.

Si ce procédé s'est montré très propice à des itérations rapides, et a mené à des découvertes intéressantes, il a également restreint la recherche en apprentissage à partir de peu d'exemples à la résolution d'un problème hypothétique. Nous observons que ce problème, créé artificiellement, est, par de nombreux aspects, non représentatif des problèmes industriels réels que nous avons rencontrés à Sicara. Dans cette thèse, nous mettons en évidence plusieurs divergences entre les hypothèses utilisées en recherche académique et les applications réelles de l'apprentissage à partir de peu d'exemples. Nous proposons des contre-mesures pour réduire cet écart.

Tout d'abord, la configuration standard de la classification d'images à partir de peu d'exemples suppose que les quelques exemples d'images disponibles (l'*ensemble support*) sont issues de la même distribution que les images à classifier (l'*ensemble des requêtes*). En réalité, cette hypothèse est souvent non vérifiée, par exemple lorsque l'ensemble support correspond à des images acquises dans un environnement contrôlé (*e.g.*, le catalogue d'un site d'e-commerce) tandis que les images requêtes sont plus chaotiques (*e.g.*, des photographies prises par des utilisateurs individuels). Nous formalisons ce problème sous la dénomination de *Few-Shot Learning under Support-Query Shift*, ou *apprentissage à partir de peu d'exemples avec changement de distribution entre le support et les requêtes*. Nous proposons des jeux de données et des procédés d'évaluation dédiés, ainsi qu'une première méthode pour faciliter les efforts de recherche consacrés à ce problème.

Par ailleurs, dans de nombreuses applications de l'apprentissage à partir de peu de données, nous ne pouvons pas assurer que les images requêtes appartiennent effectivement aux classes définies dans l'ensemble support. Ce problème, connu dans la littérature sous le nom de *Few-Shot Open-Set Recognition*, ou *reconnaissance à partir de peu d'exemples dans un ensemble ouvert*, était déjà abordé dans des précédents travaux. Cependant, les méthodes complexes développées pour ce problème ne montraient pas d'incrément notable par rapport à des méthodes naïves. Dans cette thèse, nous mettons à profit l'ensemble des requêtes par une approche simple et raisonnée pour atteindre des performances utilisables en reconnaissance à partir de peu d'exemples dans un ensemble ouvert.

Enfin, nous avons observé que les bancs de test les plus populaires dans la recherche académique présentaient un biais important. Ainsi, les modèles étaient évalués sur des tâches de classification à partir de peu d'exemples non représentatives d'applications réelles. En effet, dans ces bancs de test, nous avions tendances à demander aux modèles de classifier parmi des classes correspondant

à des concepts très distants *e.g.*, distinguer une tarte d'un serpent. À l'inverse, la plupart des applications du monde réel impliquent une distinction entre des concepts très proches *e.g.*, des bactéries d'autres bactéries, des outils d'autres outils, ou des composants électroniques d'autres composants électroniques. Dans cette thèse, nous proposons une nouvelle méthode d'évaluation pour résoudre ces biais.

# Abstract

In 2015, while deep neural networks achieved super-human performance in large-scale image recognition, we started observing that this performance could not be reproduced with small volumes of data. Deep learning algorithms showed weak results when asked to classify images among classes for which there were given only a handful of examples. In contrast, the ability to recognize new concepts from very few examples was deemed to be a signature ability of human beings. As a result, a new field emerged in the Machine Learning landscape: *Few-Shot Learning i.e.*, the ability to learn from a few examples, or *shots*. In this new field, we rapidly developed specific algorithms, designed specific benchmarks, and drew many rules and restrictive settings to evaluate Few-Shot Learning methods.

While this abstraction process was very useful for easy comparison and rapid iterations and led to many interesting findings, it also restricted Few-Shot Learning research to the resolution of a toy problem. We find that this artificially created problem is, in many ways, not representative of the real industrial use cases that we encountered at Sicara. In this thesis, we highlight several divergences between academic research and the real use cases for Few-Shot Learning and propose counter-measures to bridge this gap.

Firstly, the standard Few-Shot Image Classification setting uses the assumption that the few available example images (the *support set*) are drawn from the same distribution as the images we intend to classify (the *query set*). In reality, this assumption often breaks, when the support set corresponds to images acquired in a controlled environment (*e.g.*, the catalog of an online marketplace) while query images are taken more chaotically (*e.g.*, photos uploaded by individual users). We formalize this problem as *Few-Shot Learning under Support-Query Shift* and provide specific benchmarks, evaluation processes, and a baseline to quickstart the efforts towards its solving.

Secondly, in many applications of Few-Shot Learning, we cannot enforce that query images do indeed belong to the classes defined in the support set. This problem, known in the literature as Few-Shot Open-Set Recognition, was already addressed by a handful of previous works. However, the convoluted methods that were designed for this specific problem fail to improve on naive baselines. In this thesis, we leverage the query set through a simple and principled solution to achieve usable performance in Few-Shot Open-Set Recognition.

Finally, we observed that the most popular academic benchmarks presented an important bias, resulting in models being evaluated on few-shot classification tasks that are not representative of real-life applications. Indeed, with those benchmarks, we tend to ask the model to classify between classes that correspond to very distant concepts *e.g.*, distinguishing a pie from a snake. In contrast, most applications involve a distinction between very similar concepts *e.g.*, bacteria from bacteria, tools from tools, food from food, or electronic parts from electronic parts. In this thesis, we propose a new benchmarking method to combat this bias in our evaluation process.

# Acknowledgements

This CIFRE thesis is a collaborative work between the MICS laboratory at CentraleSupélec (Université Paris-Saclay) and the Sicara company. Sicara is a data science and engineering service company that develops tailor-made solutions to value its customers' data. A few years ago, the founders Benoît Limare and Pierre-Henri Cumenge took a leap of faith by engaging their three years old company in a three years research project. They trusted me with this ambitious project and allowed me to pursue my doctorate in the best imaginable work environment for me. I will be forever grateful for this.

I also would like to thank Professor Céline Hudelot, Associate Professor Myriam Tami, and Antoine Toubhans for their incredible support and insightful guidance during this Ph.D., as well as everyone at Sicara and the MICS laboratory for participating in creating these great environments.

Finally, to my parents, Prof. Dr. Laurence Halpern and Prof. Dr. Daniel Bennequin, thank you for never pressuring me into academic research and patiently nodding as I repeatedly stated that I would never pursue a Ph.D. Thank you for this soft and caring social reproduction.

# Contents

# 1 Introduction: The Gap between Few-Shot Learning Research and its Real Applications

## 1.1 The Few-Shot Learning Problem

The Deep Learning Revolution in Computer Vision Computer Vision is an immense area of research and applications. It regroups many tasks: classifying images, localizing and/or classifying objects in an image, segmenting objects in an image, grouping similar images, editing images, or generating them from scratch... All these tasks present a similar challenge: we need to recognize and harness visual patterns in images. Deep learning algorithms showed a lot of promise, especially the convolutional neural networks (CNN) introduced by Fukushima 1980. Indeed, their architecture based on convolutional kernels allows them to capture both local and global, low-level and high-level patterns.

The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) from Russakovsky et al. 2015, since it was launched in 2010, is the most widely used benchmark for image recognition. It consists in classifying images among the thousand classes of the popular ImageNet dataset. In 2012, the first convolutional neural network to achieve state-of-the-art on ImageNet was proposed in Krizhevsky, Sutskever, et al. 2012. The error rate of their model AlexNet was still three times higher than that of a human agent. Year after year, new convolutional networks achieved new state-of-the-art on ImageNet, and in 2015, just five years after the first edition of the challenge, He et al. 2016 obtained super-human results on ImageNet. Their proposed architecture, based on residual blocks, contained 152 layers (seven times more than its predecessor). By scaling up convolutional networks, He et al. cut the error rate in half. After this, we could think that the image recognition problem was solved. All we had to do was to tune the millions of parameters of a deep neural network using a million images.

The need for Few-Shot Learning. However, this outstanding performance of deep learning in the very specific of the ILSVRC did not generalize well when confronted with other image recognition problems. Indeed, the models used by He et al. rely on millions of parameters, which need to be optimized on a large number of examples. To get an idea of the scale, ImageNet's training set contains 1.4 million images. However, the very same year as He et al.'s ground-breaking work, B. M. Lake et al. 2015 showed that deep learning methods still performed way worse than humans when only one example is available for each concept.

What if we do not have a million images to train a deep neural network? What if we want our model to recognize concepts for which we only have five, or even one, example? The deep learning models that achieved such outstanding results in large-scale recognition would be rendered useless,

as it would be impossible to train them with such few images. We would thus need new ideas, new algorithms, and new benchmarks.

This is the displayed goal of a recent sub-field of Machine Learning: Few-Shot Learning *i.e.*, making models able to learn from only a few examples. In this context, the number of *shots* is the number of available examples for each concept or *class*. Specifically, in Few-Shot Image Classification, we want to use those few shots to learn as much as we can about the targeted concepts.

The value of Few-Shot Learning. This problem rapidly drew the attention of an increasing research community, and many novel and imaginative methods were proposed to address the Few-Shot Learning[1] problem. It is, perhaps, too early in this thesis to deep dive into these methods (we kindly ask the reader to wait patiently for Chapter 2). We will, however, state the value that the majority of these methods provide.

1. The first point comes very naturally: these methods are able to adapt to new concepts (or *classes*) from just a few examples. Therefore, they alleviate the need for extensive data acquisition for each and every new object that we need to recognize. Few-Shot Learning methods are designed to learn a *representation i.e.*, a on a *base* dataset for which there are plenty of available data, and *then* adapt to new classes for which the data is scarce. This representation is typically a projection of the image into a lower-dimensional space that captures desirable features to perform computer vision tasks.

2. The second point is not inherent to the few-data regime but rather can be seen as a valuable side effect of most strategies that were designed to address it: class-agnostic deployment. Indeed, most Few-Shot learning methods avoid any *re-training* on the few examples provided for the new classes. Instead, they provide a way to leverage these new images without any update of the parameters of the deep-learning models. As it is, the whole *support set* (the set of images provided to recognize the target classes) can be seen as an input of the model at inference time. Therefore, any change to this *support set e.g.*, the addition of an extra class, removal of an existing one, or modification of the example images, can be done seamlessly. It will not require any re-training or recalibration of the model or any other intervention of a Machine Learning expert. The model deployed in production is *class-agnostic*.

Once we have stated the value of Few-Shot Learning methods, the question that naturally comes next is: how do we measure it?

The thoroughly standardized Few-Shot setting. At this point, the Few-Shot Learning problem seems very broad: how many are "few"? how many classes are there? do we need to start from scratch or are we allowed to use previous knowledge about other concepts? are all classes "few-shot" or is it a mix-up between few-shot and large-scale classes?

All these questions were answered rather quickly since we needed standard benchmarks to compare between methods. In fact, the vast majority of contributions in Few-Shot Image Classification follow the setting standardized by Vinyals et al. 2016 and represented in Figure 1.1 *i.e.*, :

[1] Few-Shot Image Classification in most cases.

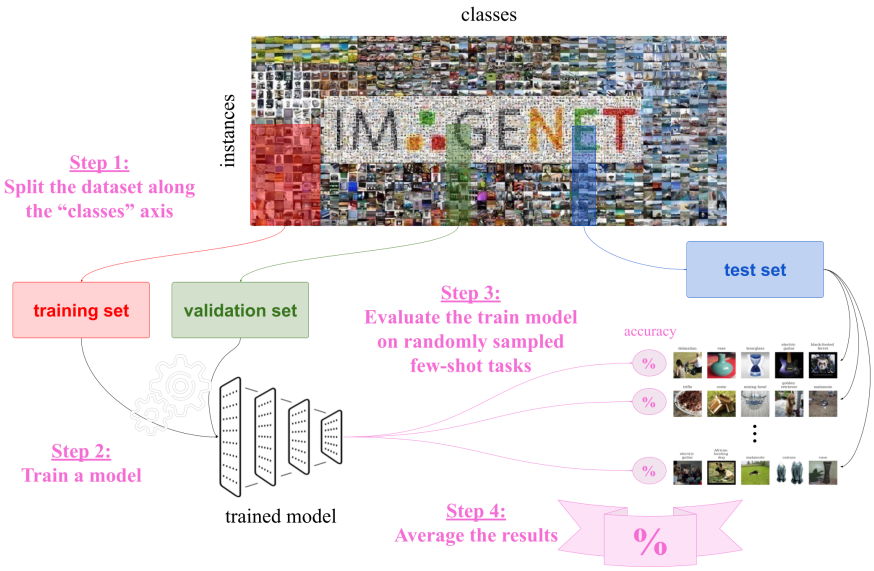



Figure 1.1: The standard Few-Shot Image Classification setting.

1. we assume a large-scale vision dataset (*e.g.*, ImageNet) and we construct a training, a validation, and a test set by splitting the datasets classes: we define specific training, validation, and test classes;

2. the model is allowed to train on all images from the training set and validate on all images from the validation set: this is a large-scale pre-training;

3. to evaluate the Few-Shot Learning ability of the model, we sample a large number of artificial few-shot tasks from the test set:

    - they are 5-way tasks, which means that each task involves 5 classes, sampled uniformly at random from the set of test classes;
    - for each class, we sample a predefined number $n$ of examples images (or *shots*) uniformly at random from the set of test images belonging to this class; these examples constitute the *support set*; in the literature, we almost always consider the cases $n = 1$ and $n = 5$;
    - we then sample query images uniformly at random from the remaining images of these classes, typically 15 for each class; these instances form the *query set*;
    - the model is asked to classify each query among the 5 classes using the information given by the $5n$ images in the support set;

4. we report the average accuracy over all test tasks.

While source datasets may vary, these four steps remain roughly the same.

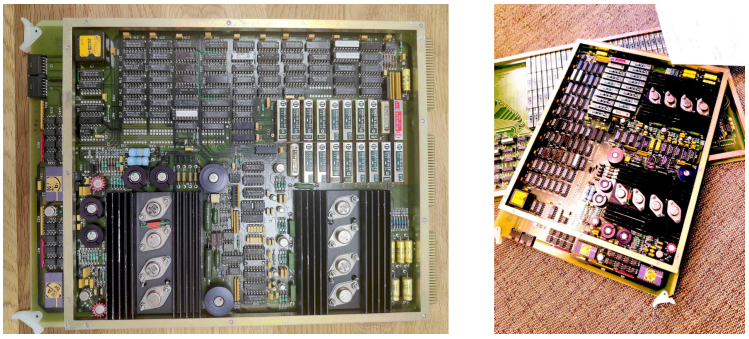


Figure 1.2: Extreme example of domain shift between a catalog image (left) and a query image (right). We notice differences in background, context, and angles, as well as different expositions and colorimetry coming from different acquisition devices.

## 1.2 Constraints in Real-Life Applications

The context of this thesis is a CIFRE collaboration between the MICS laboratory at CentraleSupélec and the French company Sicara. As a service company helping businesses develop their own custom-made computer vision solutions, Sicara has faced numerous problems involving Few-Shot Learning. Unsurprisingly, we found that real-life use cases of Few-Shot Learning do not always fit the very specific setting considered in the Few-Shot Learning literature. Let us review some of those use cases, which motivated the work presented in this thesis.

### 1.2.1 Use Case 1: Enabling Maintenance at a Factory

**The hassle of identifying specific parts of a machine.** When an industry is dealing with many different machines, each composed of many different parts and sparsely located in numerous factories, maintenance can become a hassle. Once the defective part is identified, the operator on site is likely not to know exactly what this part is, and will therefore not be able to repair it or order a new one.

**Recognizing the part in a wide database.** To solve this problem, our Sicara team chose to develop a mobile app allowing the operator to take a picture of the defective part and receive its identification from the *catalog* of existing parts. From this, they can download the reparation instructions, or order a replacement. This use case presents two main challenges:

1. Most items in the database only have one example image;

2. As the industry is evolving, the set of existing parts is expected to change. The manager of the database needs a way to add, remove, or change parts in the catalog, on the fly.

The reader will surely notice that these challenges perfectly correspond to the valuable characteristics of Few-Shot Learning methods that we stated in Section 1.1. Therefore, we naturally resolved to Few-Shot Learning methods.


Figure 1.3: Illustration of the challenges of product retrieval for an online hardware store. *(i)* The query image is a picture of a watering can taken in its natural environment, while all catalog images are taken on a clean white background. *(ii)* The class "watering can" does not appear in the catalog.

The reality of the deployed model. We found, however, that our setting presented many differences from the standard academic setting presented in Section 1.1. The first difference with the standard academic Few-Shot Learning problem is that in this case, we are facing a domain shift between the catalog of parts (most likely composed of pictures taken just after leaving the factory) and the query images, which are pictures of parts after years of use, taken with a mobile phone in an uncontrolled environment, as shown in Figure 1.2. The second problem that we identified is that we could not rely at all on academic benchmarks to identify the best algorithm for our needs, because:

- Academic benchmarks are limited to 5-way classification tasks, but our problem involved thousands of classes;
- All academic benchmarks report only top-1 accuracy *i.e.*, the proportion of instances for which the model predicted the ground truth class; yet we allowed the operator to retrieve their target between several propositions, therefore our main metric was top-5 accuracy *i.e.*, the proportion of instances for which the ground truth class belongs to the 5 most likely classes according to the model's prediction.

### 1.2.2 Use Case 2: Retrieving relevant Items in a Marketplace

A picture makes a sale. An important issue for many online marketplaces is to make the user journey to the product that they want to buy[2] as easy as possible. To that end, the idea that the customer would need to navigate through wide and deep menus to find a product is unbearable. Instead, our client, an online hardware store, intended their users to be able to simply take a picture of their tools at home and be recommended similar tools.

[2]Or that you want them to buy.

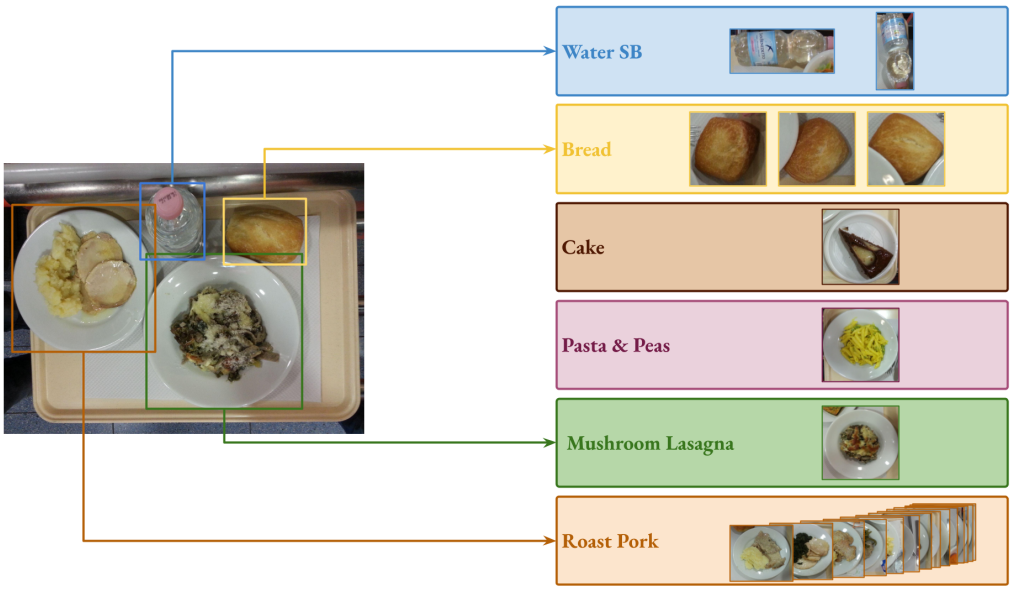



Figure 1.4: Illustration of the problem of meal tray recognition. Items on the tray must be classified among the items on the current menu. Some classes come with many examples while others come with very few examples.

Few-Shot Tool Classification. There are many similarities between this use case and the use case from Section 1.2.1 that made Few-Shot Learning methods an obvious candidate for a solution. Indeed, the marketplace's catalog also contains just a few examples for each product and is subject to evolution in time. Just as before, we do not want to re-train our model each time the marketplace releases a new land-mower.

A non-exhaustive catalog. Then again, this Few-Shot setting presents some differences from the standardized academic setting, which we show in Figure 1.3. Naturally, there is a domain shift: the images in the catalog are most often fashion shots on a clean white background, which is very different from the photos taken by the end user. Just as before, the catalogs involve thousands of classes, and top-1 accuracy is not the most relevant metric. Additionally, this problem is by nature open-set *i.e.*, query images are not guaranteed to belong to any of the known classes. When it is the case, we do not want to falsely label them as the closest entry in the catalog. Instead, we want to recognize that the item does not appear in the catalog and flag it as such.

### 1.2.3 Use Case 3: Daily Food Recognition

Recognizing new plates without retraining. A rich applicative area of Machine Learning is to automate repetitive tasks previously handled by human workers[3]. Our third use case lies in this area. The role of the cashier in cafeterias (*i.e.*, listing the items on a meal tray and editing the corresponding bill) has been seen as a promising terrain for automation. As it happens, it is

[3]The societal impact of such applications will not be discussed here.

also a very interesting playground for Few-Shot Learning. Indeed, each new service comes with its particular set of items on the menu, some of which may not have appeared on any previous menu. For instance, when a cafeteria chef decides to add *mushroom lasagnas* to the menu (even though it has never appeared on the menu before), we cannot expect to acquire sufficient data on mushroom lasagna before the day's service to fine-tune a standard classification model on this new class. Instead, we need a system that can take only one example plate prepared by the chef at the beginning of the service, and then recognize new instances of mushroom lasagna accurately. This is where Few-Shot Learning methods can add substantial value.

Popular and unknown dishes. After a few years of processing millions of meal trays, the resulting dataset would not intuitively be considered a "Few-Shot Learning" dataset. Indeed, some classes such as *roast pork* have thousands of aggregated examples, while many classes (such as the new dishes appearing on a daily basis) have only one example. This problem is illustrated in Figure 1.4. How does this mix between large-scale and few-shot classes fit into the standard academic Few-Shot Learning setting?

### 1.2.4 Use Case 4: Classification from Microscopic Images

Bridging geographical inequalities in micro-organism recognition. A Canadian start-up in biotechnologies came to Sicara with a fascinating and intensely "tech-for-good" use case. From their experience as biology teachers and practitioners, many hospitals (especially in remote regions or emerging countries) could not always guarantee students and practitioners access to some specific expertise. In that context, they could well benefit from intelligent tools to perform certain high-value tasks, such as recognizing micro-organisms (bacteria, viruses, or fungi) from microscopic images.

Few-Shot Classification of Microscopic Images. The use case was then to design a mobile application that would, from a picture taken through a microscope, automatically recognize the micro-organism, and provide an experimental process to validate the prediction using internal documentation. Our role was to design the image classification algorithm. We found that several characteristics of this problem made it an ideal use case for Few-Shot Learning:

- A specific scientific expertise is needed for both data collection and annotation. Indeed, examples of micro-organisms' images are obtained through a delicate experimental process. For instance, depending on the reactants applied to the micro-organism and the order in which they are used, it will not have the same aspects. These reactants, as well as other information such as the type of microscope or the magnification rate, must be thoroughly noted and joined to the image's annotation. This makes data collection in our context difficult and expensive, and therefore motivates the need for a model able to recognize micro-organisms from a few examples.

- As one might imagine, our classification problem is one of many classes. Additionally, this set of classes is subject to change, both in time with the collection of data for more and more micro-organisms and at the session level: indeed, based on additional information provided by the user, only a subset of classes may be considered. Few-Shot Learning algorithms are

| Real-World Problem | Exists in the use case... | | | | Addressed in... |
|---|---|---|---|---|---|
| | Maintenance | e-Shopping | Food Recognition | Microscopic FSL | |
| Support / Query Domain Shift | ✓ | ✓ | | | Chapter 3 |
| Open-Set Recognition | | ✓ | | ✓ | Chapter 4 |
| Fine-grained tasks | ✓ | ✓ | ✓ | ✓+ | Chapter 5 |
| Different shapes of Few-Shot tasks | ✓ | ✓ | | ✓ | Chapter 5 |
| Other metrics than top-1 accuracy | ✓ | ✓ | | | Chapter 5 |
| Mix large-scale & few-shot classes | | | ✓ | | - |
| Intra-class variability | | | | ✓ | Chapter 6 |

Table 1.1: Summary of the observed differences between each use case encountered at Sicara and the standardized academic setting. The rightmost column indicates the chapter of this thesis in which this difference is addressed, if any.

by nature well adapted to a changing set of classes, as they can adapt to a new set of classes without the need for re-training.

Granular yet variable classes. As in all previous use cases, our classification problem is obviously fine-grained. Indeed, we would need to distinguish between organisms from the same domain of living entities, and further on from the same kingdom, phylum, class, order, family, and even genus, making it one of the most fine-grained classification tasks. Also, as stated before, our problem involves tens of thousands of classes. Therefore, we cannot rely on benchmarks that compare Few-Shot Learning models on 5-way tasks to select the best model. Furthermore, each micro-organism presents a high intra-class variability, as its appearance depends on the type of microscope, magnification, and series of used reactants. Finally, as in most previous use cases, it is of prior importance to recognize when the query image does not belong to any of the known classes, in order not to provide falsely confident predictions.

### 1.2.5 Limitations of the standardized Few-Shot setting

We just learned that the challenges presented by real industrial use cases unveil many shortcomings of the standardized Few-Shot Learning setting, which are summarized in Table 1.1.

From just four use cases, we identified seven challenges of industrial applications that are not tackled by current Few-Shot Learning academic research.

- Domain shift between query and support images;
- Recognition of query instances which class does not appear in the support set;
- Fine-grained classification tasks;
- Variability of the evaluation setup, for instance in the shape of the few-shot classification tasks on which the models are tested;
- Other metrics than top-1 accuracy;
- Mix between large-scale and few-shot classes;

- Huge variability in the images that belong to each class.

Note that a common factor in all of these use cases is that the classes present a very fine granularity *i.e.*, they correspond to concepts that are very close to each other. This contrasts with the most popular benchmarks in Few-Shot Learning: all few-shot tasks sampled from the test set for evaluation involve 5 classes sampled uniformly at random among the many classes composing the test set. In practice, this sampling method generates almost exclusively few-shot tasks composed of concepts that are very distant from each other, as we show in Chapter 5.

These observations motivated our research team at Sicara and CentraleSupélec to focus on narrowing the gap between the hypothesis used in academic research and industrial applications of Few-Shot Learning.

## 1.3 Narrowing the Gap

The three years of work leading to this thesis have been directed at mitigating the limitations identified in Section 1.2 in order to make Few-Shot Learning research more directly applicable to industrial use cases. Our contributions can be organized into two main blocks:

1. Opening Few-Shot Learning to the challenges of real-world applications, by confronting Few-Shot Learning with other known problems in Machine Learning literature;
2. Understanding and improving Few-Shot Image Classification benchmarks.

### 1.3.1 Opening Few-Shot Learning to the Challenges of Real-World Applications

**Contribution 1: Few-Shot Learning under Support-Query Shift**

In Chapter 3, we address the problem of domain shift between support examples and query images. We formalize this problem as Few-Shot Learning under Support-Query Shift (FSQS), propose a standard evaluation protocol along with a testbed of three challenging benchmarks, introduce a novel method to solve this problem, and conduct extensive experimentation on several representative few-shot algorithms. Specifically:

1. We introduce FewShiftBed: a testbed for FSQS available at `https://github.com/ebennequin/meta-domain-shift`. The testbed includes 3 challenging benchmarks along with a protocol for fair and rigorous comparison across methods as well as an implementation of relevant baselines, and an interface to facilitate the implementation of new methods.
2. We conduct extensive experimentation using several representative few-shot algorithms. We empirically show that *Transductive* Batch-Normalization (Bronskill et al. 2020) mitigates an important part of the inopportune effect of FSQS.
3. We bridge *Unsupervised Domain Adaptation* (UDA) with FSL to address FSQS. We introduce *Transported Prototypes*, an efficient transductive algorithm that couples *Optimal Transport* (OT) from Peyré et al. 2019 with the celebrated *Prototypical*

*Networks* (Snell et al. 2017). The use of OT follows a long-standing history in UDA for aligning representations between distributions (Ben-David et al. 2007; Ganin and Lempitsky 2015). Our experiments demonstrate that OT shows a remarkable ability to perform this alignment even with only a few samples to compare distributions and provide a simple but strong baseline.

**Contribution 2: Transductive Few-Shot Open-Set Recognition**

In Chapter 4, we address a second relaxation of the assumptions made in the standardized Few-Shot Learning setting, namely the assumption that query images belong to the classes defined in the support set. This novel problem exists in the literature under the name of Few-Shot Open-Set Recognition (FSOSR).

1. We expose the specific difficulty of the FSOSR problem when using off-the-shelf pre-trained models, on a wide range of benchmarks and architectures, using our novel Mean Imposture Factor metric which measures how much the classes' distributions in a dataset are perturbed by instances from other classes.
2. 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.
3. We introduce Open-Set Transductive Information Maximization (OSTIM), an intuitive modification of the TIM method that provides an additional prototype for outliers.
4. We introduce Open-Set Likelihood Optimization (OSLO), a principled extension of the Maximum Likelihood framework that explicitly models and handles the presence of outliers. OSTIM and OSLO are interpretable and modular *i.e.*, can be applied on top of any pre-trained model seamlessly.
5. Through extensive experiments spanning five datasets and a dozen of pre-trained models, we show that our methods consistently surpass all existing methods in detecting open-set instances while competing with the strongest methods in classifying closed-set instances. Our empirical studies include long-overdue experiments on the performance of transductive methods with various sizes and shapes of the query set.

### 1.3.2 Challenges in Benchmarking Few-Shot Image Classification models

**Contribution 3: Semantic Similarity in Few-Shot Learning Benchmarks**

Following our observations on the shift between how we evaluate Few-Shot Learning methods in academic benchmarks and what we need them to do in real use cases, we provide in Chapter 5 an in-depth study focused on the similarity between the classes

appearing in a test task sampled from current benchmarks. We expose the limitations of current task sampling strategies and propose a step towards more realistic benchmarks. More specifically:

1. We use the WordNet taxonomy (Miller 1995) to evaluate *semantic distances* between classes of the popular Few-Shot Classification benchmark *tiered*ImageNet. Based on these semantic distances we put forward the concept of *coarsity* of an image classification task, which quantifies how semantically close are the classes of the task.

2. We conduct both quantitative and qualitative studies of the tasks generated from the test set of *tiered*ImageNet *i.e.*, the tasks composing the benchmark on which most papers evaluate different methods. We show that this benchmark is heavily biased towards tasks composed of semantically unrelated classes.

3. We harness the semantic distances between classes to generate the improved benchmark *better-tiered*ImageNet reestablishing the balance between fine-grained and coarse tasks. We compare state-of-the-art Few-Shot Classification methods on this new benchmark and bring out the relation between the *coarsity* of a task and its difficulty.

4. We put forward the Danish Fungi 2020 dataset (Picek et al. 2022) for evaluating Few-Shot Classification models. This dataset offers a wide range of fine-grained classes and therefore allows the sampling of tasks that we deem to be more representative of industrial applications of Few-Shot Learning. We compare state-of-the-art methods on both 5-way and 100-way tasks generated from this dataset. To the best of our knowledge, these are the first published results of few-shot methods on such wide tasks.

### Perspective: Observations on Support Set Quality

Data quality is universally known as a critical factor for performance in Machine Learning problems. Intuitively, the quality of individual samples is even more decisive when only a handful of labeled instances are available. Let us recall the context of Few-Shot Image Classification, with only one available sample per class (*i.e.*, *one-shot* classification). If the only example that we give an agent to define a class is of bad quality, we cannot hope for a good performance in recognizing this class. In Chapter 6, we report our investigations and present what we deem to be interesting perspectives for future works on the support set's quality. More precisely, we focused on the two following questions:

1. How can we characterize the quality of an example in a support set?

2. What is the impact of the selection of support set instances on the performance of Few-Shot Learning models?

## 1.4 Content's summary

In Part I, we provide the necessary background on Few-Shot Learning and Few-Shot Image Classification and draw the landscape of Few-Shot Learning's state of the art, both addressing the standardized setting and its relaxations.

In Part II, we expose our contributions towards opening Few-Shot Learning research by relaxing unrealistic assumptions and introducing:

1. Domain Shift between the support and query sets (Chapter 3);
2. Open-Set Recognition (Chapter 4).

Finally, in Part III, we expose our contribution towards improving Few-Shot Learning benchmarks using the semantic relations between classes (Chapter 5) and present promising perspectives related to the quality of the support set (Chapter 6).

# Part I

# The Foundations of Few-Shot Learning

# 2 Overview: Learning from a Few Examples

## 2.1 The many Paradigms on Learning with Limited Data

In this thesis, we focus on the paradigm of Few-Shot Learning *i.e.*, learning from few labeled samples. It is crucial to understand that this is just one of many paradigms around the vast problem of learning with limited data. In this section, we provide an overview of this family of problems and showcase their similarities and differences.

### 2.1.1 When Data is not Limited: standard Supervised Learning

Supervised Learning (SL) is a learning framework in which we assume access to the target prediction for each input example. When considering a classification task, it means that for each input $\boldsymbol{x} \in \mathcal{X}$ with $\mathcal{X}$ a given input space, a ground truth label $y \in \mathbb{C}$ is provided. A classification model is typically a mapping $\psi_{\boldsymbol{\theta}} : \mathcal{X} \longrightarrow [0,1]^{|\mathbb{C}|}$ with a set of parameters $\boldsymbol{\theta}$, that predicts for each input the probability that it belongs to each of the classes in $\mathbb{C}$.

Training a model $\psi_{\boldsymbol{\theta}}$ using Supervised Learning consists in:

1. sampling a batch of input-target pairs $\{(\boldsymbol{x}_i, y_i) \in \mathcal{X} \times \mathbb{C}\}_i$;
2. using the model's current parameters $\boldsymbol{\theta}$, predicting for each input $\boldsymbol{x}$ a prediction $\boldsymbol{p}_i = \{p_{ik}\}_{k \in \mathbb{C}}$ with $p_{ik} = \mathbb{P}(y_i = k | \boldsymbol{x}_i)$;
3. comparing the predicted labels with the ground truth, typically using the cross-entropy loss defined for a prediction $\boldsymbol{p}_i$ and a label $y_i$ as

$$\mathcal{L}_{CE} = -\sum_{k=1}^{|\mathbb{C}|} \mathbf{1}(y_i = k) \log p_{ik}$$

   with $\mathbf{1}$ the indicator function;
4. updating the parameters $\boldsymbol{\theta}$ using the computed loss;
5. repeat.

It is to this day the most commonly studied scenario, especially for Image Classification. There we assume access to a large dataset of *(image, label)* tuples. Here, we mean by "large" that we not only have many images but also many images corresponding to each existing label.

Supervised Learning allowed many breakthroughs in Image Recognition and He et al. 2016 even achieved super-human performance on the ImageNet Large Scale Visual Recognition Challenge.

However, the assumptions behind Supervised Learning were found to rarely hold in practice (see Section 1.2). The overall volume of data may be small, or it may be large but partially or fully unlabeled.

### 2.1.2 Learning with Limited Labels

Semi-Supervised Learning. In Semi-Supervised Learning (SSL), we assume access to a large dataset, but only a small (*e.g.*, 1% to 10%) proportion of labeled instances. The goal is to leverage this large body of unlabeled data to improve classification, using relevant assumptions on the structure of the data (Chapelle et al. 2009). There are many families of approaches to solving this problem, including pseudo-labeling of unlabeled instances (Lee et al. 2013; Sohn et al. 2020), consistency training (Laine and Aila 2017; Miyato et al. 2018), and generative models (Kingma et al. 2014; Odena 2016).

Unsupervised Learning. Going further into relaxing the labeling assumptions is the paradigm of Unsupervised Learning (UL). Here, we consider only unlabeled data. In the field of Image Recognition, this means that the model can only train on images, without any further information. An increasingly popular objective that can be achieved through Unsupervised Learning is Representation Learning (Bengio, Courville, et al. 2013; T. Chen et al. 2020; Oord et al. 2018; Wu et al. 2018). It consists in training a model not to output a prediction linked to a specific task (like predicting the class in Supervised Learning), but rather to map each data point (*e.g.*, an image) into a general feature space in which only the useful concepts and patterns are represented. The quality of the representation can then be measured on a variety of downstream tasks *e.g.*, image classification, object detection, instance segmentation, or even few-shot image classification.

Unsupervised Domain Adaptation. Another well-established paradigm of limited labels is Unsupervised Domain Adaptation (UDA). Domain Adaptation consists in adapting a model that was trained on data sampled from a source distribution so that it can perform accurately on data sampled from a different target distribution. In Image Recognition, many parameters can cause such a distribution shift, such as a change in the acquisition process, the lighting conditions, or simply because of a switch in the context in which source and target data are provided (see Figure 1.3). UDA has a long-standing story (Pan and Q. Yang 2009; Quionero-Candela et al. 2009). The analysis of the role of representations from Ben-David et al. 2007 has led to wide literature based on domain invariant representations (Ganin and Lempitsky 2015; Long, Y. Cao, et al. 2015). Outstanding progress has been made toward learning more domain transferable representations by looking for domain invariance. The tensorial product between representations and prediction promotes conditional domain invariance (Long, Z. Cao, et al. 2018), the use of weights (Bouvier et al. 2020; Z. Cao et al. 2018; Tachet des Combes et al. 2020; You et al. 2019) has dramatically improved the problem of label shift theoretically described in Y. Zhang, T. Liu, et al. 2019, hallucinating consistent target samples (H. Liu et al. 2019), penalizing high singular values of a batch of representations (X. Chen, S. Wang, et al. 2019) or by enforcing the favorable inductive bias of consistency through various data augmentation in the target domain (Ouali, Bouvier, et al. 2020). Recent works address the problem of adaptation without source data (Liang et al. 2020; Yeh et al. 2021). The seminal work by Courty, Flamary, Tuia, et al. 2016, followed

| Paradigm | Data volume | Labels | Domain Shift |
|---|---|---|---|
| Supervised Learning | Large | All | No |
| Semi-supervised Learning | Large | Few | No |
| Unsupervised Learning | Large | None | No |
| Unsupervised Domain Adaptation | Large | Only in source domain | Yes |
| Transfer Learning | Large in source, no assumption in target | All | Yes |
| Test-Time Adaptation | Large in source domain, small in target domain | Only in source domain | Yes |
| Few-Shot Learning | Small | All | No assumption |

Table 2.1: Overview of the differences between the learning paradigms.

by Courty, Flamary, Habrard, et al. 2017 and Bhushan Damodaran et al. 2018, brings Optimal Transport (OT) to UDA by transporting source samples in the target domain.

Transfer Learning. In UDA, we assumed that the task to be solved in the target domain was identical to the task to be solved in the source domain. When the assumption break, we talk about Transfer Learning (Pan and Q. Yang 2009). For instance, in image classification, the set of classes in which images are to be classified can be different in the target domain, resulting in a new target task. In this case, we require labeled data in the target domain as well.

Test-Time Adaptation. Test-time Adaptation (TTA) goes one step further by adapting to the target domain *at test-time*. In Y. Sun et al. 2020, adaptation is performed by test-time training of representations through a self-supervision task which consists in predicting the rotation of an image. This leads to a successful adaptation when the gradient of the fine-tuning procedure is correlated with the gradient of the cross-entropy between the prediction and the label of the target sample, which is not available. Inspired by UDA methods based on domain invariance of representations, a line of works (Nado et al. 2020; Schneider et al. 2020) aims to align the mean and the variance of train and test distribution of representations. This is simply done by updating statistics of the batch-normalization layer. In a similar spirit of leveraging the batch-normalization layer for adaptation, D. Wang et al. 2021 suggest minimizing prediction entropy on a batch of test samples, as suggested in semi-supervised learning (Grandvalet and Bengio 2005). As pointed out by authors of D. Wang et al. 2021, updating the whole network is inefficient and raises a risk of overfitting on the test batch. To address this problem, the authors suggest only updating batch-normalization parameters for minimizing the prediction's entropy. The paradigm of *Adaptative Risk Minimization* (ARM) is introduced in M. Zhang et al. 2021. ARM aims to adapt a classifier at test time by conditioning its prediction on the whole batch of test samples (not only one sample). Authors demonstrate that such a classifier is meta-trainable as long as the training data exposes a structure of group.

### 2.1.3 Few-Shot Learning: Fully Labeled but Limited Data

Now how does Few-Shot Learning fit into this landscape?

A Few-Shot Classification Task consists in recognizing images among a set of classes, given only a few labeled samples for these classes. To do so, we allow access to a large *base* dataset for training,

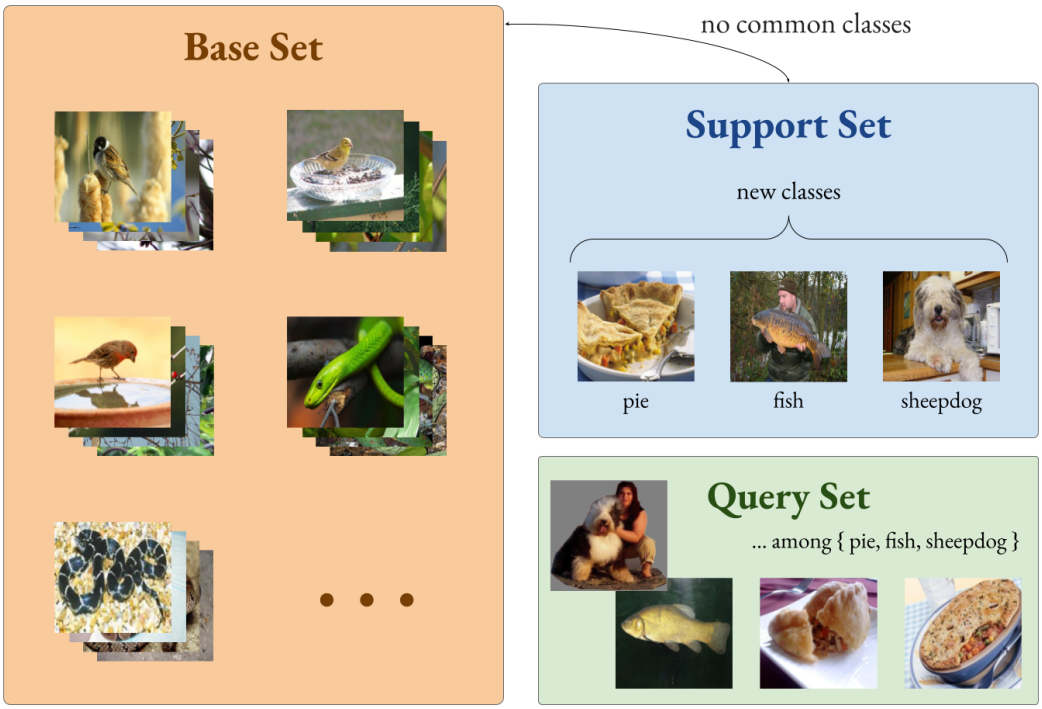



Figure 2.1: The Few-Shot Image Classification problem illustrated with images from the popular *tiered*ImageNet benchmark. To classify query images among the few-shot task's set of classes, we only have access to one example per class (this is the particular case of a 1-shot task). The model is allowed to leverage information from a base set, with the requirement that its set of classes is disjoint from those at test time.

with the only requirement that the classes involved in the few-shot task at test time do not appear in this base dataset. This problem is represented in Figure 2.1.

This differs from the previously introduced "few-labels" paradigms, as each data point involved in this problem is associated with a ground-truth label[1]. On the other hand, the volume of data is way smaller than the volumes considered in semi-supervised and unsupervised learning. This problem can also be compared with Test-Time Adaptation since we need to adapt to new data at test time. The difference is that we do not necessarily need to adapt to a new domain, but rather new classes. There is no assumption in the Few-Shot setting about domain shift, and models are usually evaluated on both in-domain and cross-domain benchmarks (see Section 2.2.4).

## 2.2 Background on Few-Shot Image Classification

We can now dive into the Few-Shot Image Classification problem, its definition, its specifics, and the vast landscape of methods designed to solve this problem.

[1]Although it is allowed to consider unsupervised training on the base set, as we do in Chapter 4.

### 2.2.1 Problem formalization

**Basics** In the following, we consider an input space $\mathcal{X}$ and a label space $\mathcal{Y}$. A representation is a learnable function $\phi_{\boldsymbol{\theta}} : \mathcal{X} \mapsto \mathcal{Z}$ with $\mathcal{Z}$ a feature space and $\boldsymbol{\theta}$ a set of learnable parameters.

**A Few-Shot Classification Task** Given a set of classes $\mathbb{C} \subset \mathcal{Y}$ with $|\mathbb{C}| = K$, a $K$-way Few-Shot Classification task[2] $\mathbb{T}_{\mathbb{S},\mathbb{Q}}$ is formed by a small support set of labeled instances $\mathbb{S} = \{(\boldsymbol{x}_i^s, y_i^s) \in \mathcal{X} \times \mathbb{C}\}_{i=1...|\mathbb{S}|}$ and a query set $\mathbb{Q} = \{\boldsymbol{x}_i^q \in \mathcal{X}\}_{i=1...|\mathbb{Q}|}$. In the standard few-shot setting, the unknown ground-truth query labels $\{y_i^q\}_{i=1...|\mathbb{Q}|}$ are assumed to be restricted to closed-set classes *i.e.*, $\forall i,\ y_i^q \in \mathbb{C}$. When $\mathbb{S}$ contains exactly $n$ instances for each class $k \in \mathbb{C}$, $\mathbb{T}_{\mathbb{S},\mathbb{Q}}$ is called a $K$-way $n$-shot classification task. The goal of Few-Shot Classification is to assign to each query instance $\boldsymbol{x}_i^q$ a prediction $\boldsymbol{p}_i^q = \{p_{ik}^q\}_{k \in \mathbb{C}}$ with $p_{ik}^q = \mathbb{P}(y_i^q = k | \boldsymbol{x}_i^q)$, with no prior knowledge about the classes in $\mathbb{C}$ except for the information in $\mathbb{S}$.

**Evaluation** Consider a base set $\mathcal{D}_{\text{base}} = \{(\boldsymbol{x}, y) \mid \boldsymbol{x} \in \mathcal{X},\ y \in \mathcal{C}_{\text{base}}\}$ and a test set $\mathcal{D}_{\text{test}} = \{(\boldsymbol{x}, y) \mid \boldsymbol{x} \in \mathcal{X},\ y \in \mathcal{C}_{\text{test}}\}$ where $\mathcal{C}_{\text{base}} \subset \mathcal{Y}$ and $\mathcal{C}_{\text{test}} \subset \mathcal{Y}$ are their respective class sets, with $\mathcal{C}_{\text{base}} \cap \mathcal{C}_{\text{test}} = \emptyset$. The standard formulation of the Few-Shot Classification problem consists in learning on $\mathcal{D}_{\text{base}}$ a classification model that can generalize to the unseen classes in $\mathcal{C}_{\text{test}}$ with only a few training examples per class. This is evaluated by sampling from $\mathcal{D}_{\text{test}}$ a large number of Few-Shot tasks. To do so, we define $\mathcal{E}_{\text{test}}(K)$ (and similarly $\mathcal{E}_{\text{base}}(K)$) as the set of $K$-way classification tasks that can be sampled from $\mathcal{D}_{\text{test}}$ *i.e.*,

$$\begin{aligned}\mathcal{E}_{\text{test}}(K) \coloneqq \{\mathbb{T}_{\mathbb{S},\mathbb{Q}} \mid \mathbb{S} = \{(\boldsymbol{x}_i^s, y_i^s) \in \mathcal{X} \times \mathbb{C}\}_{i=1...|\mathbb{S}|} \subset \mathcal{D}_{\text{test}}, \\ \mathbb{Q} = \{\boldsymbol{x}_i^q \in \mathcal{X}\}_{i=1...|\mathbb{Q}|} \subset \mathcal{D}_{\text{test}}, \\ \mathbb{S} \cap \mathbb{Q} = \emptyset,\ \mathbb{C} \subset \mathcal{C}_{\text{test}} \text{ and } |\mathcal{C}_{\text{test}}| = K\}\end{aligned} \tag{2.1}$$

In practice, most benchmarks are limited to $\mathcal{E}_{\text{test}}(5)$, and further limited to tasks with 1 or 5 support images per class, and 10 query images per class. The number of possible tasks is still untractable on most datasets. Therefore we most often evaluate few-shot classification models on a subset $\tilde{\mathcal{E}}_{\text{test}} = \{\mathbb{T} \in \mathcal{E}_{\text{test}} | \mathbb{T} \sim \mathcal{U}\}$ with $\mathcal{U}$ the uniform distribution.

**Episodic training** A large body of Few-Shot Learning methods trains on the base set $\mathcal{D}_{\text{base}}$ using episodic training. It consists in mimicking the evaluation process. Given an arbitrarily chosen $K$, episodic training minimizes over $\mathcal{E}_{\text{base}}(K)$ the cross-entropy loss on query instances:

$$\mathcal{L}_{\text{CE}} = \mathbb{E}_{\mathbb{T}_{\mathbb{S},\mathbb{Q}} \in \mathcal{E}_{\text{base}}(K)} \left[ -\frac{1}{|\mathbb{Q}|} \sum_{i=1}^{|\mathbb{Q}|} \sum_{k=1}^{K} y_{ik}^q \log(p_{ik}^q) \right] \tag{2.2}$$

with $\boldsymbol{y}_i^q \in \{0,1\}^{|\mathbb{C}|}$ the one-hot encoding of the ground truth label associated to $\boldsymbol{x}_i^q$.

Following Vinyals et al. 2016, many contributions to Few-Shot Learning (Finn et al. 2017; Snell et al. 2017; Sung et al. 2018) considered that episodic training was an essential part of the problem. It followed the very popular and exciting idea of "learning to learn", or "meta-learning", in which

[2] We may also write $\mathbb{T}_{\mathbb{C}}$ when the study focuses on the classes rather than the images composing the task.

the model is supposed to learn *across tasks* to better adapt to new tasks. "Meta-learning algorithms for few-shot computer vision" was even the topic of the preliminary work Bennequin 2019 to this thesis. In the next section, after a thorough review of the various techniques developed to solve Few-Shot Learning, we will discuss the relevance of the episodic training strategy.

### 2.2.2 Few-Shot Learning methods

It is commonly assumed, following a classification proposed in W.-Y. Chen et al. 2019, to categorize Few-Shot algorithms into three categories: metric-based, optimization-based, and hallucination-based. In the following, we describe the principle and the main contributions of each of these families.

Metric-based. Metric-based methods consist in casting both support and query images to a representation space, and then classifying query instances based on some distance to support instances in this space. Most metric-learning methods are built on the principle of the seminal work around Siamese Networks by Koch et al. 2015, which proposes a solution to the few-shot image classification problem on the Omniglot dataset[3] by using a neural network as a feature extractor for images. The model is trained on the base set using contrastive loss (Hadsell et al. 2006). Then, at inference time, query images are classified by comparison to all support images, with cosine distance. Later contributions build upon Siamese Networks, while also exploiting the episodic training paradigm: they learn a feature extractor across training tasks sampled from the base set (Vinyals et al. 2016). Prototypical Networks (Snell et al. 2017) follow the same idea but compare queries only to one prototype per class (computed as the mean of all feature vectors of the instances of this class in the support set). They choose Euclidean distance instead of cosine distance and show better results. Relation Networks (Sung et al. 2018) add another deep network on top of Prototypical Networks to compute similarities between query instances and prototypes, thus replacing the Euclidean distance. Z. Jiang et al. 2020 add an attention module to refine query embeddings using information from the support set. Ouali, Hudelot, et al. 2021 show that training with a contrastive loss can improve the generalization capabilities of few-shot methods. Laenen and Bertinetto 2021 suggest that Prototypical networks perform better without episodic training. They replace the support-query splitting of each training batch with a Neighboring Component Analysis based on all pairs of images. S. Yang et al. 2021 propagate the statistics of base classes to similar support classes to improve cluster definition. The same idea is followed by Roy et al. 2020 to augment few-shot classes in a context where some classes have few examples and others are large-scale.

To sum up, there are three degrees of liberty in the space of metric-based methods for few-shot classification:

- Training strategy: episodic training or contrastive loss. Note that it is also possible to train the feature extractor on the base set by training a classifier between all base classes, and simply removing the head.
- Comparison strategy: compare queries to every image, or simply to one prototype per class.

[3] Presented in Section 2.2.4.

- Distance: cosine, euclidean, or parameterized with a neural network.

Optimization-based. These methods usually learn to fine-tune a model on a small support set. The MAML method (Finn et al. 2017) uses episodic training to learn a good model initialization *i.e.*, model parameters that can adapt to a new task with novel classes in a small number of gradient steps. Their experiments are extended in Nichol et al. 2018, with a focus on first-order approximations. Triantafillou et al. 2020 propose an iteration named ProtoMAML, which consists in initializing the last layer of the model with the prototypes computed from the support set as in Snell et al. 2017. Meta-LSTM (Ravi and Larochelle 2017) meta-train an LSTM to perform gradient descent on new few-shot tasks. Meta-Networks (Munkhdalai and H. Yu 2017) also replaces standard gradient descent with a meta-learned optimizer. These methods are widely used in few-shot reinforcement learning, but often appear too complex for computer vision tasks, compared to metric-based methods. MAML, for instance, uses a back-propagation through a back-propagation, which makes the computation time prohibitive even with first-order approximations and makes the hyper-parameter tuning very difficult (Antoniou, Edwards, et al. 2019).

Hallucination-based. This third family of methods consists in augmenting the small support set with artificially generated samples. Hariharan and Girshick 2017 train a network to take support features as input and generate new "hallucinated" features to augment the support set. Y.-X. Wang et al. 2018 follow this idea and incorporate meta-learning to train the "hallucinator", while Antoniou, A. Storkey, et al. 2017 use Generative Adversarial Networks to augment the support set.

Discussions on episodic training. A large part of the methods presented above involve episodic training, 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. It has become a prominent part of the Few-Shot Learning literature, as many methods (some metric-based, and all optimization-based and hallucination-based methods) included architectures specifically designed to be trained at the task level. However, several recent works indicate that the results of these methods can be matched by simple fine-tuning algorithms *e.g.*, W.-Y. Chen et al. 2019 and Dhillon et al. 2020 show results indicating that simple methods based on fine-tuning can perform reasonably well on few-shot tasks. They implement a baseline where a pre-trained network's head is initialized on a new support set simply by using the class prototypes extracted from the pre-trained network. Goldblum, Reich, et al. 2020 show that metric-based methods usually provide better class-wise clustering on novel classes than classically trained networks, which could explain their good performance on comparison tasks. They show that adding a simple regularizing term in the loss when training a standard classifier can achieve the same results.

### 2.2.3 Transductive Few-Shot Image Classification

We distinguish transductive classification from inductive classification. Transductive is when we assume that we have access to the whole query set, so we can use query images as unlabeled data to

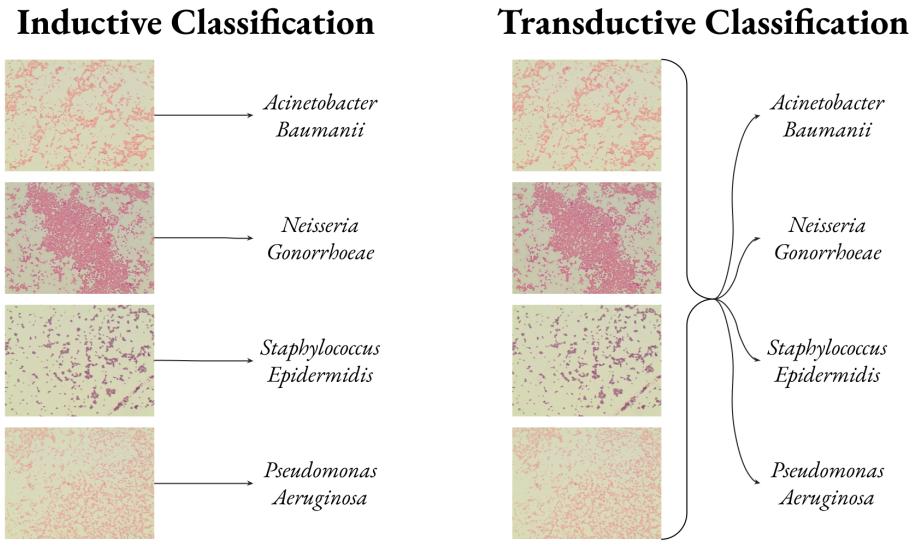



Figure 2.2: Distinction between inductive and transductive classification as defined in the Few-Shot Classification literature, illustrated with images from the Digital Image of Bacterial Species (DIBaS) dataset provided in Zieliński et al. 2017. Inductive classification is performed on each query independently, while transductive classification is performed on a batch, allowing the use of other queries as additional unlabeled data.

improve the classification. It has recently become a popular relaxation of the classical Few-Shot Classification problem. The distinction is illustrated in Figure 2.2.

Transductive Propagation Networks (Y. Liu et al. 2019) meta-learn label propagation from support to query set concurrently with the feature extractor. Antoniou and A. J. Storkey 2019 proposed to use a meta-learned critic network to further adapt a classifier on the query set in an unsupervised setting. Ren et al. 2019 extend Prototypical Networks in order to use the query set in the prototype computation. L. Liu et al. 2019 use the confidently predicted queries to refine the prototype computation, in an attempt to reduce the bias in prototype computation. Transductive Information Maximization (Boudiaf, Ziko, et al. 2020) aims at maximizing the mutual information between the features extracted from the query set and their predicted labels while also minimizing the cross-entropy loss on the support set. Transductive Fine-Tuning (Dhillon et al. 2020) consists in initializing the last layer of a pre-trained network on a few-shot task with the prototypes computed from the support set and performing fine-tuning on the support set. In this fine-tuning, they use as a regularizer the classification entropy of query set instances. Finally, PT-MAP and Transported Prototypes (Bennequin 2019; Hu et al. 2021) use Optimal Transport to align query and support sets.

The idea of maximizing the likelihood of both support and query samples under a model parameterized by class prototypes is proposed by B. Yang et al. 2020 for few-shot segmentation. However, their method relies on the closed-set assumption. Differing from previous works, the novel framework presented in Chapter 4 leverages an additional latent variable, the *inlierness* score.

### 2.2.4 Benchmarks

In the following, we describe the usual benchmarks used to assess the performance of Few-Shot Image Classification models.

Omniglot. The first and most simple few-shot classification benchmark is based on the Omniglot dataset of handwritten characters (B. Lake et al. 2011). The dataset contains 1623 handwritten characters drawn by 20 different people. Each image is associated with the sequence of strokes that were used to write the character, although most benchmarks only use the images. Models evaluated on Omniglot usually use four different settings corresponding to four shapes of the randomly sampled few-shot tasks: 1 and 5-shot, 5 and 20-way. It has been used as an MNIST-like benchmark at the early stages of Few-Shot Learning research but has recently fallen out of fashion as the classification accuracy of most methods approached 100% on the most challenging 1-shot 20-way benchmarks.

*mini*ImageNet. Rather quickly, *mini*ImageNet (Vinyals et al. 2016) became the reference dataset for Few-Shot Learning. It is a small subset of ImageNet (Deng et al. 2009) containing 60k images in 100 classes, split across classes as train/val/test. Since it was first introduced in Vinyals et al. 2016, *mini*ImageNet has only been used in the 5-way 1-shot and 5-way 5-shot settings.

*tiered*ImageNet. Later, Ren et al. 2019 introduced the larger dataset *tiered*ImageNet, also built from ImageNet but with 608 classes, which are split in a way that preserves the super-category structure of the classes. Ren et al. also upscaled in terms of the number of images per class by using all the available images in ImageNet for each class ($\sim$ 1300 images per class). Like *mini*ImageNet, *tiered*ImageNet has almost always been used in the 5-way setting with either 1 or 5 shots.

CIFAR-FS. The exact same strategy was used by Bertinetto et al. 2019 to build a benchmark sampled from the CIFAR-100 dataset (Krizhevsky, Hinton, et al. 2009), which is a dataset of 60k three-channel square images of size $32 \times 32$, evenly distributed in 100 classes. Classes are evenly distributed in 20 superclasses. CIFAR-FS is not as popular in the Few-Shot Learning community as its ImageNet counterparts[4].

CU-Bird. The classification benchmark CU-Birds 200 (Welinder et al. 2010), which happens to be a fine-grained benchmark, is now also used for Few-Shot Image Classification, especially to study cross-domain robustness (W.-Y. Chen et al. 2019). It is used in the same fashion as previous benchmarks *i.e.*, to build 5-way tasks with 1 or 5 shots.

Meta-Dataset. We established that all these benchmarks roughly follow the same process to generate few-shot classification tasks *i.e.*, we sample 5 classes [5] uniformly at random from the test set, then sample a fixed number of images per class for the support set and for the query set. However, more recently, Triantafillou et al. 2020 merged 10 computer vision datasets to build a

[4] 45 usages of CIFAR-FS in 2022 according to PapersWithCode, versus 60 for *tiered*ImageNet and 213 for *mini*ImageNet.

[5] Rarely 10, or 20 for Omniglot

gigantic benchmark for few-shot classification methods: Meta-Dataset. They introduced some randomness in the shape of the tasks (number of ways, shots, and queries) and proposed to study the hierarchy of the methods depending on the number of ways and shots. Despite its ability to benchmark methods on incredibly diverse datasets and tasks, Meta-Dataset remains underused by the community compared to a simpler and more lightweight benchmark like *mini*ImageNet[6]. In Chapter 5, we mitigate the biases of current benchmarks while avoiding additional engineering challenges that would make it harder to adopt our novel benchmarks.

## 2.3 Thinking about the Few-Shot Classification Tasks in detail

### 2.3.1 Sampling of Few-Shot Tasks

Curriculum Learning. Bengio, Louradour, et al. 2009's pioneering work on *Curriculum Learning* shows that the order in which training samples are shown to the model influences both convergence speed and the quality of the found local minimum. Q. Sun et al. 2019 propose to generate "hard" examples by biasing the sampling towards classes on which the model has shown the greatest loss during previous epochs. With the same objective but for the specific problem of episode sampling in the context of Few-Shot Learning, C. Liu et al. 2020 condition the probability of co-sampling two classes in one training episode on the average confusion between those classes. Still on the sampling of few-shot tasks, Triantafillou et al. 2020 incorporate to their benchmark the notion of task fine-graininess, *i.e.*, how "close" classes are to each other, following a predefined semantic. Sbai et al. 2020 show a correlation between the granularity of the classes in which the base set is split and the final accuracy on the test set. They show that classes can be artificially merged or further split to reach an optimum. Kaddour, Sæmundsson, et al. 2020 propose an algorithm to select the next task during episodic training for meta-reinforcement learning.

Characterizing classification tasks. Some recent works try and find a way to represent classification tasks so that they can be compared with one another (Achille et al. 2019; Nguyen et al. 2021). Here we focus on characterizing tasks using class semantics. Deselaers and Ferrari 2011 show that on ImageNet, visual and semantic similarities between classes are linked. They measure the semantic similarity with the Jiang & Conrath pseudo-distance (J. J. Jiang and Conrath 1997). It depends on the WordNet Directed Acyclic Graph, a semantic hierarchy spanning (among many other concepts) all classes in ImageNet, and on the number of images in each class. Other methods to evaluate the similarity between categories can be found in Alves et al. 2020.

### 2.3.2 Quality of the support set

To the best of our knowledge, only a handful of works address the issue of the quality of the support set. These contributions can be split into two categories.

The first category is that of methods that aim at mitigating an assumed global, not precisely defined bad quality of support sets, through the use of an additional module acting at the feature

[6] 22 usages of Meta-Dataset in 2022 according to PapersWithCode.

level *i.e.*, on the embeddings outputted by the feature extractor. MELR (Fei et al. 2020) adds a consistency regularization to the meta-training loss to encourage the model's prediction to be independent of the sampling of support examples. Lu et al. 2020 tackle the problem of outliers in the support set by adding an extra module in the feature space to mitigate their impact on the prototype computation (and therefore on the model's performance on the query set).

The second categories of contributions involve more in-depth studies of frequent flaws in support sets. Luo et al. 2021 postulate that the propensity of large-scale image recognition training strategies to use background information as a shortcut for classification arms the performance when the representations are applied to novel classes. Other works such as Bendou et al. 2022 and Hiller et al. 2022 focus on the issue of locating the object of interest in an image, which is even more important in the context of single-label images. Finally, J. Li et al. 2021 use an *erasing-inpainting* module during training to force the embeddings to store information from all regions of target objects.

Note that none of these studies propose a way to measure the quality of a support set. In Section 6.2, we draw what we deem to be useful perspectives toward quantifying the quality of support examples.

## 2.4 Opening the Few-Shot Image Classification Problem

As we established in Chapter 1, the standardized Few-Shot Image Classification setting that we described so far in this chapter does not always hold in real use cases. In particular, this standardized setting ignores very common issues like distributional shift and open-set recognition. The present thesis makes substantial contributions to these issues. However, we do not claim to be the very first to study Few-Shot Learning under Distributional Shift or Few-Shot Open-Set Recognition. In this section, we review the previous efforts made in these areas and provide useful context to understand the positioning of our contributions.

### 2.4.1 Few-Shot Classification under Distributional Shift

Recent works on few-shot classification tackle the problem of distributional shift between the base set and the test set. W.-Y. Chen et al. 2019 compare the performance of state-of-the-art solutions to few-shot classification on a cross-domain setting (training on *mini*ImageNet (Vinyals et al. 2016) and testing on Caltech-UCSD Birds 200 (Welinder et al. 2010)). A. Zhao et al. 2021 propose a Domain-Adversarial Prototypical Network in order to both align source and target domains in the feature space and maintain discriminativeness between classes. Following the "meta-learning" paradigm and considering the problem as a shift in the distribution of tasks (*i.e.* training and testing tasks are drawn from two distinct distributions), Sahoo et al. 2019 combine Prototypical Networks with adversarial domain adaptation at the task level. Goldblum, Fowl, et al. 2020 also propose adversarial data augmentation for the cross-domain scenario. While these works address the key issue of the distributional shift between the base and test set, they assume that for each task, the support set and query set are always drawn from the same distribution. We find that this assumption rarely holds in practice[7]. In Chapter 3 we consider a distributional shift both

[7] In the use cases 1.2.1 and 1.2.2, support and query images come from different distributions.

between base and test set and inside a task *i.e.*, between support and query set. Later on, Du et al. 2021 address the support-query shift problem with a multi-layer perceptron applied at each batch normalization layer of the network in order to predict relevant batch statistics. Building on the Support-Query Shift problem defined in Chapter 3, Aimen et al. 2023 propose to use adversarial projections to solve *inductive* Support-Query Shift, while S. Jiang et al. 2022 improve on our proposed Transported Prototypes approach to make it more robust to small perturbations in images.

### 2.4.2 Few-Shot Open-Set Recognition

Open-set recognition (OSR). The standard classification setting uses the assumption that all images indeed belong to the known classes. However, this assumption, while convenient, rarely holds in real-world applications. This motivated the need for 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 and Boult 2016), generating artificial outliers (Ge et al. 2017; Neal et al. 2018), improving closed-set accuracy (Vaze et al. 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 Chapter 4, we use simple yet effective adaptations of OpenMax (Bendale and Boult 2016) and PROSER (Zhou et al. 2021) as strong baselines for FSOSR.

Few-shot open-set recognition (FSOSR). In the few-shot setting, methods must detect open-set instances while only a few closed-set instances are available. B. Liu et al. 2020 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. S. 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 methods presented in Chapter 4 don't require any specific training and can be plugged into any feature extractor without further optimization.

# Part II

# Opening Few-Shot Learning to Real-World Problems

# 3 Contribution 1: Few-Shot Learning under Support-Query Shift

This chapter replicates our paper *Bridging Few-Shot Learning and Adaptation: New Challenges of Support-Query Shift*, by Etienne Bennequin, Victor Bouvier, Myriam Tami, Antoine Toubhans, and Céline Hudelot, produced in equal contribution with Victor Bouvier from the MICS laboratory, and published at ECML-PKDD 2021 (Bennequin, Bouvier, et al. 2021).

## 3.1 Introduction

The standard Few-Shot Learning setting detailed in Section 2.2.1 does not rely on any assumption about the distributions from which base instances, support instances, and query instances are sampled. However, well-adopted FSL benchmarks detailed in Section 2.2.4 (Ren et al. 2019; Triantafillou et al. 2020; Vinyals et al. 2016) commonly sample the support and query sets from the same distribution. We stress that this assumption does not hold in most use cases. When deployed in the real world, we expect an algorithm to infer on data that may shift, resulting in an acquisition system that deteriorates, lighting conditions that vary, or real-world objects evolving (Amodei et al. 2016). This is, for instance, one of the problems occurring in the use-case presented in Section 1.2.1.

The situation of *Distribution Shift* (DS) *i.e.*, when training and testing distributions differ, is ubiquitous and has dramatic effects on deep models (Hendrycks and Dietterich 2019), motivating works in *Unsupervised Domain Adaptation* (Pan and Q. Yang 2009), *Domain Generalization* (Gulrajani and Lopez-Paz 2021) or *Test-Time Adaptation* (D. Wang et al. 2021). However, the state of the art brings insufficient knowledge on few-shot learners' behaviors when facing distribution shift. Some pioneering works demonstrate that advanced FSL algorithms do not handle cross-domain generalization better than more naive approaches (W.-Y. Chen et al. 2019). Despite its great practical interest, FSL under distribution shift between the support and query sets is an under-investigated problem and attracts very recent attention (Du et al. 2021). We refer to it as ***F**ew-Shot Learning under **S**upport/**Q**uery **S**hift* (**FSQS**) and provide an illustration in Figure 3.1. It reflects a more realistic situation where the algorithm is fed with a support set at the time of deployment and infers continuously on data subject to shift. The first solution is to re-acquire a support set that follows the data's evolution. Nevertheless, it implies human intervention to select and annotate data to update an already deployed model, reacting to a potential drop in performance. The second solution consists in designing an algorithm that is robust to the distribution shift encountered during inference. This is the subject of this chapter.


(a) Standard FSL (b) FSL under Support / Query Shift

Figure 3.1: Illustration of the FSQS problem with a 5-way 1-shot classification task sampled from the mini-ImageNet dataset (Vinyals et al. 2016). In (a), a standard FSL setting where support and query sets are sampled from the same distribution. In (b), the same task but with shot-noise and contrast perturbations from Hendrycks and Dietterich 2019 applied on support and query sets (respectively) that results in a support-query shift. In the latter case, a similarity measure based on the Euclidean metric (Snell et al. 2017) may become inadequate.

**Chapter's Contributions**

1. We introduce FewShiftBed: a testbed for FSQS available at `https://github.com/ebennequin/meta-domain-shift`. The testbed includes 3 challenging benchmarks along with a protocol for fair and rigorous comparison across methods as well as an implementation of relevant baselines, and an interface to facilitate the implementation of new methods.

2. We conduct extensive experimentation of a representative set of few-shot algorithms. We empirically show that *Transductive* Batch-Normalization (Bronskill et al. 2020) mitigates an important part of the inopportune effect of FSQS.

3. We bridge *Unsupervised Domain Adaptation* (UDA) with FSL to address FSQS. We introduce *Transported Prototypes*, an efficient transductive algorithm that couples *Optimal Transport* (OT) from Peyré et al. 2019 with the celebrated *Prototypical Networks* (Snell et al. 2017). The use of OT follows a long-standing history in UDA for aligning representations between distributions (Ben-David et al. 2007; Ganin and Lempitsky 2015). Our experiments demonstrate that OT shows a remarkable ability to perform this alignment even with only a few samples to compare distributions and provide a simple but strong baseline.

In Section 3.2 we provide a formal statement of FSQS, and we position this new problem among existing learning paradigms. In Section 3.3, we present FewShiftBed. We detail the datasets, the chosen baselines, and a protocol that guarantees a rigorous and reproducible evaluation. In Section 3.4, we present a method that couples Optimal Transport with Prototypical Networks (Snell et al. 2017). Finally, in Section 3.5, we conduct an extensive evaluation of baselines and our proposed method using the testbed.

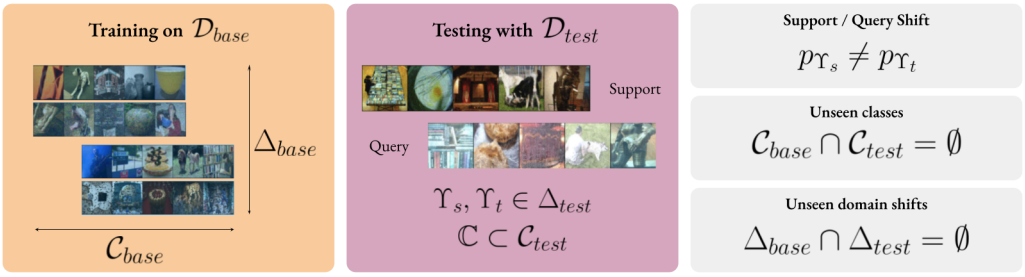



Figure 3.2: A Few-Shot Learning algorithm is trained on a base set $\mathcal{D}_{\text{base}}$ made of images from a set of domains $\Delta_{\text{base}}$ and labels from a set of classes $\mathcal{C}_{\text{base}}$. At test-time, the trained model is fed with a support set sampled from a new (source) domain $\Upsilon_s \in \Delta_{\text{test}}$ and new classes $\mathbb{C} \subset \mathcal{C}_{\text{test}}$ and is asked to classify query samples from another (target) domain $\Upsilon_t$ with $p_{\Upsilon_t} \neq p_{\Upsilon_s}$. Importantly, both classes and domain shifts are not seen during training ($\mathcal{C}_{\text{base}} \cap \mathcal{C}_{\text{test}} = \Delta_{\text{base}} \cap \Delta_{\text{test}} = \emptyset$), making Few-Shot Learning under Support-Query Shift a challenging problem of generalization.

## 3.2 The Support-Query Shift problem

### 3.2.1 Statement

**Domain Shift.** Similarly to the definition of $\mathcal{C}_{\text{base}}$ and $\mathcal{C}_{\text{test}}$ in Section 2.2.1, we define $\Delta_{\text{base}}$ (resp. $\Delta_{\text{test}}$) the set of domains represented in the base set (resp. the test set), with $\Delta_{\text{base}} \cap \Delta_{\text{test}} = \emptyset$. A domain $\Upsilon \subset \mathcal{X}$ is a set of Independent and Identically Distributed (IID) realizations from a distribution noted $p_\Upsilon$. Following this formalization, for this chapter we expand the definition of $\mathcal{D}_{\text{base}}$ and $\mathcal{D}_{\text{test}}$ to account for domain shift between the base and test set:

$$\mathcal{D}_{\text{base}} = \{(\boldsymbol{x}, y) \,|\, \boldsymbol{x} \in \Upsilon,\ y \in \mathcal{C}_{\text{base}},\ \Upsilon \in \Delta_{\text{base}}\}$$
$$\mathcal{D}_{\text{test}} = \{(\boldsymbol{x}, y) \,|\, \boldsymbol{x} \in \Upsilon,\ y \in \mathcal{C}_{\text{test}},\ \Upsilon \in \Delta_{\text{test}}\}$$

For two domains $\Upsilon, \Upsilon' \subset \mathcal{X}$, the distribution shift is characterized by $p_\Upsilon \neq p_{\Upsilon'}$. For instance, if the data consists of images of letters handwritten by several users, $\Upsilon$ can consist of samples from a specific user. Referring to the well-known Unsupervised Domain Adaptation (UDA) terminology of source/target (Pan and Q. Yang 2009), we define a couple of source-target domains as a couple $(\Upsilon_s, \Upsilon_t)$ with $p_{\Upsilon_s} \neq p_{\Upsilon_t}$, thus presenting a distribution shift.

**Few-Shot Classification under Support-Query Shift (FSQS).** We expand the definition of a Few-Shot Classification task given in Section 2.2.1 to define the FSQS task. Given a set of classes $\mathbb{C} \subset \mathcal{Y}$ with $|\mathbb{C}| = K$, a source domain $\Upsilon_s \subset \mathcal{X}$ and a target domain $\Upsilon_t \subset \mathcal{X}$, a $K$-way FSQS task $\mathbb{T}^{\text{FSQS}}_{\mathbb{S},\mathbb{Q}}$ is defined with a small *source* support set of labeled instances $\mathbb{S} = \{(\boldsymbol{x}_i^s, y_i^s) \in \Upsilon_s \times \mathbb{C}\}_{i=1\ldots|\mathbb{S}|}$ and a *target* query set $\mathbb{Q} = \{\boldsymbol{x}_i^q \in \Upsilon_t\}_{i=1\ldots|\mathbb{Q}|}$. Note that this paradigm provides an additional challenge compared to classical Few-Shot Classification tasks since at test time, the model is expected to generalize to both new classes and new domains while the support set and query set are sampled from different distributions. This paradigm is illustrated in Figure 3.2.

Evaluation and episodic training. We follow the same logic to expand the definition of the sets of $K$-way Few-Shot Classification tasks to FSQS tasks $\mathcal{E}_{\text{test}}(K)$.

$$\begin{aligned}\mathcal{E}_{\text{test}}(K) = \{\mathbb{T}^{\text{FSQS}}_{\mathbb{S},\mathbb{Q}} \mid & \mathbb{S} = \{(\boldsymbol{x}^s_i, y^s_i) \in \Upsilon_s \times \mathbb{C}\}_{i=1\ldots|\mathbb{S}|},\\ & \mathbb{Q} = \{\boldsymbol{x}^q_i \in \Upsilon_t\}_{i=1\ldots|\mathbb{Q}|},\\ & \mathbb{S} \cap \mathbb{Q} = \emptyset,\\ & \mathbb{C} \subset \mathcal{C}_{\text{test}}, \text{ and } |\mathcal{C}_{\text{test}}| = K,\\ & \{\Upsilon_s, \Upsilon_t\} \in \Delta_{\text{test}}, \text{ and } p_{\Upsilon_s} \neq p_{\Upsilon_t}\}\end{aligned}$$

In this chapter, we use the episodic training strategy described in Section 2.2.1, which means that at train time, we use $\mathcal{D}_{\text{base}}$ to sample tasks that are meant to replicate the tasks from $\mathcal{E}_{\text{test}}(K)$ and train the model to minimize its loss on each task's query set based on the information from the task's support set. Therefore, we define $\mathcal{E}_{\text{base}}(K)$ similarly to $\mathcal{E}_{\text{test}}(K)$:

$$\begin{aligned}\mathcal{E}_{\text{base}}(K) = \{\mathbb{T}^{\text{FSQS}}_{\mathbb{S},\mathbb{Q}} \mid & \mathbb{S} = \{(\boldsymbol{x}^s_i, y^s_i) \in \Upsilon_s \times \mathbb{C}\}_{i=1\ldots|\mathbb{S}|},\\ & \mathbb{Q} = \{\boldsymbol{x}^q_i \in \Upsilon_t\}_{i=1\ldots|\mathbb{Q}|},\\ & \mathbb{S} \cap \mathbb{Q} = \emptyset,\\ & \mathbb{C} \subset \mathcal{C}_{\text{base}}, \text{ and } |\mathcal{C}_{\text{base}}| = K,\\ & \{\Upsilon_s, \Upsilon_t\} \in \Delta_{\text{base}}, \text{ and } p_{\Upsilon_s} \neq p_{\Upsilon_t}\}\end{aligned}$$

Note that since base and test domains are disjoint, we ensure that the model will be tested to adapt to domain shifts that were not seen during training.

### 3.2.2 Positioning FSQS among Support-Query problems

To highlight FSQS's novelty, our discussion revolves around the problem of inferring on a given *Query Set* provided with the knowledge of a *Support Set*. We refer to this class of problems as *SQ problems*. Intrinsically, FSL falls into the category of SQ problems. Interestingly, *Unsupervised Domain Adaptation* (Pan and Q. Yang 2009) (UDA), defined as labeling a dataset sampled from a target domain based on labeled data sampled from a source domain (see Section 2.4.1), is also an SQ problem. Indeed, in this case, the source domain plays the role of support, while the target domain plays the query's role. Notably, an essential line of study in UDA leverages the target data distribution for aligning source and target domains, reflecting the importance of transduction in a context of adaptation (Ben-David et al. 2007; Ganin and Lempitsky 2015) *i.e.*, performing prediction by considering all target samples together. Transductive algorithms also have a special place in FSL (Dhillon et al. 2020; Y. Liu et al. 2019; Ren et al. 2019) and show that leveraging a query set as a whole brings a significant boost in performance (see Section 2.2.3). Nevertheless, UDA and FSL exhibit fundamental differences. UDA addresses the problem of distribution shift using important source data and target data (typically thousands of instances) to align distributions. In contrast, FSL focuses on the difficulty of learning from a few samples. To this purpose, we frame UDA as both an SQ problem with *large* transductivity and Support / Query Shift, while Few-Shot Learning is an SQ problem, eventually with *small* transductivity for transductive FSL. Thus, FSQS

| | Support-Query Problems | Train-Time | | | | Test-Time | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Support | | Query | | Support | | Query | New classes | New domains |
| | | Size | Labels | Size | Labels | Size | Labels | Transductivity | | |
| No SQS | Few-Shot Learning (FSL)[1] | Few | ✓ | Few | ✓ | Few | ✓ | Point-wise | ✓ | ✗ |
| | Transductive FSL[2] | Few | ✓ | Few | ✓ | Few | ✓ | Small | ✓ | ✗ |
| | Cross-Domain FSL[3] | Few | ✓ | Few | ✓ | Few | ✓ | Point-wise | ✓ | ✓ |
| SQS | Unsupervised Domain Adaptation[4] | | | | | Large | ✓ | Large | | |
| | Test-Time Adaptation[5] | Large | ✓ | | | | | Small | | ✓ |
| | Adaptive Risk Minimization[6] | Large | ✓ | Few | ✓ | | | Small | | ✓ |
| | Inductive FSQS | Few | ✓ | Few | ✓ | Few | ✓ | Point-wise | ✓ | ✓ |
| | Transductive FSQS | Few | ✓ | Few | ✓ | Few | ✓ | Small | ✓ | ✓ |

Table 3.1: An overview of SQ problems. We divide SQ problems into two categories, presence or not of **S**upport-**Q**uery shift; **No SQS** *vs* **SQS**. We consider three classes of transductivity: point-wise transductivity that is equivalent to inductive inference, small transductivity when inference is performed at batch level (typically in D. Wang et al. 2021; M. Zhang et al. 2021), and large transductivity when inference is performed at dataset level (typically in UDA). New classes (resp. new domains) describe if the model is evaluated at test-time on novel classes (resp. novel domains). Note that we frame UDA as a fully test-time algorithm. Notably, Cross-Domain FSL (CDFSL) (W.-Y. Chen et al. 2019) assumes that the support set and query set are drawn from the same distribution, thus No SQS.

combines both challenges: distribution shift and small transductivity. This new perspective allows us to establish fruitful connections with related learning paradigms, presented in Table 3.1.

## 3.3 FewShiftBed: A PyTorch testbed for FSQS

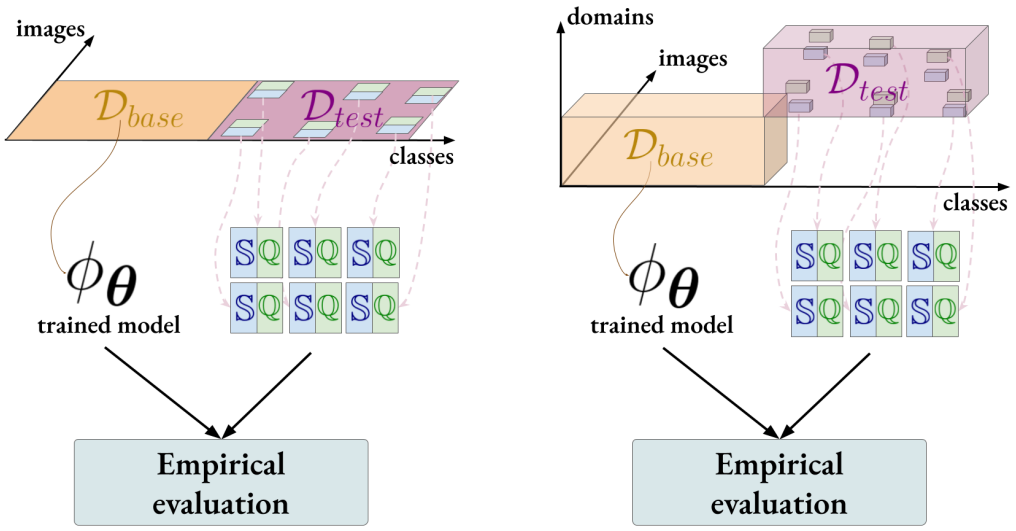
### 3.3.1 Datasets

We designed three new image classification datasets adapted to the FSQS problem. These datasets have two specificities.

1. They are dividable into groups of images, assuming that each group corresponds to a distinct domain. A key challenge is that each group must contain enough images with a sufficient variety of class labels so that it is possible to sample FSQS episodes.

2. They are delivered with a train/val/test split ($\mathcal{D}_{\text{base}}, \mathcal{D}_{\text{val}}, \mathcal{D}_{\text{test}}$), along both the class and the domain axis. This split is performed following the principles detailed in Section 3.2. Therefore, these datasets provide true few-shot tasks at test time, in the sense that the model will not have seen any instances of test classes and domains during training. Note that since

[1] Finn et al. 2017; Snell et al. 2017.
[2] Y. Liu et al. 2019; Ren et al. 2019.
[3] W.-Y. Chen et al. 2019.
[4] Pan and Q. Yang 2009; Quionero-Candela et al. 2009.
[5] Schneider et al. 2020; Y. Sun et al. 2020; D. Wang et al. 2021.
[6] M. Zhang et al. 2021.




(a) Benchmarking standard FSL (b) Benchmarking FSQS

Figure 3.3: Comparison between benchmarks for standard Few-Shot Learning (FSL) and for Few-Shot Learning under Support-Query Shift (FSQS). In standard FSL, we ensure that base and test classes are disjoint. At test time, to sample an FSL task, we sample a subset of classes, and from these classes, we sample some images for the support set and some others for the query set. FSQS adds a third dimension which represents the variety of domains. We must ensure that both the sets of classes and domains for base and test sets are disjoint. At test time, to sample an FSQS class, we also sample a subset of classes, but also a source and target domain. From these classes, we sample some images from the source domain for the support set and some images from the target domain for the query set. For simplicity, we only represented the split between base and test sets, not including a potential validation set.

we split along two axes, some data may be discarded (for instance images from a domain in $\Delta_{\text{base}}$ with a label in $\mathcal{C}_{\text{test}}$). Therefore it is crucial to find a split that minimizes this loss of data.

The differences between a benchmark for standard Few-Shot Learning and for Few-Shot Learning under Support-Query Shift are highlighted in Figure 3.3.

META-CIFAR100-CORRUPTED (MC100-C). CIFAR-100 (Krizhevsky, Hinton, et al. 2009) is a dataset of 60k three-channel square images of size $32 \times 32$, evenly distributed in 100 classes. Classes are evenly distributed in 20 superclasses. We use the same method used to build CIFAR-100-C (Hendrycks and Dietterich 2019), which makes use of 19 image perturbations, each one being applied with 5 different levels of intensity, to evaluate the robustness of a model to domain shift. We modify their protocol to adapt it to the FSQS problem: (i) we split the classes with respect to

Figure 3.4: Common perturbations from Hendrycks and Dietterich 2019 applied to a drill image. These perturbations are used in our benchmarks MC100-C and mIN-C to simulate a large variety of domains.

the superclass structure, and assign 13 superclasses (65 classes) to the training set, 2 superclasses (10 classes) to the validation set, and 5 superclasses (25 classes) to the testing set; (ii) we also split image perturbations (acting as domains), following the split of M. Zhang et al. 2021. We obtain 2,184k transformed images for training, 114k for validation, and 330k for testing.

*mini*ImageNet-Corrupted (mIN-C). *mini*ImageNet (Vinyals et al. 2016) is a popular benchmark for few-shot image classification. It contains 60k images from 100 classes from the ImageNet dataset. 64 classes are assigned to the training set, 16 to the validation set, and 20 to the test set. Like MC100-C, we build mIN-C using the image perturbations proposed by Hendrycks and Dietterich 2019 to simulate different domains. We use the original split from Vinyals et al. 2016 for classes, and use the same domain split as for MC100-C. Although the original *mini*ImageNet uses $84 \times 84$ images, we use $224 \times 224$ images. This allows us to re-use the perturbation parameters calibrated in Hendrycks and Dietterich 2019 for ImageNet. Finally, we discard the 5 most time-consuming perturbations. We obtain a total of 1.2M transformed images for training, 182k for validation, and 228k for testing. The detailed split for mIN-C, as well as MC100-C, are available in the documentation of our code repository[7].

FEMNIST-FewShot (FEMNIST-FS). EMNIST (Cohen et al. 2017) is a dataset of images of handwritten digits and uppercase and lowercase characters. Federated-EMNIST (Caldas et al. 2018) is a version of EMNIST where images are sorted by writer (or user). FEMNIST-FS consists of a split of the FEMNIST dataset adapted to few-shot classification. We separate both users and classes between training, validation, and test sets. We build each group as the set of images written by one user. The detailed split is available in the code. Note that in FEMNIST, many users provide several instances for each digit, but less than two instances for most letters. Therefore it is hard to find enough samples from a user to build a support set or a query set. As a result, our experiments are limited to classification tasks with only one sample per class in both the support and query sets.

[7] https://github.com/ebennequin/meta-domain-shift

### 3.3.2 Protocol

To prevent the pitfall of misinterpreting a performance boost, we draw three recommendations to isolate the causes of improvement rigorously.

- **How important is episodic training?** Despite its wide adoption in meta-learning for FSL, in some situations episodic training does not perform better than more naive approaches (W.-Y. Chen et al. 2019). Therefore we recommend reporting both the result obtained using episodic training and standard Empirical Risk Minimization (see the documentation of our code repository).

- **How does the algorithm behave in the absence of Support-Query Shift?** In order to assess that an algorithm designed for distribution shift does not provide degraded performance in an ordinary concept, and to provide a top-performing baseline, we recommend reporting the model's performance when we do not observe, at test-time, a support-query shift. Note that it is equivalent to evaluating the performance in cross-domain generalization, as firstly described in W.-Y. Chen et al. 2019.

- **Is the algorithm transductive?** The assumption of transductivity has been responsible for several improvements in FSL (Bronskill et al. 2020; Ren et al. 2019) while it has been demonstrated in Bronskill et al. 2020 that MAML (Finn et al. 2017) benefits strongly from the Transductive Batch-Normalization (TBN). Thus, we recommend specifying if the method is transductive and adapting the choice of the batch-normalization accordingly (Conventional Batch Normalization (Ioffe and Szegedy 2015) and Transductive Batch Normalization for inductive and transductive methods, respectively) since transductive batch normalization brings a significant boost in performance (Bronskill et al. 2020).

## 3.4 Transported Prototypes: A baseline for FSQS

### 3.4.1 Overall idea

We present a novel method that brings UDA to FSQS. As aforementioned, FSQS presents new challenges since we no longer assume that we sample the support set and the query set from the same distribution. As a result, it is unlikely that the support set and query sets share the same representation space region (non-overlap). In particular, the Euclidean distance, adopted in the celebrated Prototypical Network (Snell et al. 2017), may not be relevant for measuring similarity between query and support instances, as presented in Figure 3.1. To overcome this issue, we develop a two-phase approach that combines Optimal Transport (Transportation Phase) and the celebrated Prototypical Network (Prototype Phase). We give some background about Optimal Transport (OT) in Section 3.4.2 and the whole procedure is presented in Algorithm 1.

### 3.4.2 Background on Optimal Transport

Definition. We provide some basics about Optimal Transport (OT). A thorough presentation of OT is available at Peyré et al. 2019. Let $p_s$ and $p_t$ be two distributions on $\mathcal{X}$, we note $\Pi(p_s, p_t)$

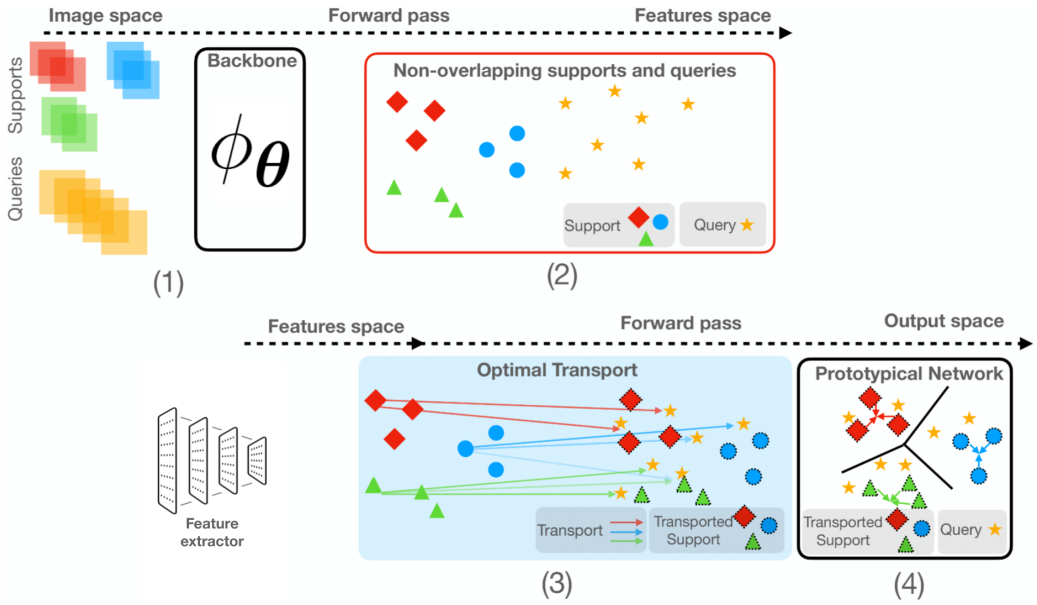



Figure 3.5: Overview of *Transported Prototypes*. (1) A support set and a query set are fed to a trained backbone that embeds images into a feature space. (2) Due to the shift between distributions, support, and query instances are embedded in non-overlapping areas. (3) We compute the Optimal Transport from support instances to query instances to build the transported support set. Note that we represent the transport plan only for one instance per class to preserve clarity in the schema. (4) Provided with the transported support, we apply the Prototypical Network (Snell et al. 2017) *i.e.*, $L^2$ similarity between transported support and query instances.

the set of joint probability with marginal $p_s$ and $p_t$ *i.e.*, $\forall \pi \in \Pi(p_s, p_t), \forall \boldsymbol{x} \in \mathcal{X}, \pi(\cdot, \boldsymbol{x}) = p_s$ and $\pi(\boldsymbol{x}, \cdot) = p_t$. The *Optimal Transport*, associated to cost $c$, between $p_s$ and $p_t$ is defined as:

$$W_c(p_s, p_t) := \min_{\pi \in \Pi(p_s, p_t)} \mathbb{E}_{(\boldsymbol{x}_s, \boldsymbol{x}_t) \sim \pi}[c(\boldsymbol{x}_s, \boldsymbol{x}_t)] \tag{3.1}$$

with $c(\cdot, \cdot)$ any metric. We note $\pi^\star(p_s, p_t)$ the joint distribution that achieves the minimum in equation 3.1. It is named the *transportation plan* from $p_s$ to $p_t$. When there is no confusion, we simply note $\pi^\star$. For our applications, we will use as metric the Euclidean distance in the representation space $\mathcal{Z}$ obtained from a representation $\phi_{\boldsymbol{\theta}}$ *i.e.*, $c_{\boldsymbol{\theta}}(\boldsymbol{x}_s, \boldsymbol{x}_t) := \|\phi_{\boldsymbol{\theta}}(\boldsymbol{x}_s) - \phi_{\boldsymbol{\theta}}(\boldsymbol{x}_t)\|_2$.

Discrete OT. When $p_s$ and $p_t$ are only accessible through a finite set of samples, respectively $(\boldsymbol{x}_1^s, ..., \boldsymbol{x}_{|\mathbb{S}|}^s)$ and $(\boldsymbol{x}_1^t, ..., \boldsymbol{x}_{|\mathbb{Q}|}^t)$ we introduce the empirical distributions

$$\hat{p}_s := \sum_{i=1}^{|\mathbb{S}|} w_i^s \delta_{\boldsymbol{x}_i^s}$$
$$\hat{p}_t := \sum_{j=1}^{|\mathbb{Q}|} w_j^t \delta_{\boldsymbol{x}_j^t}$$

**Algorithm 1** Transported Prototypes. Blue lines highlight the OT's contribution in the computational graph of an episode compared to the standard Prototypical Network (Snell et al. 2017).

**Input:** Support set $\mathbb{S} := (\boldsymbol{x}_i^s, y_i^s)_{1 \leq i \leq |\mathbb{S}|}$, query set $\mathbb{Q} := (\boldsymbol{x}_j^q, y_j^q)_{1 \leq j \leq |\mathbb{Q}|}$, classes $\mathbb{C}$, backbone $\phi_{\boldsymbol{\theta}}$.

**Output:** Loss $\mathcal{L}(\boldsymbol{\theta})$ for a randomly sampled episode.

$\boldsymbol{z}_i^s,\ \boldsymbol{z}_j^q \leftarrow \phi_{\boldsymbol{\theta}}(\boldsymbol{x}_i^s),\ \phi_{\boldsymbol{\theta}}(\boldsymbol{x}_j^q),\ \forall i, j$ ▷ Get representations.

$\boldsymbol{C}_{\boldsymbol{\theta}}(i,j) \leftarrow \left\| \boldsymbol{z}_i^s - \boldsymbol{z}_j^q \right\|^2,\ \forall i, j$ ▷ Cost-matrix.

$\boldsymbol{\pi}_{\theta}^{\star} \leftarrow$ Solve Equation 3.2 ▷ Transportation plan.

$\hat{\boldsymbol{\pi}}_{\theta}^{\star}(i,j) \leftarrow \boldsymbol{\pi}_{\theta}^{\star}(i,j) / \sum_j \boldsymbol{\pi}_{\theta}^{\star}(i,j),\ \forall i, j$ ▷ Normalization.

$\hat{\mathbb{S}}^{\mathcal{Z}} = (\hat{\boldsymbol{z}}_i^s)_i \leftarrow$ Given by Equation 3.4 ▷ Get transported support set.

$\hat{\boldsymbol{\mu}}_k \leftarrow \frac{1}{|\hat{\mathbb{S}}_k^{\mathcal{Z}}|} \sum_{\hat{\boldsymbol{z}}^s \in \hat{\mathbb{S}}_k^{\mathcal{Z}}} \hat{\boldsymbol{z}}^s$, for $k \in \mathbb{C}$. ▷ Get transported prototypes.

$\boldsymbol{p}_j^q \leftarrow$ From Equation 3.6, $\forall j \leq |\mathbb{Q}|$

**Return:** $\mathcal{L}(\boldsymbol{\theta}) := -\frac{1}{|\mathbb{Q}|} \sum_{i=1}^{|\mathbb{Q}|} \sum_{k=1}^{K} y_{ik}^q \log(p_{ik}^q)$.

where $w_i^s$ (resp. $w_j^t$) is the mass probability put in sample $\boldsymbol{x}_i^s$ (resp. $\boldsymbol{x}_j^t$) *i.e.*, $\sum_{i=1}^{|\mathbb{S}|} w_i^s = 1$ (resp. $\sum_{j=1}^{|\mathbb{Q}|} w_j^t = 1$) and $\delta_{\boldsymbol{x}}$ is the Dirac distribution in $\boldsymbol{x}$. The discrete version of the OT is derived by introducing the set of couplings

$$\boldsymbol{\Pi}(p_s, p_t) := \left\{ \boldsymbol{\pi} \in \mathbb{R}^{|\mathbb{S}| \times |\mathbb{Q}|} | \boldsymbol{\pi} \mathbf{1}_{|\mathbb{S}|} = \mathbf{p}_s, \boldsymbol{\pi}^\top \mathbf{1}_{|\mathbb{Q}|} = \mathbf{p}_t \right\}$$

where $\mathbf{p}_s := (w_1^s, \cdots, w_{|\mathbb{S}|}^s)$, $\mathbf{p}_t := (w_1^t, \cdots, w_{|\mathbb{Q}|}^t)$, and $\mathbf{1}_{|\mathbb{S}|}$ (respectively $\mathbf{1}_{|\mathbb{Q}|}$) is the unit vector with dimension $|\mathbb{S}|$ (respectively $|\mathbb{Q}|$). The discrete transportation plan $\boldsymbol{\pi}_{\boldsymbol{\theta}}^{\star}$ is then defined as:

$$\boldsymbol{\pi}_{\boldsymbol{\theta}}^{\star} := \underset{\boldsymbol{\pi} \in \boldsymbol{\Pi}(p_s, p_t)}{\operatorname{argmin}} \langle \boldsymbol{\pi}, \boldsymbol{C}_{\boldsymbol{\theta}} \rangle_F \tag{3.2}$$

where $\boldsymbol{C}_{\boldsymbol{\theta}}(i,j) := c_{\boldsymbol{\theta}}(\boldsymbol{x}_i^s, \boldsymbol{x}_j^t)$ and $\langle \cdot, \cdot \rangle_F$ is the Frobenius dot product. Note that $\boldsymbol{\pi}_{\boldsymbol{\theta}}^{\star}$ depends on both $p_s$ and $p_t$, and $\boldsymbol{\theta}$ since $\boldsymbol{C}_{\boldsymbol{\theta}}$ depends on $\boldsymbol{\theta}$. In practice, we use Entropic regularization (Cuturi 2013) that makes OT easier to solve by promoting a smoother transportation plan with a computationally efficient algorithm, based on Sinkhorn-Knopp's scaling matrix approach (Knight 2008). It is defined as

$$\boldsymbol{\pi}_{\boldsymbol{\theta}}^{\star}(\hat{p}_s, \hat{p}_t) := \arg\min_{\boldsymbol{\pi} \in \boldsymbol{\Pi}} \langle \boldsymbol{\pi}, \boldsymbol{C}_{\boldsymbol{\theta}} \rangle_F + \varepsilon \mathcal{H}(\boldsymbol{\pi}) \tag{3.3}$$

with $\varepsilon > 0$ and $\mathcal{H}(\boldsymbol{\pi}) = \sum_{i,j=1}^{|\mathbb{S}|,|\mathbb{Q}|} \boldsymbol{\pi}(i,j) \log \boldsymbol{\pi}(i,j)$ is the negative entropy. In our experiments, we set $\varepsilon = 0.05$.

### 3.4.3 Method

Transportation Phase. At each episode, we are provided with a source support set $\mathbb{S}$ and a target query set $\mathbb{Q}$. We work on top of extracted features *i.e.*, $\boldsymbol{z} = \phi_{\boldsymbol{\theta}}(\boldsymbol{x})$. We note $\mathbb{S}^{\mathcal{Z}} \subset \mathcal{Z}$

(resp. $\mathbb{Q}^{\mathcal{Z}} \subset \mathcal{Z}$) the set of representations extracted from instances in $\mathbb{S}$ (resp. $\mathbb{Q}$). As these two sets are sampled from different distributions, $\mathbb{S}^{\mathcal{Z}}$ and $\mathbb{Q}^{\mathcal{Z}}$ are likely to lie in different regions of the representation space $\mathcal{Z}$. In order to adapt the source support set $\mathbb{S}$ to the target domain $\Upsilon_t$, which is only represented by the target query set $\mathbb{Q}$, we follow Courty, Flamary, Tuia, et al. 2016 to compute $\hat{\mathbb{S}}^{\mathcal{Z}}$ the *barycenter mapping* of $\mathbb{S}^{\mathcal{Z}}$, that we refer to as the *transported support set*, defined as follows:

$$\hat{\mathbb{S}}^{\mathcal{Z}} := \hat{\boldsymbol{\pi}}^{\star}_{\boldsymbol{\theta}} \mathbb{Q}^{\mathcal{Z}} \tag{3.4}$$

where $\boldsymbol{\pi}^{\star}_{\boldsymbol{\theta}}$ is the transportation plan from $\mathbb{S}^{\mathcal{Z}}$ to $\mathbb{Q}^{\mathcal{Z}}$ and $\hat{\boldsymbol{\pi}}^{\star}_{\boldsymbol{\theta}}(i,j) := \boldsymbol{\pi}^{\star}_{\boldsymbol{\theta}}(i,j) / \sum_{j=1}^{|\mathbb{Q}|} \boldsymbol{\pi}^{\star}_{\boldsymbol{\theta}}(i,j)$. The *transported* support set $\hat{\mathbb{S}}^{\mathcal{Z}}$ estimates labeled examples in the target domain using labeled examples in the source domain. The success relies on the fact that transportation preserves labels, *i.e.*, a query instance close to $\hat{\boldsymbol{z}}^s \in \hat{\mathbb{S}}^{\mathcal{Z}}$ should share the same label with $\boldsymbol{x}^s$, where $\hat{z}^s$ is the barycenter mapping of $\boldsymbol{z}^s = \phi_{\boldsymbol{\theta}}(\boldsymbol{x}^s) \in \mathbb{S}^{\mathcal{Z}}$. See step (3) of Figure 3.5 for a visualization of the transportation phase.

Prototype Phase. For each class $k \in \mathbb{C}$, we compute the *transported prototypes*

$$\hat{\boldsymbol{\mu}}_k := \frac{1}{|\hat{\mathbb{S}}^{\mathcal{Z}}_k|} \sum_{\hat{\boldsymbol{z}}^s \in \hat{\mathbb{S}}^{\mathcal{Z}}_k} \hat{\boldsymbol{z}}^s \tag{3.5}$$

(where $\hat{\mathbb{S}}^{\mathcal{Z}}_k = \{(\hat{\boldsymbol{z}}, y) \in \hat{\mathbb{S}}^{\mathcal{Z}} \mid y = k\}$ is the subset of the transported support set composed of instances with label $k \in \mathbb{C}$ and $\mathbb{C}$ are classes of the current task). We classify each query $\boldsymbol{x}^q_j$ with representation $\boldsymbol{z}^q_j = \phi_{\boldsymbol{\theta}}(\boldsymbol{x}^q_j)$ using its euclidean distance to each transported prototypes:

$$p^q_{jk} = \mathbb{P}(y^q_j = k | \boldsymbol{x}^q_j) \propto \exp\left( -\left\| z^q_j - \hat{\boldsymbol{\mu}}_k \right\|^2_2 \right) \tag{3.6}$$

Crucially, the standard Prototypical Networks (Snell et al. 2017) computes Euclidean distance to each prototype while we compute the Euclidean to each *transported* prototype, as presented in step (4) of Figure 3.5. Note that our formulation involves the query set in the computation of $(\hat{\boldsymbol{\mu}}_k)_{k \in \mathbb{C}}$, which means that our method is *transductive*.

Genericity of OT. FewShiftBed implements OT as a stand-alone module that can be easily plugged into any FSL algorithm. We report additional baselines in Appendix 7.3 where other FSL algorithms are equipped with OT. This technical choice reflects our insight that OT may be ubiquitous for addressing FSQS and makes its usage in the testbed straightforward.

## 3.5 Experiments

[8] Snell et al. 2017.
[9] Vinyals et al. 2016.
[10] Y. Liu et al. 2019.
[11] Dhillon et al. 2020.

| Strategy | Meta-CIFAR100-C | | miniImageNet-C | | FEMNIST-FS |
|---|---|---|---|---|---|
| | **1-shot** | **5-shot** | **1-shot** | **5-shot** | **1-shot** |
| ProtoNet[8] | $30.02 \pm 0.40$ | $42.77 \pm 0.47$ | $36.37 \pm 0.50$ | $47.58 \pm 0.57$ | $84.31 \pm 0.73$ |
| MatchingNet[9] | $30.71 \pm 0.38$ | $41.15 \pm 0.45$ | $35.26 \pm 0.50$ | $44.75 \pm 0.55$ | $84.25 \pm 0.71$ |
| TransPropNet†[10] | $\mathbf{34.15 \pm 0.39}$ | $47.39 \pm 0.42$ | $24.10 \pm 0.27$ | $27.24 \pm 0.33$ | $86.42 \pm 0.76$ |
| FTNet†[11] | $28.91 \pm 0.37$ | $37.28 \pm 0.40$ | $39.02 \pm 0.46$ | $51.27 \pm 0.45$ | $86.13 \pm 0.71$ |
| TP† (ours) | $\mathbf{34.00 \pm 0.46}$ | $\mathbf{49.71 \pm 0.47}$ | $\mathbf{40.49 \pm 0.54}$ | $\mathbf{59.85 \pm 0.49}$ | $\mathbf{93.63 \pm 0.63}$ |
| TP w/o OT † | $32.47 \pm 0.41$ | $48.00 \pm 0.44$ | $40.43 \pm 0.49$ | $53.71 \pm 0.50$ | $90.36 \pm 0.58$ |
| TP w/o TBN † | $33.74 \pm 0.46$ | $49.18 \pm 0.49$ | $37.32 \pm 0.55$ | $55.16 \pm 0.54$ | $92.31 \pm 0.73$ |
| TP w. OT-TT † | $32.81 \pm 0.46$ | $48.62 \pm 0.48$ | $\mathbf{44.77 \pm 0.57}$ | $\mathbf{60.46 \pm 0.49}$ | $\mathbf{94.92 \pm 0.55}$ |
| TP w/o ET † | $\mathbf{35.94 \pm 0.45}$ | $48.66 \pm 0.46$ | $42.46 \pm 0.53$ | $54.67 \pm 0.48$ | $94.22 \pm 0.70$ |
| *TP w/o SQS* † | *85.67 ± 0.26* | *88.52 ± 0.17* | *64.27 ± 0.39* | *75.22 ± 0.30* | *99.72 ± 0.07* |

Table 3.2: Top-1 accuracy of few-shot learning models in various datasets and numbers of shots with 8 instances per class in the query set (except for FEMNIST-FS: 1 instance per class in the query set), with 95% confidence intervals. The top half of the table is a comparison between existing few-shot learning methods and Transported Prototypes (TP). The bottom half is an ablation study of TP. OT denotes Optimal Transport, TBN is Transductive Batch-Normalization, OT-TT refers to the setting where Optimal Transport is applied at test time but not during episodic training, and ET means episodic training *i.e.*, w/o ET refers to the setting where training is performed through standard Empirical Risk Minimization. TP w/o SQS reports the model's performance in the absence of support-query shift. † flags if the method is transductive. For each setting, the best accuracy among existing methods is shown in bold, as well as the accuracy of an ablation if it improves TP.

We compare the performance of baseline algorithms with *Transported Prototypes* on various datasets and settings. We also offer an ablation study in order to isolate the source to the success of *Transported Prototypes*. Extensive results are detailed in Appendix 7.3. Instructions to reproduce these results can be found in the code's documentation.

Setting and details. We conduct experiments on all methods and datasets implemented in FewShiftBed. We use a standard 4-layer convolutional network for our experiments on Meta-CIFAR100-C and FEMNIST-FewShot, and a ResNet18 for our experiments on miniImageNet. Transductive methods are equipped with a Transductive Batch-Normalization. All episodic training runs contain 40k episodes, after which we retrieve the model state with the best validation accuracy. We run each individual experiment on three different random seeds. All results presented in this paper are the average accuracies obtained with these random seeds.

Analysis. The top half of Table 3.2 reveals that Transported Prototypes (TP) outperform all baselines by a strong margin on all datasets and settings. Importantly, baselines perform poorly on FSQS, demonstrating they are not equipped to address this challenging problem, stressing our study's significance. It is also interesting to note that the performance of transductive approaches, which is significantly better in a standard FSL setting (Dhillon et al. 2020; Y. Liu et al. 2019), is here similar to inductive methods (notably, TransPropNet (Y. Liu et al. 2019) fails loudly without

| | | Meta-CIFAR100-C | | miniImageNet-C | | FEMNIST-FS |
|---|---|---|---|---|---|---|
| | Training | **1-shot** | **5-shot** | **1-shot** | **5-shot** | **1-shot** |
| TP | Standard ERM | $36.17 \pm 0.47$ | $50.45 \pm 0.47$ | $45.41 \pm 0.54$ | $57.82 \pm 0.48$ | $93.60 \pm 0.68$ |
| MAP | Standard ERM | $35.96 \pm 0.44$ | $49.55 \pm 0.45$ | $43.51 \pm 0.47$ | $56.10 \pm 0.43$ | $92.86 \pm 0.67$ |
| TP | Episodic | $32.13 \pm 0.45$ | $46.19 \pm 0.47$ | $45.77 \pm 0.58$ | $59.91 \pm 0.48$ | $94.92 \pm 0.56$ |
| MAP | Episodic | $32.38 \pm 0.41$ | $45.96 \pm 0.43$ | $43.81 \pm 0.47$ | $57.70 \pm 0.43$ | $87.15 \pm 0.66$ |

Table 3.3: Top-1 accuracy with 8 instances per class in the query set when applying Transported Prototypes and MAP with or without episodic training. Transported Prototypes perform equally or better than MAP (Hu et al. 2021). Here TP includes power transform in the feature space.

Transductive Batch-Normalization showing that propagating label with non-overlapping support/query can have a dramatic impact, see Appendix 7.3). Thus, FSQS deserves a fresher look to be solved. Transported Prototypes mitigate a significant part of the performance drop caused by support-query shift while benefiting from the simplicity of combining a popular FSL method with a time-tested UDA method. This gives us strong hopes for future works in this direction.

Ablation study. Transported Prototypes (TP) combine three components: Optimal Transport (OT), Transductive Batch-Normalization (TBN), and episode training (ET). Which of these components are responsible for the observed gain? Following recommendations from Section 3.3.2, we ablate those components in the bottom half of Table 3.2. We observe that both OT and TBN individually improve the performance of ProtoNet for FSQS and that the best results are obtained when the two of them are combined. Importantly, OT without TBN performs better than TBN without OT (except for 1-shot mIN-C), demonstrating the superiority of OT compared to TBN for aligning distributions in the few samples regime. Note that the use of TaskNorm (Bronskill et al. 2020) is beyond the scope of the paper[12]; we encourage future work to dig into that direction and we refer the reader to the very recent work Du et al. 2021. We observe that there is no clear evidence that using OT at train time is better than simply applying it at test time on a ProtoNet trained without OT. Additionally, the value of Episodic Training (ET) compared to standard Empirical Risk Minimization (ERM) is not obvious. For instance, simply training with ERM and applying TP at test time is better than adding ET on 1-shot MC100-C, 1-shot mIN-C, and FEMNIST-FS, making it another element to add to the study from Laenen and Bertinetto 2021 who put into question the value of ET. Understanding why and when we should use ET or only OT at test time is interesting for future works. Additionally, we compare TP with MAP (Hu et al. 2021) which implements an OT-based approach for transductive FSL. Their approach includes a power transform to reduce the skew in the distribution, so for fair comparison, we also implemented it into Transported Prototypes for these experiments[13]. We also used the OT module only at test time and compared with two backbones, respectively trained with ET and ERM. Interestingly, our experiments in Table 3.3 show that MAP is able to handle SQS. Finally, in order to evaluate the performance drop related to Support-Query Shift compared to a setting with support and query instances sampled from the same distribution, we test Transported Prototypes

[12]These normalizations are implemented in FewShiftBed for future works.
[13]Therefore results in Table 3.3 differ from results in Table 3.2.

on few-shot classification tasks without SQS (TP w/o SQS in Table 3.2), making a setup equivalent to CDFSL. Note that in both cases, the model is trained in an episodic fashion on tasks presenting a Support-Query Shift. These results show that SQS presents a significantly harder challenge than CDFSL, while there is considerable room for improvement.

## 3.6 Conclusion

We release FewShiftBed, a testbed for the under-investigated and crucial problem of Few-Shot Learning when the support and query sets are sampled from related but different distributions, named FSQS. FewShiftBed includes three datasets, relevant baselines, and a protocol for reproducible research. Inspired by the recent progress of Optimal Transport (OT) to address Unsupervised Domain Adaptation, we propose a method that efficiently combines OT with the celebrated Prototypical Network (Snell et al. 2017). Following the protocol of FewShiftBed, we bring compelling experiments demonstrating the advantage of our proposal compared to transductive counterparts. We also isolate factors responsible for improvements. Our findings suggest that Batch-Normalization is ubiquitous, as described in related works Bronskill et al. 2020; Du et al. 2021, while episodic training, even if promising on paper, is questionable.

Perspectives. As a lead for future works, FewShiftBed could be improved by using different datasets to model different domains, instead of using artificial transformations. Since we are talking about domain adaptation, we also encourage the study of accuracy as a function of the size of the target domain, *i.e.*, the size of the query set. Moving beyond the transductive algorithm, as well as understanding when meta-learning brings a clear advantage to address FSQS remains an open and exciting problem. FewShiftBed brings the first step towards its progress.

Where are they now? Following the original publication of this work in 2021, several works addressed our proposed FSQS problem. Du et al. 2021 predict relevant batch statistics by applying a multi-layer perceptron at each batch normalization. Our call for moving beyond transductive methods was heard by Aimen et al. 2023 who propose to use adversarial projections to solve *inductive* Support-Query Shift. Finally, S. Jiang et al. 2022 improve on our proposed Transported Prototypes approach to make it more robust to small perturbations in images.

# 4 Contribution 2: Transductive Few-Shot Open-Set Recognition

This chapter is an aggregation of two closely linked works, delivered in equal contribution with Malik Boudiaf from the LIVIA laboratory:

1. *Model-Agnostic Few-Shot Open-Set Recognition*, by Malik Boudiaf, Etienne Bennequin, Myriam Tami, Celine Hudelot, Antoine Toubhans, Pablo Piantanida, and Ismail Ben Ayed (Boudiaf, Bennequin, Tami, Hudelot, et al. 2022), made available as an arxiv Preprint;

2. *Open-Set Likelihood Maximization for Few-Shot Learning*, by Malik Boudiaf, Etienne Bennequin, Myriam Tami, Celine Hudelot, Antoine Toubhans, Pablo Piantanida, and Ismail Ben Ayed (Boudiaf, Bennequin, Tami, Toubhans, et al. 2023), published in CVPR 2023.

These two works are motivated by common observations on the necessity of an open-set transductive method for Few-Shot Learning and the desirable features of such a method. However, each one proposed a different solution to the problem.

## 4.1 Introduction

Most few-shot methods listed in Section 2.2.2 classify the unlabeled query samples of a given task based on their similarity to the support instances in the feature space. 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, as described in Section 2.2.1. However, the real world is open and this closed-set assumption may not hold in practice, especially for limited support sets. In fact, two of the real use cases described in Section 1.2 (*i.e.*, e-shopping and bacteria recognition) involve *open-set instances i.e.*, query instances which belong to none of the support classes. A closed-set classifier will falsely label these open-set instances 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, as described in Section 2.4.2. In large-scale settings, the literature abounds of methods designed specifically to detect open-set instances while maintaining good accuracy on closed-set instances (Bendale and Boult 2016; Scheirer et al. 2012; Zhou et al. 2021). However, recent studies of the Few-Shot Open-Set Recognition (FSOSR) setting (S. Huang et al. 2022; Jeong et al. 2021; B. Liu et al. 2020) expose it to be a difficult task in the inductive setting. Indeed, their reported results, which we reproduced and expose in Table 4.2, indicate that sophisticated inductive methods specialized for FSOSR do not show any improvement with respect to simple baselines such as $k$ nearest neighbors (Ramaswamy et al. 2000). In Section 4.2, we explore potential causes for the specific difficulty of FSOSR.

To help alleviate the scarcity of labeled data, transduction (Vapnik 2013) was recently explored for few-shot classification (Y. Liu et al. 2019), and has since become a prominent research direction, fueling a large body of works described in Section 2.2.3. In this chapter, we seek to explore transduction for the FSOSR setting. We argue that theoretically, transduction has the potential to enable both classification and outlier detection (OD) modules to act symbiotically. Indeed, the classification module can reveal valuable structure of the inlier's marginal distribution that the OD module seeks to estimate, such as the number of modes or conditional distributions, while the OD part indicates the "usability" of each unlabelled sample. However, transductive principles currently adopted for few-shot learning heavily rely on the closed-set assumption in the unlabelled data, leading them to match the classification confidence for open-set instances with that of closed-set instances. In the presence of outliers, this not only harms their predictive performance on closed-set instances, but also makes prediction-based outlier detection substantially harder than with simple inductive baselines.

In this chapter, we propose two simple yet powerful methods to reconcile transduction with the open nature of the FSOSR problem. Both methods are fully model-agnostic, in the sense that they can be applied on top of any pre-trained model seamlessly. We argue that this is an important feature for a few-shot method, because 1) such a method can be seamlessly integrated into an existing pipeline without the need to re-train the model; 2) it can scale up to the latest and most significant advances in representation learning (*e.g.*, ViTs or self-supervised learning) without any additional effort; and 3) the difficult reproduction of episodic training has been shown to be an obstacle to the fair comparison between methods (Antoniou, Edwards, et al. 2019; Bennequin 2019), while this is not an issue when comparing methods using the same trained parameters.

Open-Set Transductive Information Maximization (OSTIM). Our diagnostic of the InfoMax principle at the core of the TIM method (Boudiaf, Ziko, et al. 2020) indicates that this state-of-the-art transductive method tends to enforce confident predictions for all samples, regardless of whether they are closed-set or open-set. We, therefore, propose a modification to the original method using an additional *outlier prototypes*. This additional prototype allows predictions to be confident towards a new implicitly defined *outlier class*. We name this simple method Open-Set Transductive Information Maximization (OSTIM).

Open-Set Likelihood Optimization (OSLO). 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.

Empirically, we show that both methods significantly surpass their 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 their parameters, OSTIM and OSLO's improvements over a strong baseline remain large and consistent. This modularity allows our methods to fully benefit from the latest advances in standard image recognition.

**CHAPTER'S CONTRIBUTIONS**

1. We expose the specific difficulty of the FSOSR problem when using off-the-shelf pre-trained models, on a wide range of benchmarks and architectures, using our novel Mean Imposture Factor metric which measures how much the classes' distributions in a dataset are perturbed by instances from other classes.
2. 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.
3. We introduce Open-Set Transductive Information Maximization (OSTIM), an intuitive modification of the TIM method that provides an additional prototype for outliers.
4. We introduce Open-Set Likelihood Optimization (OSLO), a principled extension of the Maximum Likelihood framework that explicitly models and handles the presence of outliers. OSTIM and OSLO are interpretable and modular *i.e.*, can be applied on top of any pre-trained model seamlessly.
5. Through extensive experiments spanning five datasets and a dozen of pre-trained models, we show that our methods consistently surpass both inductive and existing transductive methods in detecting open-set instances while competing with the strongest transductive methods in classifying closed-set instances. Our empirical studies include long-overdue experiments on the performance of transductive methods with various sizes and shapes of the query set.

## 4.2 Few-Shot Open-Set Recognition

**SETUP AND FORMALIZATION.** The standard Few-Shot Classification task described in Section 2.2.1 is under the *closed-set* assumption, which means that the unknown ground-truth query labels $\{y_i^q\}_{i=1...|\mathbb{Q}|}$ are assumed to be restricted to closed-set classes *i.e.*, $\forall i,\ y_i^q \in \mathbb{C}_{\text{CS}}$ with $\mathbb{C}_{\text{CS}} = \mathbb{C}$ is the set of classes represented in the support set $\mathbb{S}$. In FSOSR, however, query labels may also belong to an additional set $\mathbb{C}_{\text{OS}}$ of *open-set* classes *i.e.*, $\forall i,\ y_i^q \in \mathbb{C}_{\text{CS}} \cup \mathbb{C}_{\text{OS}}$ with $\mathbb{C}_{\text{CS}} \cap \mathbb{C}_{\text{OS}} = \emptyset$. For easy referencing, we refer to query samples from the closed-set classes $\mathbb{C}_{\text{CS}}$ as *inliers* and to query samples from open-set classes $\mathbb{C}_{\text{OS}}$ as *outliers*. For each query image $\boldsymbol{x}_i^q$, the goal of FSOSR is to simultaneously assign a closed-set prediction $p_{ik}^q = \mathbb{P}(y_i^q = k|\boldsymbol{x}_i^q),\ k \in \mathbb{C}_{\text{CS}}$ and an *outlierness* score $\mathbb{P}(y_i^q \notin \mathbb{C}_{\text{CS}}|\boldsymbol{x}_i^q)$.

**MEASURING THE DIFFICULTY OF OUTLIER DETECTION ON NOVEL CLASSES.** As an anomaly detection problem, open-set recognition consists in detecting samples that differ from the popula-

Figure 4.1: 2-dimensional reduction with T-SNE of feature extracted from ImageNet's validation set using a ResNet12 trained on *mini*ImageNet. (Left): images from 20 randomly selected classes represented in *mini*ImageNet's base set. (Right): Images from the 20 classes represented in *mini*ImageNet's test set. Each color corresponds to a distinct class.

tion that is known by the classification model. However, in FSOSR, neither closed-set classes nor open-set classes have been seen during the training of the feature extractor *i.e.*,

$$\mathcal{C}_{\text{base}} \cap \mathbb{C}_{\text{CS}} = \mathcal{C}_{\text{base}} \cap \mathbb{C}_{\text{OS}} = \emptyset$$

In that sense, both the inliers and the outliers of our problem can be considered outliers from the perspective of the feature extractor. Intuitively, this makes it harder to detect open-set instances, since the model doesn't know well the distribution from which they are supposed to diverge. Here we empirically demonstrate and quantify the difficulty of OSR in a setting where closed-set classes have not been represented in the training set. Specifically, we estimate the gap in terms of the quality of the classes' definition in the feature space, between classes that were represented during the training of the feature extractor *i.e.*, $\mathcal{C}_{\text{base}}$, and the classes of the test set, which were not represented in the training set. To do so, we introduce the novel Mean Imposture Factor measure and use the intra-class to inter-class variance ratio $\rho$ as a complementary measure. Note that the following study is performed on whole datasets, *not* few-shot tasks.

Mean Imposture Factor (MIF). Let $\mathcal{D}_{\phi_{\boldsymbol{\theta}}} \subset \mathcal{Z} \times \mathcal{C}$ be a labeled dataset of extracted feature vectors, with $\phi_{\boldsymbol{\theta}}$ a fixed feature extractor and $\mathcal{C}$ a finite set of classes. For any feature vector $\boldsymbol{z}$ and a class $k$ to which $\boldsymbol{z}$ does not belong, we define the Imposture Factor $IF_{\boldsymbol{z}|k}$ as the proportion of the instances of class $k$ in $\mathcal{D}_{\phi_{\boldsymbol{\theta}}}$ that are further than $\boldsymbol{z}$ from their class centroid $\boldsymbol{\mu}_k$. Then the MIF is the average IF over all instances in $\mathcal{D}_{\phi_{\boldsymbol{\theta}}}$.

$$\boxed{MIF = \frac{1}{|\mathcal{C}|} \sum_{k} \frac{1}{|\mathcal{D}_{\phi_{\boldsymbol{\theta}}} \setminus \mathcal{D}_k|} \sum_{\boldsymbol{z} \notin \mathcal{D}_k} IF_{\boldsymbol{z}|k}} \text{ with } IF_{\boldsymbol{z}|k} = \frac{1}{|\mathcal{D}_k|} \sum_{\boldsymbol{z}' \in \mathcal{D}_k} \mathbf{1}_{\|\boldsymbol{z}' - \boldsymbol{\mu}_k\|_2 > \|\boldsymbol{z} - \boldsymbol{\mu}_k\|_2} \tag{4.1}$$

with $\mathcal{D}_k$ the set of instances in $\mathcal{D}_{\phi_{\boldsymbol{\theta}}}$ with label $k$, and $\mathbf{1}$ the indicator function. The MIF is a measure of how perturbed the clusters corresponding to the ground truth classes are. A MIF

Table 4.1: Contrast between datasets made of images from classes represented (*base*) or not represented (*test*) in the feature extractor's training set, on three benchmarks and with several backbones (RN12: ResNet12, WRN: WideResNet1810, ViT, RN50: ResNet50, and MX: MLP-Mixer), following the MIF (in percents) and the variance ratio ($\rho$). Best result for each column is shown in bold.

| Classes | miniImageNet | | | | tieredImageNet | | | | ImageNet → Aircraft | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\rho$ | | MIF (%) | | $\rho$ | | MIF (%) | | $\rho$ | | | MIF (%) | | |
| | RN12 | WRN | RN12 | WRN | RN12 | WRN | RN12 | WRN | ViT | RN50 | MX | ViT | RN50 | MX |
| *base* | **0.93** | **0.84** | **0.89** | **1.03** | **1.09** | **0.78** | **0.78** | **0.81** | **0.96** | **1.36** | **2.54** | **0.09** | **0.29** | **0.31** |
| *test* | 2.10 | 2.07 | 5.56 | 7.36 | 2.10 | 1.54 | 4.39 | 5.18 | 3.20 | 4.88 | 5.35 | 18.08 | 21.58 | 17.27 |

of zero means that all instances are closer to their class centroid than any outsider. Note that $\text{MIF} = 1 - \text{AUROC}(\psi)$ where $\text{AUROC}(\psi)$ is the area under the ROC curve for an outlier detector $\psi$ that would assign to each instance an outlier score equal to the distance to the ground truth class centroid. To the best of our knowledge, the MIF is the first tool allowing the measure of the class-wise integrity of a projection in the feature space. As a sanity check for MIF, we also report the intra-class to inter-class variance ratio $\rho$, used in the previous work Goldblum, Reich, et al. 2020, to measure the compactness of a clustering solution.

Base classes are better defined than test classes. We experiment on three widely used Few-Shot Learning benchmarks: *mini*ImageNet (Vinyals et al. 2016), *tiered*ImageNet (Ren et al. 2019), and ImageNet → Aircraft (Maji et al. 2013). We use the validation set of ImageNet in order to obtain novel instances for ImageNet, *mini*ImageNet, and *tiered*ImageNet's base classes. We also use it for test classes for consistency. In Figure 4.1, we present a visualization of the ability of a ResNet12 trained on *mini*ImageNet to project images of both base and test classes into clusters. While we are able to obtain well-separated clusters for base classes after the 2-dimensional T-SNE reduction, this is clearly not the case for test classes, which are more scattered and overlapping. Such results are quantitatively corroborated by Table 4.1, which shows that both MIF and $\rho$ are systematically lower for base classes across 3 benchmarks and 5 feature extractors. This demonstrates the difficulty of defining in the feature space the distribution of a class that was not seen during the training of the feature extractor, and therefore the difficulty of defining clear boundaries between inliers and outliers *i.e.*, closed-set images and open-set images, all the more when only a few samples are available.

## 4.3 Open-Set Transductive Information Maximization

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. In practice, transductive methods have achieved impressive improvements over inductive methods in standard closed-set FSC (Boudiaf, Ziko, et al. 2020; Dhillon et al. 2020; Hu et al. 2021; Ziko et al. 2020). Considering the difficulty posed by the FSOSR problem, detailed in Section 4.2, we expect that transductive methods can help us improve outlier detection while still achieving super-inductive closed-set

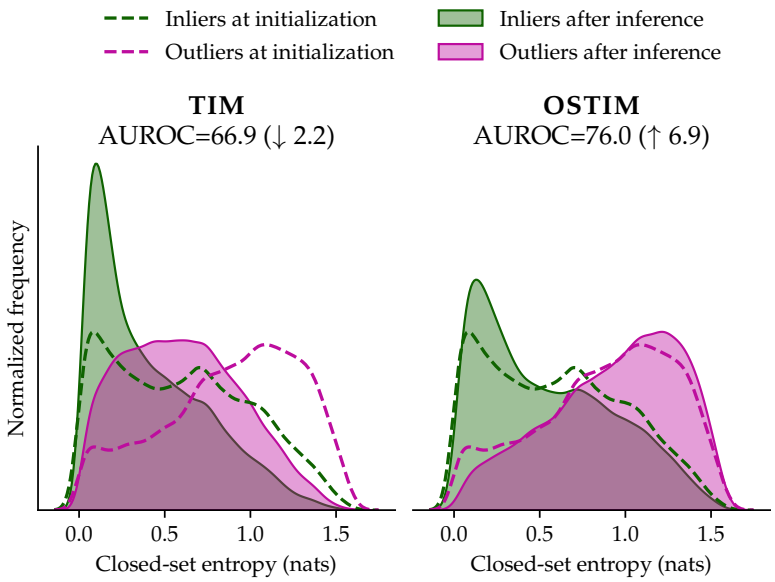



Figure 4.2: **Closed v.s. Open-Set InfoMax.** (Left) Minimizing closed-set entropy (Boudiaf, Ziko, et al. 2020) on all samples degrades prediction-based outlier detection. (Right) OSTIM tends to bin outliers in a $(K+1)^{th}$ category. Therefore, their *open-set* conditional entropy in Eq. (4.4) decreases while their closed-set entropy increases.

predictive performance. Unfortunately, we empirically show in Sec. 4.5 that significant accuracy gains systematically come along with significant outlier detection degradation.

DIAGNOSING THE INFOMAX TRANSDUCTION. Among the 5 transductive methods evaluated, we find TIM (Boudiaf, Ziko, et al. 2020) offers the best trade-off between performances in closed-set classification and outlier detection, although the latter still falls far below inductive alternatives. Indeed, as part of the InfoMax principle, TIM (Boudiaf, Ziko, et al. 2020) systematically enforces confident predictions on each query sample through conditional entropy minimization, whether this sample is an outlier or not. To make things worse, outliers' initial predictions typically fall in the region of the simplex where the magnitude of entropy's gradients is the highest (Veilleux et al. 2021), meaning the model prioritizes minimizing the entropy of outliers over inliers. Altogether, those ingredients lead to a degradation of the discriminability between inliers' and outliers' predictions during inference. This situation is depicted in the left plot of Fig. 4.2, where we observe the entropy histogram of outliers (purple) shifting significantly towards the left (low-entropy) after inference. Following these observations, we seek to instantiate the InfoMax principle in a way that simultaneously benefits closed-set predictive performance and outlier detection.

INTRODUCING OSTIM. To help remediate the issue, we propose a simple yet highly effective modification to the original closed-set TIM (Boudiaf, Ziko, et al. 2020) method, that retains TIM's high closed-set accuracy while drastically improving outlier detection. Importantly, it does not introduce any computational overhead or tunable hyperparameter. Relaxing the closed-set

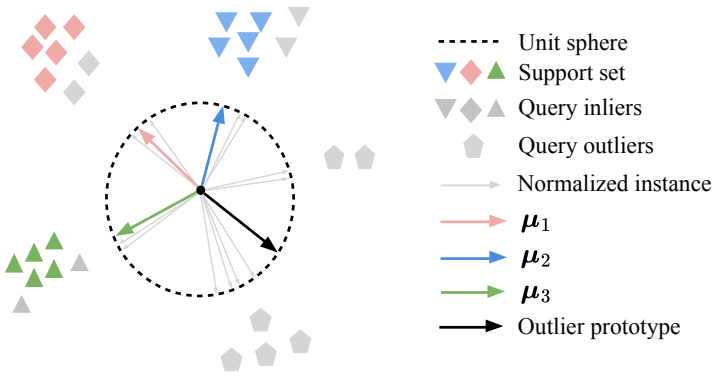



Figure 4.3: **Geometric intuition behind OSTIM**. The CE term encourages colored arrows to align with support samples, while $\mathcal{I}_\alpha$ encourages grey arrows to either align with a colored arrow (inlier prototypes) or with the black arrow (outlier prototype).

assumption, we consider a $K+1$-way classification problem, where the added class represents the broad *outlier* category. We observe in Sec. 4.5.2 that introducing additional learnable parameters, e.g. a new prototype as in Zhou et al. 2021 to represent the *outlier* class in such low-data regimes yields poor performances. Consequently, we propose an implicit definition of the *outlier* class that reuses existing parameters and remains differentiable. We name our method OSTIM for *Open Set Transductive Information Maximization*.

IMPLICIT OUTLIER PROTOTYPE. As part of the model-agnostic setting, we abstract the base model and work directly on top of extracted features $\boldsymbol{z} = \phi_{\boldsymbol{\theta}}(\boldsymbol{x})$ with $\boldsymbol{\theta}$ considered frozen. Additionally, we center-normalize all features and prototypes such that all operations are performed on the unit sphere. This choice of center-normalization is developed in Section 4.5. We define the similarity $l_{ik}$ between a sample $\boldsymbol{z}_i$ and a class prototype $\boldsymbol{\mu}_k$ as their dot product:

$$l_{ik} = \langle \boldsymbol{z}_i,\ \boldsymbol{\mu}_k \rangle \tag{4.2}$$

We now define the *outlier* logit as the negative average of inliers class logits (with $K$ the number of classes):

$$l_{i,K+1} = -\frac{1}{K}\sum_{k=1}^{K} l_{ik} = \langle \boldsymbol{z}_i,\ \underbrace{-\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{\mu}_k}_{\text{implicit } \textit{outlier} \text{ prototype}} \rangle. \tag{4.3}$$

The outlier logit can be interpreted as the similarity between some point and an *outlier* prototype corresponding to the diametrical opposite of the average of inlier prototypes. To clarify this intuition, a geometrical description of the problem is provided in Figure 4.3. For a center-normalized query point represented by a gray arrow, inlier logits $\{l_k\}_{k=1}^{K}$ correspond to measuring the an-

gles between respective colored arrows and the gray arrow, while $l_{K+1}$ measures the similarity with the black arrow. The concatenation of *inlier* logits and the *outlier* logit forms the final logit vector $\boldsymbol{l}_i = [l_{i1}, \dots, l_{i,K+1}]^T$, which is translated into a probability vector $\boldsymbol{p}_i$ over the $K+1$ outcomes through a standard softmax operation. The first $K$ components of this probability vector $\{p_{ik}\}_{k=1}^{K}$ are used for closed-set classification, while the last $p_{i,K+1}$ is used as the outlierness score.

PROTOTYPE REFINEMENT. The prototypes $\{\boldsymbol{\mu}_k\}_{k=1}^{K}$ are initialized as the class-centroids using labeled samples from $\mathbb{S}$, and further refined by minimizing an open-set version of TIM's transductive loss:

$$\boxed{\min_{\boldsymbol{\mu}} \text{CE} - \widehat{\mathcal{I}}_\alpha} \text{ with } \text{CE} := -\frac{1}{|\mathbb{S}|}\sum_{i=1}^{|\mathbb{S}|}\sum_{k=1}^{K+1} y_{ik}^s \log(p_{ik}^s),$$

$$-\widehat{\mathcal{I}}_\alpha := \underbrace{\sum_{k=1}^{K+1} \widehat{p}_k \log \widehat{p}_k}_{\substack{\text{marginal entropy} \\ \text{prevents trivial solutions}}} \underbrace{- \frac{\alpha}{|\mathbb{Q}|}\sum_{i=1}^{|\mathbb{Q}|}\sum_{k=1}^{K+1} p_{ik}^q \log(p_{ik}^q)}_{\substack{\text{conditional entropy} \\ \text{forces query samples into} \\ \text{inlier category or outlier group}}}, \tag{4.4}$$

where $y_{ik}^s = \mathbb{1}[y_i^s = k], k \in [1, K]$ is a one-hot encoded version of the ground-truth label, complemented with a last outlier component $y_{i,K+1}^s = 0$, and $\hat{p}_k = \frac{1}{|\mathbb{Q}|}\sum_i p_{ik}^q$ denotes the marginal prediction for class $k$. Following Boudiaf, Ziko, et al. 2020, $\alpha \in \mathbb{R}$ is found through validation. Note that unlike in the standard TIM, the introduction of an additional $(K+1)^{th}$ class, represented by the implicit prototype, makes it possible to minimize the entropy of all samples without losing discriminability between inliers and outliers. In other words, outliers can simply be predicted in the $(K+1)^{th}$ category with high confidence, and inliers in their associated inlier category. As a matter of fact, Fig. 4.2 shows that inliers' closed-set entropy decreases, indicating that they tend to get closer to some inlier prototype, while outliers' closed-set entropy increases, indicating that they are on average moving away from their closest inlier prototype. We emphasize that closed-set entropy is simply used for diagnosis, and neither corresponds to the outlierness score used by TIM (Boudiaf, Ziko, et al. 2020) nor OSTIM in Sec. 4.5.

## 4.4 OPEN-SET LIKELIHOOD

In Section 4.3, we introduced OSTIM *i.e.*, a first method to address Few-Shot Open-Set Recognition in a transductive fashion, by leveraging an implicit outlier prototype to mitigate the closed-set bias of standard transductive methods. In this section, we go one step further we introduce OSLO, a novel extension of the standard likelihood designed for transductive FSOSR. Unlike existing transductive methods including OSTIM, OSLO explicitly models and handles the potential presence of outliers, which allows it to outperform inductive baselines on both aspects of the open-set scenario.

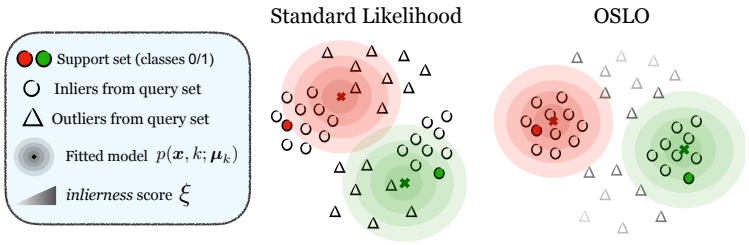



Figure 4.4: **Intuition behind OSLO.** Standard transductive likelihood (left) tries to enforce high likelihood for all points, including outliers. OSLO (right) instead treats the *outlierness* of each sample as a latent variable to be solved alongside the parametric model. Besides yielding a principled *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.

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 $\boldsymbol{\zeta}_i \in \Delta^K$ describing the membership of sample $i$ to each closed-set class[1], with $\Delta^K = \{\boldsymbol{\zeta} \in [0,1]^K : \boldsymbol{\zeta}^T \mathbf{1} = 1\}$ 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 $\boldsymbol{\zeta} = \{\boldsymbol{\zeta}_i\}_{i=1}^{|\mathbb{S}|+|\mathbb{Q}|}$ 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 $\boldsymbol{\zeta}_i = \boldsymbol{y}_i$ 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 $\boldsymbol{\mu} = \{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K\}$:

$$p(\boldsymbol{x}, k; \boldsymbol{\mu}) = p(k)p(\boldsymbol{x}|k) \propto \exp(-\frac{\|\phi_{\boldsymbol{\theta}}(\boldsymbol{x}) - \boldsymbol{\mu}_k\|^2}{2}) \tag{4.5}$$

As mentioned in section 4.2, the feature extractor's parameters $\boldsymbol{\theta}$ are kept frozen, and only $\boldsymbol{\mu}$ will be optimized.

[1] In the original paper, latent class assignments are noted $\boldsymbol{z}$. Here we changed the notation to $\boldsymbol{\zeta}$ to avoid any conflict with the notation of feature vectors.

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

$$p(\boldsymbol{X}, \boldsymbol{\zeta}; \boldsymbol{\mu}) = \prod_{i=1}^{|\mathbb{S}|+|\mathbb{Q}|} \prod_{k=1}^{K} p(\boldsymbol{x}_i, k; \boldsymbol{\mu})^{\zeta_{ik}} \tag{4.6}$$

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

$$\log(p(\boldsymbol{X}, \boldsymbol{\zeta}; \boldsymbol{\mu})) = \sum_{i=1}^{|\mathbb{S}|+|\mathbb{Q}|} \sum_{k=1}^{K} \zeta_{ik} \log(p(\boldsymbol{x}_i, k; \boldsymbol{\mu})) \tag{4.7}$$

Eq. (4.7) 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 4.4 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:

$$\mathcal{L}_{\text{O}}(\boldsymbol{X}, \boldsymbol{\zeta}, \boldsymbol{\xi}; \boldsymbol{\mu}) = \sum_{i=1}^{|\mathbb{S}|+|\mathbb{Q}|} \xi_i \sum_{k=1}^{K} \zeta_{ik} \log(p(\boldsymbol{x}_i, k; \boldsymbol{\mu})) \tag{4.8}$$

Eq (4.8) can be interpreted as follows: samples believed to be inliers *i.e.*, $\xi_i \approx 1$ will be required to have 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 $\boldsymbol{\zeta}$. Finally, to prevent overconfident latent scores, we consider a *penalty* term on both $\boldsymbol{\zeta}$ and $\boldsymbol{\xi}$:

$$\mathcal{L}_{\text{soft}} = \sum_{i=|\mathbb{S}|+1}^{|\mathbb{S}|+|\mathbb{Q}|} \lambda_z \mathcal{H}(\boldsymbol{\zeta}_i) + \lambda_\xi \mathcal{H}(\boldsymbol{\xi}_i) \tag{4.9}$$

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.

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

$$\begin{aligned} \max_{\boldsymbol{\mu}, \boldsymbol{\zeta}, \boldsymbol{\xi}} \quad & \mathcal{L}_{\text{O}}(\boldsymbol{\zeta}, \boldsymbol{\xi}, \boldsymbol{\mu}) + \mathcal{L}_{\text{soft}}(\boldsymbol{\zeta}, \boldsymbol{\xi}) \\ \text{s.t} \quad & \boldsymbol{\zeta}_i \in \Delta^K, \quad \xi_i \in [0, 1] \quad \forall\, i \\ & \boldsymbol{\zeta}_i = \boldsymbol{y}_i, \quad \xi_i = 1, \quad i \leq |\mathbb{S}| \end{aligned} \tag{4.10}$$

Problem (4.10) 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* (4.10) *can be minimized by alternating the following updates, with $\sigma$ denoting the sigmoid operation:*

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

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

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{1}{\sum_{i=1}^{|\mathbb{S}|+|\mathbb{Q}|} \xi_i^{(t+1)} \zeta_{ik}^{(t+1)}} \sum_{i=1}^{|\mathbb{S}|+|\mathbb{Q}|} \xi_i^{(t+1)} \zeta_{ik}^{(t+1)} \phi_\theta(\boldsymbol{x}_i)$$

*Proof.* We denote by $\nabla_\cdot(\mathcal{L}_\text{O} + \mathcal{L}_\text{soft})$ the partial derivative of OSLO's optimization problem. We calculate the updates of $\xi_i$ and $\zeta_{ik}$ for $i > |S|$, and of $\boldsymbol{\mu}_k$, by finding the annulation point of their partial derivative.

$$\begin{aligned} & \boxed{\nabla_{\xi_i}(\mathcal{L}_\text{O} + \mathcal{L}_\text{soft}) = 0} \\ \Leftrightarrow \quad & \sum_{k=1}^{K} \zeta_{ik} \log(p(\boldsymbol{x}_i, k; \boldsymbol{\mu})) = \lambda_\xi((\log \xi_i + 1) - (\log(1 - \xi_i) + 1)) \\ \Leftrightarrow \quad & \frac{1}{\lambda_\xi} \sum_{k=1}^{K} \zeta_{ik} \log(p(\boldsymbol{x}_i, k; \boldsymbol{\mu})) = \log\left(\frac{\xi_i}{1 - \xi_i}\right) \\ \Leftrightarrow \quad & \xi_i = \sigma\left(\frac{1}{\lambda_\xi} \sum_{k=1}^{K} \zeta_{ik} \log(p(\boldsymbol{x}_i, k; \boldsymbol{\mu}))\right) \end{aligned}$$

$$\begin{aligned} & \boxed{\nabla_{\zeta_{ik}}(\mathcal{L}_\text{O} + \mathcal{L}_\text{soft}) = 0} \\ \Leftrightarrow \quad & \xi_i \log(p(\boldsymbol{x}_i, k; \boldsymbol{\mu})) = \lambda_z(\log \zeta_{ik} + 1) \\ \Rightarrow \quad & \zeta_{ik} \propto \exp\left(\frac{\xi_i}{\lambda_z} \log(p(\boldsymbol{x}_i, k; \boldsymbol{\mu}))\right) \end{aligned}$$

$$\boxed{\nabla_{\boldsymbol{\mu}_k}(\mathcal{L}_{\text{O}} + \mathcal{L}_{\text{soft}}) = 0}$$
$$\Leftrightarrow \quad \sum_{i=1}^{|\mathbb{S}|+|\mathbb{Q}|} \xi_i \zeta_{ik} (\phi_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - \boldsymbol{\mu}_k) = 0$$
$$\Leftrightarrow \quad \boldsymbol{\mu}_k = \frac{\sum_{i=1}^{|\mathbb{S}|+|\mathbb{Q}|} \xi_i \zeta_{ik} \phi_{\boldsymbol{\theta}}(\boldsymbol{x}_i)}{\sum_{i=1}^{|\mathbb{S}|+|\mathbb{Q}|} \xi_i \zeta_{ik}}$$

□

The optimal solution for the *inlierness* score $\xi_i$ 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 $\xi_i$ allows to refine and improve the closed-set parametric model**. In particular, $\xi_i$ acts as a sample-wise temperature in the update of $\boldsymbol{\zeta}_i$, 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 $\xi_i$.

## 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 et al. 2000), (ii) Inductive Few-Shot classifiers, e.g. SimpleShot (Y. Wang, Chao, et al. 2019) (iii) Inductive Open-Set methods formed by standard methods such as OpenMax (Bendale and Boult 2016) and Few-Shot methods such as Snatcher (Jeong et al. 2021) (iv) Transductive classifiers, e.g. TIM (Boudiaf, Ziko, et al. 2020), that implicitly rely on the closed-set assumption, and finally (v) Transductive Open-Set introduced in this work through Ostim and Oslo. Following Jeong et al. 2021, closed-set few-shot classifiers are 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 $\psi_{\boldsymbol{\upsilon}} : \boldsymbol{x} \mapsto (\boldsymbol{x} - \boldsymbol{\upsilon})/||\boldsymbol{x} - \boldsymbol{\upsilon}||_2$ on the features extracted by $\phi_{\boldsymbol{\theta}}$ benefited all methods. Therefore, we apply it to the features before applying any method, using

- an inductive *Base centering* (Y. Wang, Chao, et al. 2019) for inductive methods $\boldsymbol{\upsilon}_{Base} = \frac{1}{|\mathcal{D}_{base}|} \sum_{\boldsymbol{x} \in \mathcal{D}_{base}} \phi_{\boldsymbol{\theta}}(x)$,
- and a transductive *Task centering* (Hu et al. 2021) $\boldsymbol{\upsilon}_{Task} = \frac{1}{|\mathbb{S} \cup \mathbb{Q}|} \sum_{\boldsymbol{x} \in \mathbb{S} \cup \mathbb{Q}} \phi_{\boldsymbol{\theta}}(x)$ 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 for Oslo, which we show in Appendix 1 remains a valid minimizer of Eq. (4.10) 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 4.6 are produced using the pre-trained ResNet-12 and Wide-ResNet 28-10 checkpoints provided by the authors from Ye et al. 2020. As for episodically-finetuned models required by Snatcher (Jeong et al. 2021) and FEAT (Ye et al. 2020), checkpoints are obtained from the authors' respective repositories. Finally, to challenge the model-agnosticity of our methods, 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. These models 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. 2016) dataset with 100 classes and the larger *tiered*-ImageNet (Ren et al. 2019) 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 et al. 2013) with 100 classes, and the Fungi classification challenge (Schroeder and Cui 2018) with 1394 classes. Following standard FSOSR protocol, support sets contain $|\mathbb{C}_{\text{CS}}| = 5$ 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 $|\mathbb{C}_{\text{OS}}| = 5$ open-set classes. We follow this setting for a fair comparison with previous works Jeong et al. 2021 B. Liu et al. 2020 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

Benchmarking the state of the art.

Simplest inductive methods are competitive. The first surprising result comes from analyzing the performances of standard OOD detectors on the FSOSR problem. Fig. 4.2 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 (Y. Zhao et al. 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, Fig. 4.5 shows that across 5 scenarios, the combination SimpleShot (Y. Wang, Chao, et al. 2019)+ $k$-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 4.5 and 4.6. Additional results for the Wide-ResNet architecture are provided in Appendix 3.

Table 4.2: **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.

| | | ***mini*-ImageNet** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Strategy | Method | 1-shot | | | | 5-shot | | | |
| | | Acc ±0.72 | AUROC ±0.79 | AUPR ±0.69 | Prec@0.9 ±0.47 | Acc ±0.44 | AUROC ±0.73 | AUPR ±0.61 | Prec@0.9 ±0.56 |
| OOD detection | *k*-NN (Ramaswamy et al. 2000) | - | 70.86 | 70.43 | 58.23 | - | 76.22 | 76.36 | 61.48 |
| | IForest (F. T. Liu et al. 2008) | - | 55.59 | 55.24 | 52.18 | - | 62.80 | 61.62 | 54.77 |
| | OCVSM (Schölkopf et al. 2001) | - | 69.67 | 69.71 | 57.35 | - | 68.49 | 65.60 | 59.24 |
| | PCA (Shyu et al. 2003) | - | 67.23 | 66.50 | 56.67 | - | 75.24 | 75.53 | 60.73 |
| | COPOD (Z. Li et al. 2020) | - | 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 (Y. Wang, Chao, et al. 2019) | 65.90 | 64.99 | 63.78 | 55.77 | 81.72 | 70.61 | 70.06 | 57.91 |
| | Baseline ++ (W.-Y. Chen et al. 2019) | 65.81 | 65.15 | 63.85 | 55.87 | 81.86 | 66.37 | 65.58 | 56.33 |
| | FEAT (Ye et al. 2020) | 67.23 | 52.45 | 54.44 | 50.00 | 82.00 | 53.25 | 56.48 | 50.00 |
| Inductive Open-Set | PEELER⋆ (B. Liu et al. 2020) | 65.86 | 60.57 | - | - | 80.61 | 67.35 | - | - |
| | TANE-G⋆ (S. 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 and 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 (J. Liu et al. 2020) | 69.35 | 57.95 | 58.58 | 52.71 | 82.66 | 61.27 | 62.17 | 53.26 |
| | TIM-GD (Boudiaf, Ziko, et al. 2020) | 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 (Y. Wang, C. Xu, et al. 2020) | 68.24 | 49.96 | 51.61 | 50.45 | 81.77 | 51.82 | 53.49 | 50.80 |
| Transductive Open-Set | OSTIM (ours) | 69.36 | **74.57** | **75.24** | 60.13 | 82.62 | **83.76** | **84.10** | **67.64** |
| | OSLO (ours) | **71.73** | **74.92** | **74.61** | **60.95** | **83.40** | 82.59 | 82.34 | 66.98 |
| | | ***tiered*-ImageNet** | | | | | | | |
| | | ±0.74 | ±0.76 | ±0.71 | ±0.52 | ±0.52 | ±0.68 | ±0.75 | ±0.57 |
| OOD detection | *k*-NN (Ramaswamy et al. 2000) | - | 73.97 | 73.15 | 60.74 | - | 80.22 | 80.06 | 65.47 |
| | IForest (F. T. Liu et al. 2008) | - | 54.57 | 54.24 | 51.85 | - | 62.31 | 60.82 | 54.72 |
| | OCVSM (Schölkopf et al. 2001) | - | 71.22 | 71.17 | 58.81 | - | 71.20 | 68.23 | 61.09 |
| | PCA (Shyu et al. 2003) | - | 68.30 | 67.02 | 57.66 | - | 76.26 | 76.41 | 61.81 |
| | COPOD (Z. Li et al. 2020) | - | 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 (Y. Wang, Chao, et al. 2019) | 70.27 | 69.78 | 67.89 | 58.54 | 84.94 | 77.38 | 76.28 | 63.21 |
| | Baseline ++ (W.-Y. Chen et al. 2019) | 70.21 | 69.73 | 67.80 | 58.50 | 85.10 | 73.77 | 72.39 | 61.05 |
| | FEAT (Ye et al. 2020) | 69.94 | 52.49 | 56.74 | 50.00 | 83.96 | 53.30 | 59.81 | 50.00 |
| Inductive Open-Set | PEELER⋆ (B. Liu et al. 2020) | 69.51 | 65.20 | - | - | 84.10 | 73.27 | - | - |
| | TANE-G⋆ (S. 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 and 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 (J. Liu et al. 2020) | 74.07 | 62.13 | 61.84 | 54.53 | 85.65 | 67.41 | 67.57 | 56.30 |
| | TIM-GD (Boudiaf, Ziko, et al. 2020) | 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 (Y. Wang, C. Xu, et al. 2020) | 73.80 | 49.32 | 51.41 | 50.35 | 85.21 | 51.65 | 53.85 | 50.79 |
| Transductive Open-Set | OSTIM (ours) | 73.77 | **78.86** | **78.99** | **63.96** | 85.73 | **87.62** | **87.95** | **72.81** |
| | OSLO (ours) | **76.64** | **79.06** | **79.07** | **64.36** | **86.35** | 86.92 | **87.28** | 71.98 |

TRANSDUCTIVE METHODS STILL IMPROVE ACCURACY BUT DEGRADE OUTLIER DETECTION. As shown in Table 4.2, 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 (Y. Chen et al. 2020; Killamsetty et al. 2021; Saito et al. 2021;

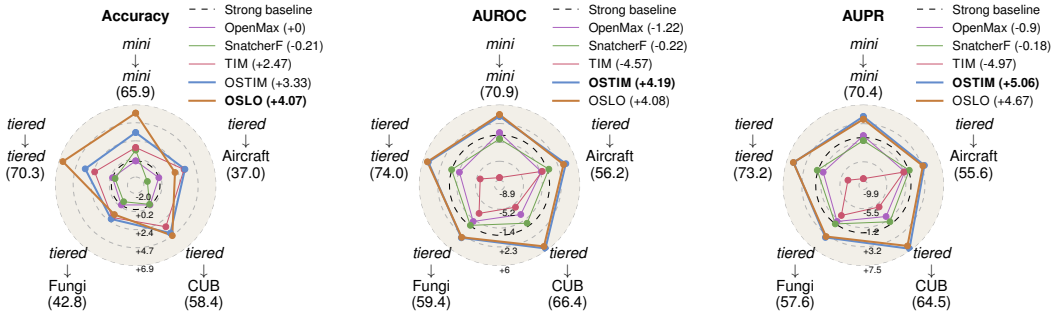



Figure 4.5: **Oslo improves open-set performances on a wide variety of tasks.** Relative 1-shot performance of the best methods of each family w.r.t the *Strong baseline* using a ResNet-12, across a set of 5 scenarios, including 3 with domain-shift. Each vertex represents one scenario, e.g. *tiered*→Fungi ($x$) 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 5-shot setting and using a WideResNet backbone.

Q. Yu et al. 2020). 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, 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.

### Outstanding performance of Transductive Open-Set methods.

**Ostim and Oslo above all.** Benchmark results in Fig. 4.2 show that both Oslo and Ostim propose a trade-off between closed-set accuracy and outlier detection performance which cannot be achieved with existing methods. Ostim competes and Oslo surpasses the best transductive methods in terms of closed-set accuracy, while both methods consistently outperform 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 Ostim and Oslo remain largely superior to their inductive competitors even in the 5-shot scenario, where a consistent 3-7% gap in AUROC and AUPR with the third-best method can be observed. We accumulate further evidence of Ostim and Oslo's superiority by introducing 3 additional cross-domain scenarios in Fig. 4.5, corresponding to a base model pre-trained on *tiered*-ImageNet, but tested on CUB, Aircraft, and Fungi datasets. In such challenging scenarios, where both feature and class distributions shift, both methods remain competitive in closed-set classification and maintain consistent improvements in outlier detection. They even widen the gap in the tiered CUB setting, achieving a strong AUPR improvement (more than 7%) over the Strong Baseline.

Accuracy or Auroc? Interestingly, while both methods present similar improvements to the state-of-the-art, they propose a different trade-off between closed-set accuracy and outlier detection. Ostim slightly outperforms Oslo in OOD metrics in the 5-shot scenarios but presents similar results in the 1-shot scenarios. The main difference between both methods resides in the closed-set performance: Oslo significantly outperforms Ostim in closed-set accuracy on both *mini* and *tiered*ImageNet in the 1 and 5-shot scenarios. However, Oslo's closed-set performance drops in the cross-domain scenarios shown in Figure 4.5, falling below Ostim's and even classical TIM's performance. These results indicate that the choice of the ideal method between Ostim and Oslo strongly depends on the specifics of the problem.

We step toward model-agnosticity. We evaluate our methods' *model-agnosticity* by their 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 the empirical demonstration of our methods' 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 4.6 show the relative improvement of Ostim and Oslo with respect to the strong baseline in the 1-shot scenario on the $* \rightarrow$ Fungi benchmark[2]. Without any tuning, both our methods remain able to leverage the strong expressive power of large-scale models, and even consistently widen the gap with the strong baseline, achieving remarkable performance with the ViT-B/16 trained in a supervised fashion. This set of results testifies to how easy obtaining highly competitive results on difficult specialized tasks can be by combining transductive open-set methods with the latest models.

Robustness to the shape of the query set.

The benefits of 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 4.7 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, Ostim and Oslo gain from additional queries but stays above the baseline even with a small number of queries. Interestingly enough, all closed-set transductive methods present a drop in outlier detection performance when the number of queries increases. Our proposed open-set transductive methods are the only transductive methods to improve their outlier detection ability when the number of queries increases.

Broad Open-Set. In the standard FSOSR setting (Jeong et al. 2021; B. Liu et al. 2020):

- support sets contain $|\mathbb{C}_{CS}| = 5$ closed-set classes with 1 or 5 instances, or *shots*, per class;

[2]Some experiments were re-run after the publication of the original paper and thus may marginally differ from the original results. They still systematically lie in the original confidence intervals.

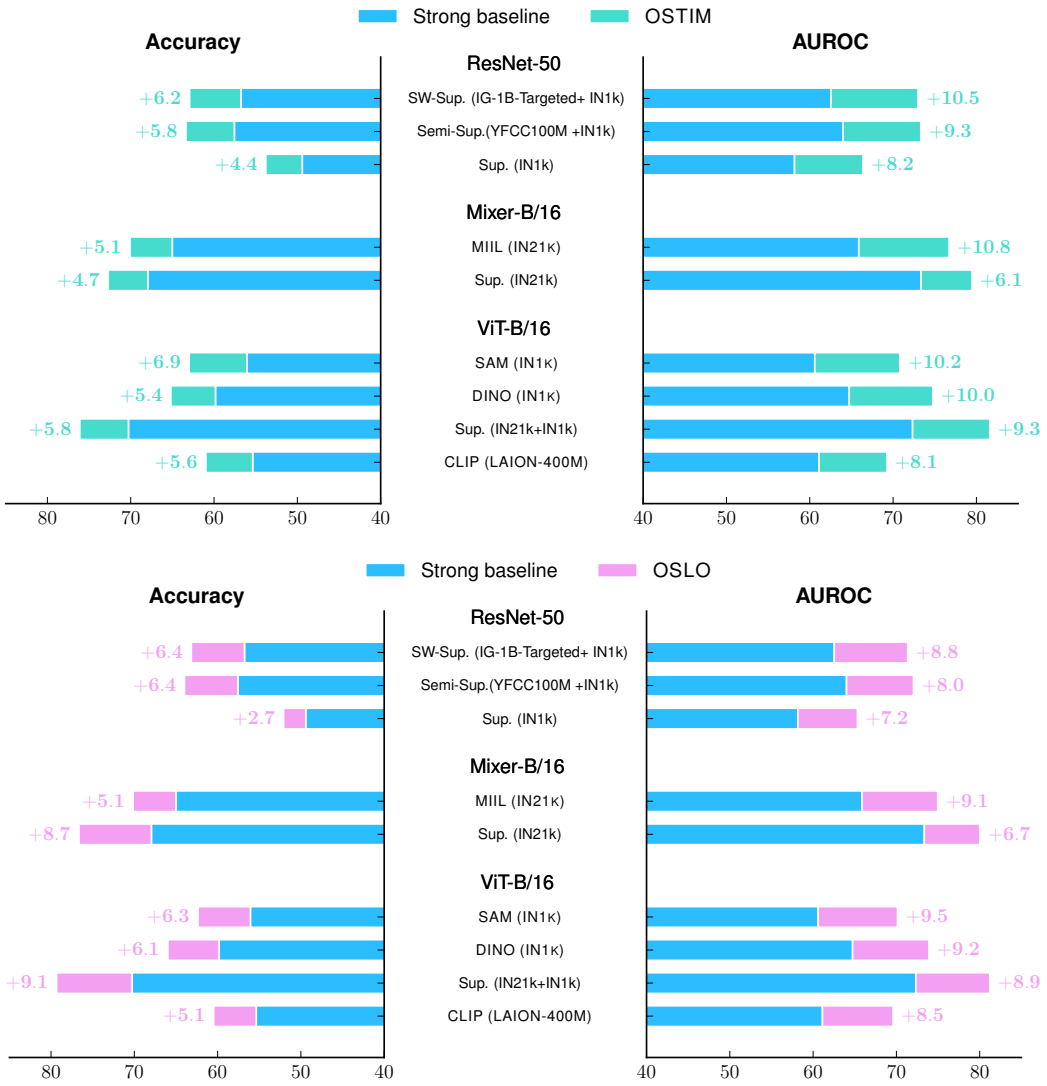



Figure 4.6: **Ostim and Oslo's improvement are consistent across many architectures and training strategies.** To evaluate model-agnosticity, we compare our methods to the Strong baseline on challenging 1-shot Fungi tasks. We experiment across 3 largely distinct architectures: ResNet-50 (CNN) (He et 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 (X. Chen, Hsieh, 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 4.5.

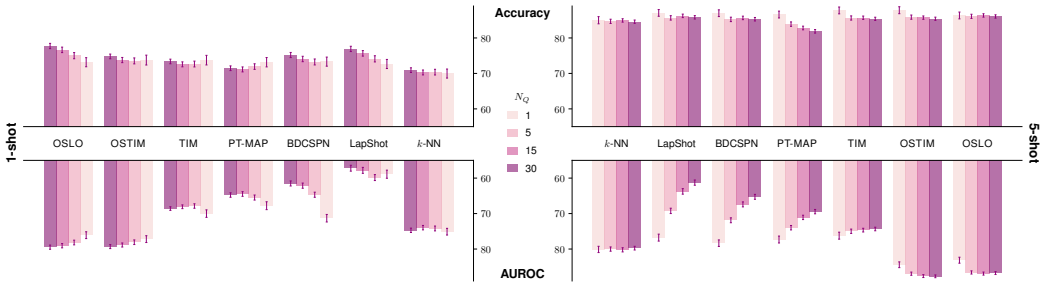



Figure 4.7: **Transductive Open-Set improves performance even with few queries.** We study the closed-set (accuracy) and open-set (AUROC) performance of transductive methods depending on the size of the query set on *tiered*-ImageNet in the 1-shot and 5-shot settings. The total size $|Q|$ of the query 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 Appendix 3.

- query sets are formed by sampling 15 instances per class, from a total of ten classes: the five closed-set classes $\mathbb{C}_{\text{CS}}$ and an additional set of $|\mathbb{C}_{\text{OS}}| = 5$ open-set classes.

This is a very strong assumption on the distribution of open-set samples. While this will not affect an inductive method, it is likely to impact the performance of both closed-set and open-set transductive methods. In this section, we provide additional results in a more realistic setting. In this new setting, the query set is formed by sampling 15 instances for each of the 5 closed-set classes, plus $5 \times 15 = 75$ open-set instances, which are sampled indifferently from all remaining classes in the test set. Results in Figure 4.8 show that the distribution of open-set queries is indeed a major factor in both closed-set and open-set performances for most transductive methods. Interestingly enough, some methods like Laplacian Shot (Ziko et al. 2020) or BDCSPN (J. Liu et al. 2020) benefit from this relaxation of the previous open-set assumption. However, while the closed-set accuracy of open-set transductive methods (Oslo and Ostim) increases in the new setting, their open-set recognition ability decreases (while still achieving the best results across the benchmark).

### Ablations.

**Ostim takes the best of both worlds.** We refer to the results in Figure 4.9 to motivate the design choices of Ostim: (i) Even at initialization, Ostim achieves high outlier detection performances but requires prototype refinement through the mutual information maximization in Eq. (4.4) to improve its closed-set accuracy. (ii) The inductive bias that consists in implicitly defining the *outlier* prototype as the diametrically opposite of the average of support prototypes is crucial. Introducing and optimizing an independent prototype as in the large-scale open-set PROSER (Zhou et al. 2021) only adds up to the ambiguity of the few-shot problem and ends up achieving poor outlier detection performances.

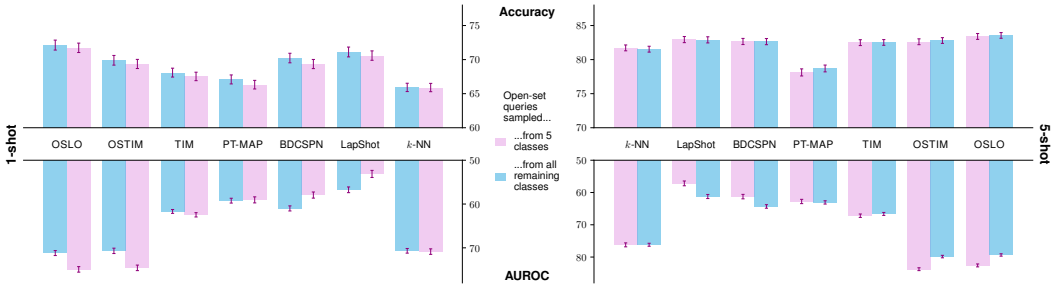
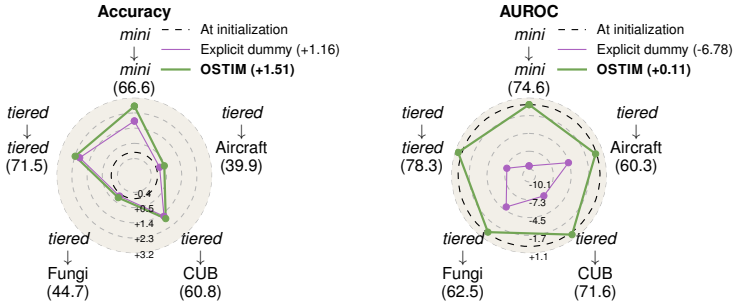
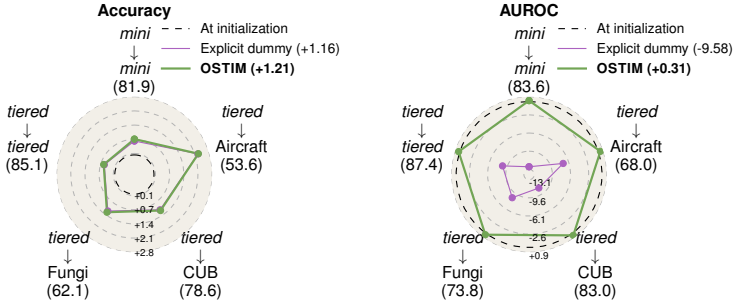



Figure 4.8: **Performance of transductive methods in the broad open-set setting.** We study the closed-set (accuracy) and open-set (AUROC) performance of transductive methods depending on the size of the query set on *mini*-ImageNet in the 1-shot and 5-shot settings. We add the inductive method $k$-NN + SimpleShot to compare with a method that is by nature independent of the number of queries.

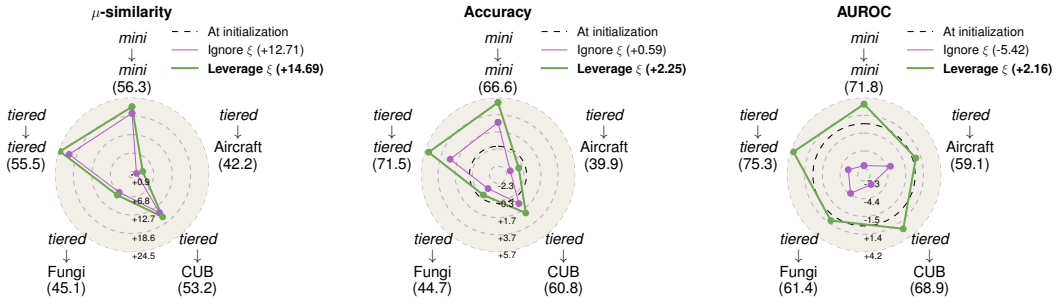



(a) 1-shot

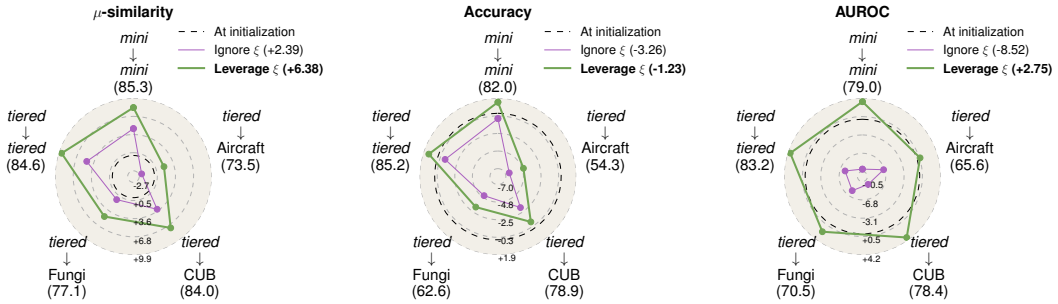



(b) 5-shot

Figure 4.9: **Ostim's ablation study.** Effects of optimizing the prototypes with Information Maximization (Eq. (4.7)) **and** using an *implicit* outlier prototype (Eq (4.2)) on the closed-set accuracy and the open-set performance measured with the AUROC. We compare the full Ostim method to a version of Eq. (4.7) with an explicit dummy prototype and to the model at initialization (before information maximization). This figure follows the same logic as Figure 4.5.




(a) 1-shot




(b) 5-shot

Figure 4.10: **Oslo's ablation study.** Effects of leveraging the inlier latent $\boldsymbol{\xi}$ on the quality of the model's both closed-set parameters $\boldsymbol{Z}$ (measured with the accuracy) and $\boldsymbol{\mu}$ (measured with the cosine similarity between $\boldsymbol{\mu}$ and the ground truth prototypes computed as the average of all support and query embeddings for each class) and the open-set performance measured with the AUROC. We compare the full Oslo method from Eq. (4.8) (Leverage $\boldsymbol{\xi}$) with the standard likelihood objective from Eq. (4.7) (Ignore $\boldsymbol{\xi}$) and no optimization (At initialization). This figure follows the same logic as Figure 4.5.

The inlier latent is essential to Oslo. We perform an ablation study on the important ingredients of Oslo. As a core contribution of our work, we show in Figure 4.10 that the presence *and* optimization of the latent *inlierness* scores is crucial. In particular, the closed-form latent score $\xi$ 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 $72\%$ to $65\%$ on 1-shot *mini*ImageNet. 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 in both accuracy and outlier detection across all in-domain benchmarks. However, we note that the optimization of the inlier latent, while still improving the parameters $\boldsymbol{\mu}$ of the closed-set parametric model, does not benefit the closed-set accuracy in cross-domain scenarios, accrediting the idea that such scenarios would need specifically tuned hyperparameters of the optimization model.

## 4.6 Discussion and Limitations

In this study, we advocate for Transduction as a promising avenue to address the difficult FSOSR problem. Through the proposed implicit prototype idea, we show that the InfoMax method TIM can be successfully *opened*. Going further, we show that a simple principle optimization framework can just be just as effective for Few-Shot Open-Set Recognition. We further insist that the proposed techniques do not necessitate any particular training process or model-specific parameter optimization, and can therefore be plugged effortlessly into the most expressive feature extractors. We hope that the promising results we obtained in this setting, using the latest advances in representation learning, will encourage the community to go beyond small residual networks and leverage these advances in our methods more often than we do today.

In this chapter, we also provided some long-overdue experiments about a widely anticipated limitation of transductive methods: the potential performance drop in the context of few queries. Our experiments show that most transductive methods still perform better than the inductive baseline in terms of closed-set accuracy, even in a 1-query regime.

Taking a step back, we showed that transductive methods which achieved outstanding results in the closed-set scenario suffer a serious drop in performance when faced with open-set instances. Therefore, we argue that the open-set scenario should be included in the standard benchmarks for few-shot transductive methods, and hope that our findings will encourage the community to follow this practice.

# Part III

# Challenges in Benchmarking Few-Shot Image Classification models

# 5 Contribution 3: Semantic Similarity in Few-Shot Learning Benchmarks

This chapter replicates our paper *Few-Shot Image Classification Benchmarks are Too Far From Reality: Build Back Better with Semantic Task Sampling*, by Etienne Bennequin, Myriam Tami, Antoine Toubhans, and Céline Hudelot, published in the Workshop on Vision Datasets Understanding at CVPR 2022 (Bennequin, Tami, et al. 2022).

## 5.1 Introduction

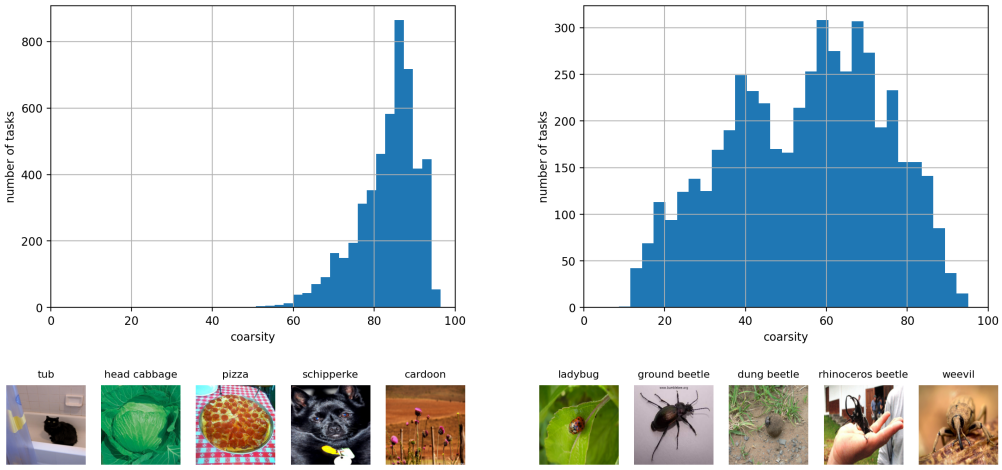
In the last four years, our Sicara team has been involved in a variety of industrial use cases for Few-Shot Learning. In Section 1.2, we observed that all of these use cases (namely industrial part recognition, retrieval in a marketplace's catalog, daily food recognition, and microorganisms recognition) are semantic fine-grained classification tasks: they all consisted in recognizing an object among many classes that were semantically similar to one another.

On the contrary, standardized academic benchmarks generate tasks using uniform random sampling from a wide range of semantically dissimilar or unrelated classes (*e.g.*, *tiered*ImageNet), which leads to evaluating our models mostly on tasks composed of objects that we would never need to distinguish in real-life use cases. In Figure 5.1a, we show an example of a task representative of the *tiered*ImageNet benchmark. This task obviously cannot be related to any real-life use case. More specifically, compared to the use cases presented in Section 1.2, we notice that the compared classes are much more semantically distant from one another. In Figure 5.2 (resp. 5.3), we show four more tasks representative of *mini*ImageNet (resp. *tiered*ImageNet) that share this same drawback. To allow the reader to verify these observations by themself, we provide in this project's webpage[1] a tool to manually sample tasks from *mini*ImageNet and *tiered*ImageNet.

Additionally, the Few-Shot Learning community has chosen to formalize the Few-Shot Image classification problem as an accumulation of $K$-way $n$-shot classification tasks *i.e.*, classifying query images, assuming that they belong to one of $K$ classes for which we have $n$ labeled examples each. In practice, most works compared their methods on benchmarks for which they fixed $K = 5$ (sometimes $K = 10$) and $n = 1$ or $n = 5$[2]. To the best of our knowledge, only one method for Few-Shot Image Classification was evaluated with $K > 50$ (Ramalho and Garnelo 2019). The choice made by the community, while relevant to facilitate experiments in the early stages of Few-Shot Learning research, casts a dark shadow on the robustness of state-of-the-art few-shot learning methods when discriminating between a large number of classes.

[1] https://semantic-task-sampling.streamlit.app/

[2] https://paperswithcode.com/task/few-shot-image-classification




(a) Coarsity histogram (top) and an example of task (bottom) of a testbed designed from *tiered*ImageNet with uniform class sampling. This task presents a coarsity of 85.1, which is the median coarsity for this testbed. "We really need a machine to distinguish bathroom tubs from cabbage, pizzas, cardoon, and some very specific kind of dog!" said no one in the history of humankind.

(b) Coarsity histogram (top) and an example of task (bottom) of our testbed *better-tiered*ImageNet. This task presents a coarsity of 15.8. Tasks with this coarsity never occur in the uniformly sampled testbed, although they are more representative of real few-shot classification use cases.

Figure 5.1: Comparison, in terms of our coarsity measure defined in Section 5.4.1, between a testbed designed with uniform class sampling (left) and a testbed designed with semantic awareness (right, ours). Our testbed gives a better representativity to tasks with low coarsity *i.e.*, composed of classes semantically relevant to one another.

Because of these limitations, we could not rely on the most popular Few-Shot Classification benchmarks to identify the most appropriate method for our use cases. How can we improve our current evaluation processes to better fit real-life use cases?

### Chapter's Contributions

In this chapter, we bring out some limitations of current Few-Shot Classification benchmarks with both quantitative and qualitative studies and propose new benchmarks to get past these limitations. More specifically:

1. We use the WordNet taxonomy (Miller 1995) to evaluate *semantic distances* between classes of the popular Few-Shot Classification benchmark *tiered*ImageNet. Based on these semantic distances we put forward the concept of *coarsity* of an image classification task, which quantifies how semantically close are the classes of the task.

2. We conduct both quantitative and qualitative studies of the tasks generated from the test set of *tiered*ImageNet *i.e.*, the tasks composing the benchmark on which most papers evaluate different methods. We show that this benchmark is heavily biased towards tasks composed of semantically unrelated classes.

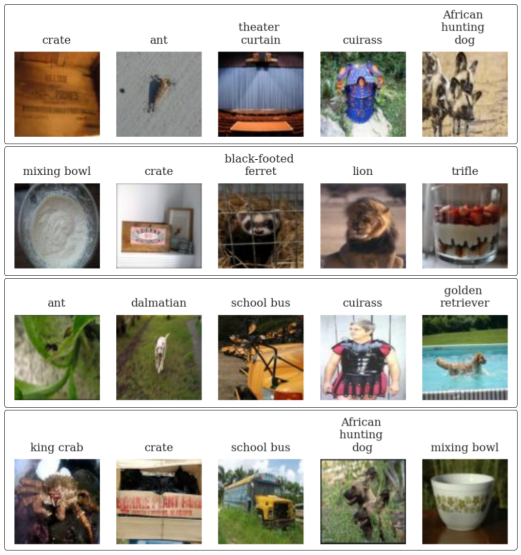



Figure 5.2: Four tasks sampled uniformly at random from the *mini*ImageNet benchmark.

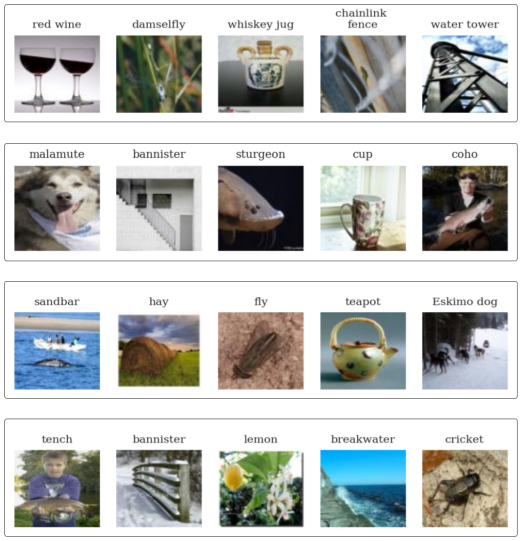



Figure 5.3: Four tasks sampled uniformly at random from the *tiered*ImageNet benchmark.

3. We harness the semantic distances between classes to generate the improved benchmark *better-tiered*ImageNet reestablishing the balance between fine-grained and coarse tasks. We compare state-of-the-art Few-Shot Classification methods on this new benchmark and bring out the relation between the *coarsity* of a task and its difficulty.

4. We put forward the Danish Fungi 2020 dataset (Picek et al. 2022) for evaluating Few-Shot Classification models. This dataset offers a wide range of fine-grained classes and therefore allows the sampling of tasks that we deem to be more representative of industrial applications of Few-Shot Learning. We compare state-of-the-art methods on both 5-way and 100-way tasks generated from this dataset. To the best of our knowledge, these are the first published results of few-shot methods on such wide tasks.

All our implementations, datasets, and experiments are publicly available[3]. We hope that our work will drive the research community towards a better awareness of the biases in few-shot evaluation processes and that the new benchmarks that we propose to counterbalance some of these biases will find echoes in the community and be further improved in future works.

[3] https://github.com/sicara/semantic-task-sampling

## 5.2 Positioning with respect to fine-graininess in Few-Shot Learning

Recent works proposed specific methods for Fine-Grained Few-Shot Image Classification (Ruan et al. 2021; Tang et al. 2020; J. Xu et al. 2021; Y. Zhu et al. 2020). These methods are typically compared on CU-Birds (Welinder et al. 2010) or FGVC Aircraft (Maji et al. 2013). These datasets propose respectively 50 and 25 test classes. In this work, we propose to use Danish Fungi 2020 (Picek et al. 2022), a fine-grained image classification dataset equipped with 1604 classes, which allows us to compare methods on tasks composed of a large number of classes. Also note that among many other contributions, the original paper for Meta-Dataset (Triantafillou et al. 2020) proposed a small study of the performance of state-of-the-art few-shot classifiers depending on a measure of *task fine-graininess* on ImageNet, as pointed out in Section 2.3.1. However, they failed to highlight any correlation between the fine-graininess and difficulty of a task, leaving it for future works. We claim to be this future work: compared to (Triantafillou et al. 2020), we use more precise tools to define the fine-graininess of a task and decorrelate the fine-graininess from the shape of the task (*i.e.*, the number of ways). Thereby we successfully show in Section 5.6.2 the correlation between fine-graininess and difficulty on *tiered*ImageNet.

## 5.3 Problem formalization

We follow the formalization defined in Section 2.2.1. Since it is of the utmost importance in this section, here we rewrite the definition of the set of $K$-way classification tasks that can be sampled from the test set $\mathcal{D}_{\text{test}}$:

$$\begin{aligned}\mathcal{E}_{\text{test}}(K) = \{\mathbb{T}_{\mathbb{S},\mathbb{Q}} \,|\, &\mathbb{S} = \{(\boldsymbol{x}_i^s, y_i^s) \in \mathcal{X} \times \mathbb{C}\}_{i=1\ldots|\mathbb{S}|} \subset \mathcal{D}_{\text{test}}, \\ &\mathbb{Q} = \{\boldsymbol{x}_i^q \in \mathcal{X}\}_{i=1\ldots|\mathbb{Q}|} \subset \mathcal{D}_{\text{test}}, \\ &\mathbb{S} \cap \mathbb{Q} = \emptyset,\ \mathbb{C} \subset \mathcal{C}_{\text{test}} \text{ and } |\mathcal{C}_{\text{test}}| = K\}\end{aligned}$$

with $\mathcal{D}_{\text{test}}$ the test dataset, $\mathcal{C}_{\text{test}}$ its set of classes, and $\mathbb{S}$ and $\mathbb{Q}$ respectively the support and query set for a particular task $\mathbb{T}$ with classes $\mathbb{C}$.

As we established in Section 2.2.1, even though most benchmarks are limited to $\mathcal{E}_{\text{test}}(5)$ (which we will note $\mathcal{E}_{\text{test}}$ when there is no ambiguity), and further limited to tasks with 1 or 5 support images per class ($|\mathbb{S}| = 5$ or 25) and 10 query images per class ($|\mathbb{Q}| = 50$), the number of possible tasks is still untractable ($\sim 10^{173}$ on *tiered*ImageNet). For this reason, it is common practice in the community to evaluate few-shot classification models on a subset $\tilde{\mathcal{E}}_{\text{test}} = \{\mathbf{T} \in \mathcal{E}_{\text{test}} | \mathbf{T} \sim \mathcal{U}\}$ with $\mathcal{U}$ the uniform distribution. This design choice gives the same weight to all tasks from $\mathcal{E}_{\text{test}}$, regardless of how informative they are on a model's ability to perform on real-life few-shot learning problems. In this chapter, we study alternative distributions for $\tilde{\mathcal{E}}_{\text{test}}$.

**Class sampling and instance sampling** Here we limit ourselves to the sampling of the classes that constitute each episode. We iterate only on the probability distribution over classes that constitute a task. Once the classes are sampled, we use uniform sampling over all instances for each class to constitute the support and query sets.

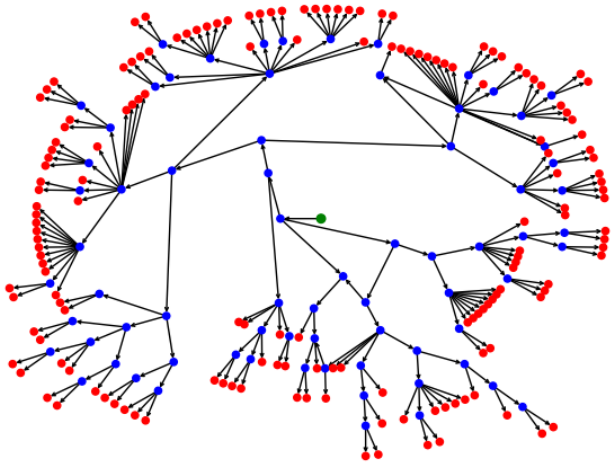


Figure 5.4: Directed acyclic subgraph of WordNet spanning the 160 classes of *tiered*ImageNet's test set, which are shown in red. The root (in green) corresponds to the concept of "entity".

## 5.4 Building the *better-tiered*Imagenet benchmark

### 5.4.1 Measuring task coarsity with WordNet taxonomy

Since *tiered*ImageNet is a subset of ImageNet, its classes are the leaves of a directed acyclic graph which is a subgraph of the WordNet graph (Miller 1995). This graph is represented in Figure 5.4. Using this graph, it is possible to establish a semantic similarity between classes. We use the Jiang & Conrath pseudo-distance between classes (J. J. Jiang and Conrath 1997), which is defined for two classes $k_1$ and $k_2$ as:

$$D^{JC}(k_1, k_2) = 2 \log |lso(k_1, k_2)| - (\log |k_1| + \log |k_2|) \tag{5.1}$$

where we note $|k|$ the number of instances of the dataset with class $k$, and $lso(k_1, k_2)$ is the lowest superordinate, *i.e.*, the most specific common ancestor of $k_1$ and $k_2$ in the directed acyclic graph. Our choice of pseudo-distance[4] was motivated by the results of Deselaers and Ferrari 2011, who showed that this semantic pseudo-distance between classes is correlated with the visual similarity between images of these classes on ImageNet. We insist, however, that the definition of the coarsity and the subsequent sampling strategy are agnostic of the choice of the distance, and that other semantic distances may be used in future works.

From this pseudo-distance, we define the *coarsity* $\kappa$ of a task $\mathbb{T}_{\mathbb{C}}$ constituted of instances from a set of classes $\mathbb{C}$ as the mean square distance between all pairs of classes in $\mathbb{C}$ *i.e.*,

[4] We call it a pseudo-distance since it is positive, symmetric and separated. However, due to the weak assumption on the directed acyclic graph, it is possible to find $k_1, k_2, k_3$ such that $D^{JC}(k_1, k_3) > D^{JC}(k_1, k_2) + D^{JC}(k_2, k_3)$. In practice, our datasets are sufficiently balanced for this case not to occur.

$$\kappa(\mathbb{T}_{\mathbb{C}}) = \underset{k_i,k_j\in\mathbb{C},\, k_i\neq k_j}{\text{mean}} D^{JC}(k_i,k_j)^2 \tag{5.2}$$

This coarsity is an indicator of how semantically close are the classes that constitute a task. As shown in Deselaers and Ferrari 2011, on datasets derived from ImageNet, the semantic distance is closely linked to the visual similarity between items of these classes, and therefore the coarsity of a task is closely linked to the average visual similarity between the classes that compose this task.

Note that since we use the mean of square distances, a task with one class very distant from the others will have a higher coarsity than a task composed of 5 reasonably distant classes.

### 5.4.2 Generating a more informative benchmark using class semantics

As discussed in Section 5.3, finding an appropriate subset of $\mathcal{E}_{\text{test}}$ is a key point to ensure that we provide an accurate evaluation of a model. In the literature, testing tasks are sampled uniformly at random from $\mathcal{E}_{\text{test}}$ (Vinyals et al. 2016). However, we observed that the resulting testbeds are biased towards tasks with high coarsity *i.e.*, composed of classes semantically far from each other with respect to the Jiang & Conrath pseudo-distance (see Figure 5.1a). We argue that in practice, few-shot learning models are often used to distinguish between similar objects rather than distinguishing between objects that have nothing to do with one another (*e.g.*, circuit boards from circuit boards, carpets from carpets, or people from people). This is, in fact, the case of all the real use cases that we presented in Section 1.2. A testbed presenting this type of bias is therefore irrelevant to evaluate a model's ability to solve this family of problems.

In this work, we define a unique, reproducible set of testing tasks to evaluate all models. This testbed is built with a dual objective:

- We want tasks with a smooth repartition in terms of coarsity to ensure that the testbed also evaluates the ability of a model to distinguish between classes close to each other. Providing a good span of coarsities also allows to compare models on different types of tasks: a model might be better for coarse tasks but not for fine-grained tasks.
- This first objective inherently creates a bias towards classes with many neighboring classes. However we want our testbed to be balanced, *i.e.*, all images must be sampled roughly as many times as the others[5].

To achieve these goals, we define a semantic task sampler based on the Jiang & Conrath pseudo-distance. As presented in Section 2.3.1, C. Liu et al. 2020 propose to condition the probability of co-sampling two classes in one training episode using a potential matrix for pairs of classes. Building on their framework, we build an initial potential matrix $\boldsymbol{\mathcal{P}}^0$ such that

$$\boldsymbol{\mathcal{P}}^0_{i,j} = \exp(-\alpha D^{JC}(k_i,k_j)) \tag{5.3}$$

with $\alpha \in \mathbb{R}_+$ an arbitrary scalar. For the first task, the probability for a pair of classes $(k_i,k_j)$ to be sampled together is proportional to $\boldsymbol{\mathcal{P}}^0_{i,j}$. To enforce that the testbed is balanced, once the

[5] In the case of *tiered*ImageNet, which presents as many images for each class, this is equivalent to ensuring the balance between classes.

$t-1^{\text{th}}$ task is sampled we update the number $\nu_i^t$ of occurrences of class $k_i$ in previous tasks. Then we update the potential matrix to penalize classes with higher values of $\nu_t$:

$$\boldsymbol{\mathcal{P}}_{i,j}^t = \boldsymbol{\mathcal{P}}_{i,j}^0 \times \exp(-\beta \frac{\nu_i^t + \nu_j^t}{\max_k(\nu_k^t)}) \tag{5.4}$$

with $\beta \in \mathbb{R}_+$ an arbitrary scalar. Intuitively, a larger $\alpha$ gives more weight to pairs of semantically close classes, while a larger $\beta$ forces a stricter balance between classes. The class sampling process is detailed in Algorithm 2.

We then sample instances from these classes uniformly at random. As shown in Figure 5.1b, our 5000-tasks testbed gives far greater representation to fine-grained tasks compared to a uniformly sampled testbed. Our testbed offers a wide range and balance of task coarsities, allowing to test models on both coarse and fine-grained tasks, while the uniformly sampled testbed only allows the evaluation on coarse tasks. Figure 5.5 shows that the occurrences-based penalty successfully enforces the balance between classes in our testbed.

**Algorithm 2** How classes of a task are sampled using the potential matrix, following C. Liu et al. 2020

**Input:** potential matrix $\boldsymbol{\mathcal{P}}^0$, number of tasks $T$, number of classes per task $K$
**Output:** set of sampled tasks $\tilde{\mathcal{E}}_{\text{test}}$

$\tilde{\mathcal{E}}_{\text{test}} \longleftarrow \{\}$
$\boldsymbol{\nu} \longleftarrow \mathbf{1}$
**for** $t < T$ **do**
  $\mathbf{p} \longleftarrow \exp(-\beta \frac{\boldsymbol{\nu}}{\max(\boldsymbol{\nu})})$
  $\mathbb{C} \longleftarrow \{k_0\}$ with $k_0$ sampled according to a distribution of probability proportional to $\mathbf{p}$
  $\mathbf{p} \longleftarrow \mathbf{p} \odot \boldsymbol{\mathcal{P}}_{k_0}^0$
  **while** $|\mathbb{C}| < K$ **do**
    $\mathbb{C} \longleftarrow \mathbb{C} \cup \{k\}$ with $k$ sampled according to a distribution of probability proportional to $\mathbf{p}$
    $\mathbf{p} \longleftarrow \mathbf{p} \odot \boldsymbol{\mathcal{P}}_k^0$
  **end while**
  $\tilde{\mathcal{E}}_{\text{test}} \longleftarrow \tilde{\mathcal{E}}_{\text{test}} \cup \{\mathbb{C}\}$
  $\forall i \in \mathbb{C}, \nu_i \longleftarrow \nu_i + 1$
**end for**

## 5.5 Fungi: a large fine-grained dataset for Few-Shot Image Classification

### 5.5.1 Danish Fungi 2020

Danish Fungi 2020 (DF20) (Picek et al. 2022) is an image recognition dataset of 295 938 images of fungi distributed in 1604 fine-grained classes, with no overlap with ImageNet. The dataset offers

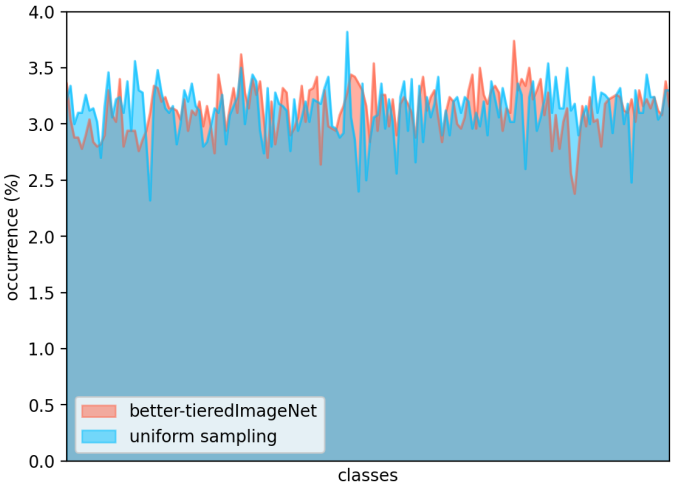



Figure 5.5: Comparison of class imbalance in testbeds. We show the proportion of tasks containing each one of the 160 classes of *tiered*ImageNet's test set, with uniform task sampling (blue) and our semantic task sampling (red). A perfectly balanced testbed will show a flat line *i.e.*, each class is equally represented in the testbed. We see that our sampling does not raise class imbalance compared to uniform sampling.

insightful metadata, such as the object's geographical location, habitat, and substrate. DF20's classes are equipped with a seven-level hierarchical structure. Note that Meta-Dataset (Triantafillou et al. 2020) also includes a Fungi dataset, from the FGVCx 2018 Fungi classification challenge[6]. This Fungi dataset comes from the same source as DF20 but offers fewer images, fewer classes, no metadata, and no taxonomy, which convinced us to push forward DF20. The semantic tree of DF20 is shown in Figure 5.6.

### 5.5.2 DF20 for Few-Shot Image Classification Benchmark

Why do we use it? Following our observations on the high coarsity of tasks sampled from *tiered*ImageNet, we propose to use DF20 as a test set for few-shot image classification models. This dataset allows sampling a wide variety of fine-grained few-shot classification tasks, which are to our experience more representative of real-world applications. It also offers a taxonomy allowing to further study the performance of few-shot models depending on the coarsity of the task. Since we consider the whole dataset as a test set, we allow the comparison of methods with parameters optimized on various training sets, provided that they do not overlap with DF20. This brings the few-shot learning methodology closer to the neighboring field of transfer learning, in which training data is not part of the benchmark (Dumoulin et al. 2021). This can also be considered as a generalization of the cross-domain few-shot learning setting (W.-Y. Chen et al. 2019), in which we train a model on the training set of one benchmark (*e.g.*, *mini*ImageNet) and test it on the testing set of another benchmark (*e.g.*, CUB). Note that in this work, compared methods all share the same backbone with parameters classically trained on ImageNet's training set. This was a

[6] https://github.com/visipedia/fgvcx_fungi_comp

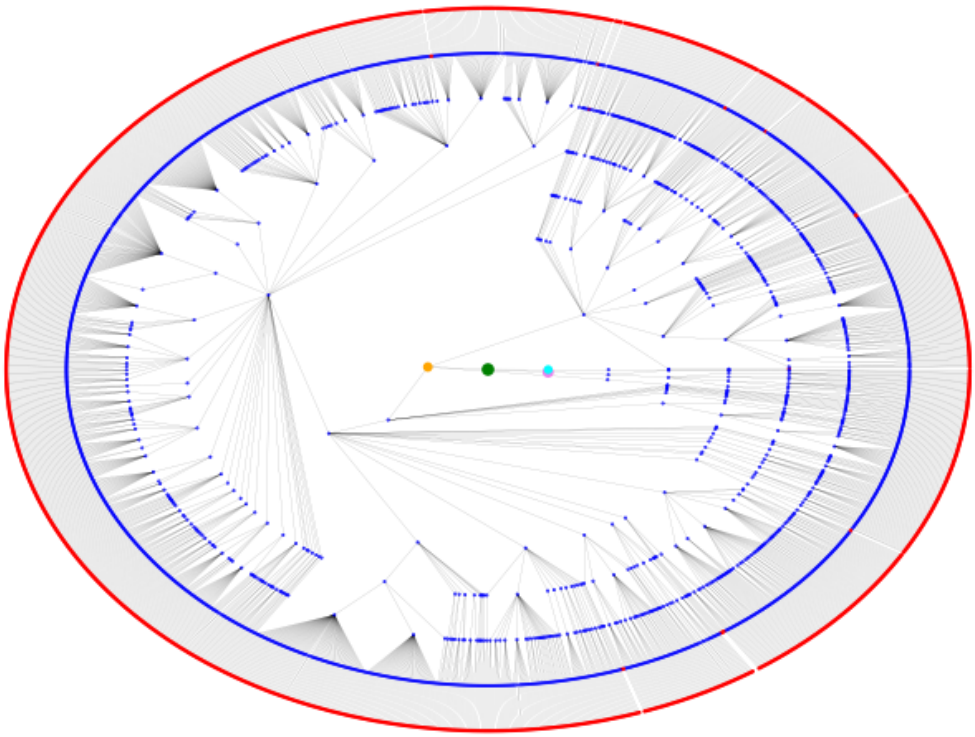


Figure 5.6: Semantic tree spanning the 1604 classes of DF20 (in red). The root is shown in green. DF20 contains images of species that belong to three kingdoms: Chromista (purple), Protozoa (cyan), and Fungi (orange). As we can see from the tree, the vast majority of classes are spanned from the Fungi kingdom.

| | *better-tiered*ImageNet | | | | | Uniform testbed |
|---|---|---|---|---|---|---|
| | Whole testbed | 1st. Qtl | 2nd. Qtl | 3rd Qtl | 4th Qtl | |
| **1-shot** | | | | | | |
| ProtoNet (Snell et al. 2017) | 53.10 ± 0.40 | 42.01 | 51.11 | 56.35 | 62.88 | 65.24 ± 0.35 |
| Finetune (W.-Y. Chen et al. 2019) | 56.60 ± 0.41 | 44.56 | 54.29 | 60.12 | 67.43 | 69.96 ± 0.35 |
| BD-CSPN (J. Liu et al. 2020) | **59.99 ± 0.45** | **46.69** | **57.50** | **63.80** | **71.95** | **74.55 ± 0.37** |
| TIM (Boudiaf, Ziko, et al. 2020) | **59.19 ± 0.43** | **46.74** | **56.76** | **63.03** | 70.21 | 73.09 ± 0.35 |
| Transductive Finetuning (Dhillon et al. 2020) | 53.01 ± 0.40 | 41.98 | 51.74 | 56.31 | 62.82 | 65.27 ± 0.35 |
| PT MAP (Hu et al. 2021) | 56.74 ± 0.38 | 45.81 | 54.38 | 59.76 | 66.96 | 69.58 ± 0.37 |
| **5-shot** | | | | | | |
| ProtoNet (Snell et al. 2017) | 70.77 ± 0.37 | 59.72 | 68.95 | 74.05 | 80.41 | 82.79 ± 0.25 |
| Finetune (W.-Y. Chen et al. 2019) | 71.60 ± 0.37 | 60.63 | 69.68 | 74.86 | 81.26 | 83.66 ± 0.25 |
| BD-CSPN (J. Liu et al. 2020) | 72.50 ± 0.37 | 61.23 | 70.47 | **75.94** | **82.38** | 84.70 ± 0.25 |
| TIM (Boudiaf, Ziko, et al. 2020) | **73.32 ± 0.37** | **62.40** | **71.48** | **76.53** | **82.92** | **85.49 ± 0.24** |
| Transductive Finetuning (Dhillon et al. 2020) | 70.79 ± 0.37 | 59.73 | 68.98 | 74.08 | 80.43 | 82.79 ± 0.25 |
| PT MAP (Hu et al. 2021) | 69.45 ± 0.36 | 58.97 | 67.40 | 72.36 | 79.09 | 81.54 ± 0.26 |

Table 5.1: Top-1 accuracy of various few-shot learning methods on 1 and 5-shot tasks sampled from the tieredImageNet test set with uniform and semantic sampling strategies. For better-tieredImageNet, we show the average accuracy on the whole testbed, along with the average accuracy on the four quartiles of the testbed, when sorted by coarsity of the task (1st quartile contains the tasks with the smallest coarsity). For the testbed of uniformly sampled tasks, we only show the average accuracy on the whole testbed, as the results on each quartile are very similar to one another. For each setting, we separate inductive (top) from transductive (bottom) methods. Best method(s) in each column is shown in bold.

convenient baseline, but we insist that this does not make it mandatory for future work to train on ImageNet. In fact, we believe that it is of prior importance to study the effect of the choice of the training data on the few-shot classification model's performance.

How do we use it? We sampled four benchmarks: 5-way 1-shot, 5-way 5-shot, 100-way 1-shot, and 100-way 5-shot. The 5-way settings are very common in the few-shot learning literature (Triantafillou et al. 2020). However, to the best of our knowledge, very few works evaluate their method on *wide* (*i.e.*, more than 10-way) few-shot classification tasks[7]. We believe that such tasks are at least equally interesting as 5-way tasks. We assume that this setting was avoided by early works because GPU memory limitations made it very hard to use back-propagation on batches mimicking 100-way tasks during episodic training. We claim that these constraints do not justify overlooking such an interesting problem. Specifically for DF20, the task of recognizing an image among a wide variety of fungi makes way more sense than recognizing an image among 5 random species[8].

## 5.6 Experiments on new benchmarks

### 5.6.1 Implementation details

We conducted the necessary experiments to bring out the need for novel few-shot classification benchmarks and showcase the limitations of state-of-the-art methods on more challenging settings. We restricted the comparison to methods allowing classical training (*i.e.*, non episodic) and to a unique set of hyper-parameters. All parameters of our experiments can be found on our publicly available code [9].

*tieredImageNet* We followed the original split of Ren et al. 2019. All methods tested in our benchmark use a common ResNet12 with parameters trained for 500 epochs with classical cross-entropy among the 351 classes of the train set using stochastic gradient descent with a batch size of 512 and learning rate of 0.1 with a decreasing factor of 0.1 after 350, 450 and 480 epochs. The trained weights are directly downloadable from our code. We built two testbeds with uniform class sampling (1-shot and 5-shot), and two testbeds (1-shot and 5-shot) with semantic task sampling (see Section 5.4.2) with $\alpha = 0.383$ and $\beta = 100.0$. These hyperparameters were selected to enforce the sampling of tasks with small coarsity while ensuring that all classes were equally represented in the testbed (with a small margin). This selection was monitored with visual observations shown in Figures 5.1b and 5.5. We upsampled 10000 tasks, then we removed all duplicate tasks and downsampled them to 5000 tasks. All tasks present 10 queries per class.

Danish Fungi 2020 All methods tested in our benchmark use the built-in ResNet18 from `PyTorch` with weights trained on ImageNet. Since DF20 is already fine-grained, we built four 5000-task testbeds (5-way 1-shot, 5-way 5-shot, 100-way 1-shot, and 100-way 5-shot) with uniform class sampling. All tasks present 10 queries per class.

Methods For Finetune (W.-Y. Chen et al. 2019) we use 10 fine-tuning steps with a learning rate of $10^{-3}$. For Transductive Information Maximization (TIM) (Boudiaf, Ziko, et al. 2020) we use 100 fine-tuning steps with a learning rate of $10^{-3}$ and put a 0.1 weight on the conditional entropy term of the loss. For Transductive Finetuning (Dhillon et al. 2020) we use 25 inference steps with a learning rate of $5 \times 10^{-5}$. These hyperparameters were selected to fit the original implementations of these methods. Following the discussion on the value of model-agnosticity for Few-Shot Learning methods in Chapter 4, we insist that we didn't put any additional effort into further optimizing any few-shot method on our benchmarks.

### 5.6.2 Results

*tieredImageNet* Results for *tiered*ImageNet are shown in Table 5.1. The immediate observation that we can make is that our benchmark *better-tiered*ImageNet is much more challenging than uniform task sampling, with a performance drop of 12 to 15% in top-1 accuracy for all settings and methods. For further details, we sorted all 5000 tasks with respect to their coarsity and grouped

[7] `https://paperswithcode.com/task/few-shot-image-classification`

[8] As of 2020, experts have identified ∼148 000 species of fungi (Cheek et al. 2020).

[9] `https://github.com/sicara/semantic-task-sampling`

| | 5-way | | 100-way | | | |
|---|---|---|---|---|---|---|
| | 1-shot | 5-shot | 1-shot | | 5-shot | |
| | Top-1 | | Top-1 | Top-5 | Top-1 | Top-5 |
| ProtoNet (Snell et al. 2017) | 37.55 ± 0.25 | 60.53 ± 0.27 | 7.81 ± 0.08 | 20.12 ± 0.13 | 17.69 ± 0.12 | 40.20 ± 0.16 |
| Finetune (W.-Y. Chen et al. 2019) | 47.00 ± 0.30 | 65.06 ± 0.29 | **9.70 ± 0.09** | 25.94 ± 0.15 | **19.60 ± 0.12** | **44.25 ± 0.17** |
| BD-CSPN (J. Liu et al. 2020) | 47.81 ± 0.33 | **66.32 ± 0.30** | **9.75 ± 0.09** | 24.11 ± 0.15 | **19.52 ± 0.13** | 41.79 ± 0.17 |
| TIM (Boudiaf, Ziko, et al. 2020) | 40.73 ± 0.28 | 62.89 ± 0.28 | 8.36 ± 0.09 | 21.30 ± 0.14 | 18.53 ± 0.12 | 41.47 ± 0.17 |
| Trans. Finetuning (Dhillon et al. 2020) | 37.54 ± 0.25 | 60.54 ± 0.27 | 7.71 ± 0.08 | 20.13 ± 0.13 | 17.69 ± 0.12 | 40.21 ± 0.16 |
| PT MAP (Hu et al. 2021) | **52.08 ± 0.35** | **66.78 ± 0.29** | 9.54 ± 0.09 | **26.37 ± 0.15** | 18.50 ± 0.12 | 43.36 ± 0.16 |

Table 5.2: Accuracy of various few-shot learning methods on DF20. For 100-way tasks, we report both top-1 and top-5 accuracy. For 5-way tasks, we do not report top-5 accuracy as we found that it was always 100%. Best method(s) in each column is shown in bold.

them into four quartiles. The 1st quartile contains the most fine-grained tasks, and the 4th quartile contains the coarsest tasks. From these results, we can confirm that coarsity is indeed correlated to the difficulty of the task since the performance consistently improves when moving towards coarser tasks. We finally observed that even the 4th quartile seems to be more challenging than the uniform benchmark. This is consistent with the demography of tasks shown in Figure 5.1, since the tasks constituting the 4th quartile of our testbed show a smaller average coarsity than the uniform testbed.

We observed that transductive methods (Boudiaf, Ziko, et al. 2020; Dhillon et al. 2020; Hu et al. 2021; J. Liu et al. 2020), which use the unlabeled information from the query set, unsurprisingly show the best results on both set-ups but especially in 1-shot classification. The leaderboard seems to be consistent on all quartiles, suggesting that none of these methods are "specialized" towards a particular demographic of tasks.

Danish Fungi 2020 Results for DF20 are shown in Table 5.2. They show that while being more challenging than *tiered*ImageNet and *better-tiered*ImageNet, our DF20 benchmark still constitutes an achievable task. We also report results showing that all methods struggle in the more challenging problem of 100-way classification, especially in the 1-shot setting (less than 10% top-1 accuracy for the best model, less than 20% in the 5-shot setting). We believe that this should stand as a red flag regarding the ability of state-of-the-art few-shot classification methods to scale to real-life use cases.

To complete the study, we show in Figure 5.7 the correlation between the coarsity of a task (based on the taxonomy of DF20) and the accuracy of the PT-MAP (Hu et al. 2021) method. We observe that, as was the case for *better-tiered*ImageNet, the closest the classes sampled from Fungi are from one another, the harder the task composed of these classes.

## 5.7 Conclusion

We showed that the widely used *tiered*ImageNet benchmark with a uniform sampling of classes led to evaluating few-shot learning models on disproportionately coarse tasks. We used semantic task sampling to generate a more informative testbed from *tiered*ImageNet's test set. We also pushed forward as a new benchmark for Few-shot Image Classification models the Danish Fungi 2020

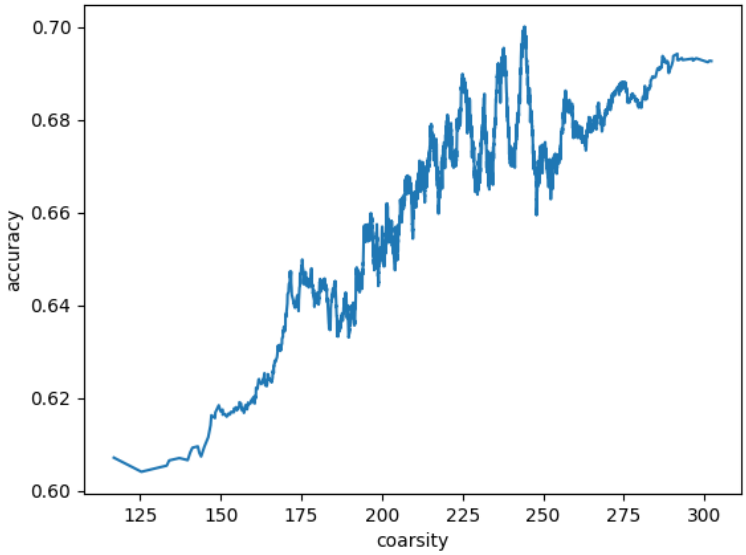



Figure 5.7: Correlation between the coarsity of the 5000 tasks sampled from DF20 for our 5-way 5-shot testbed and the accuracy of the PT-MAP (Hu et al. 2021) method. The plot is smoothed using the rolling average over a window of 200 tasks.

dataset, which we believe to be an incredibly promising playing field for future research in Few-Shot Learning. Finally, we showed that state-of-the-art methods dramatically fail when confronted with many-way classification (here many is 100) of fine-grained objects. We insist that this setting is **not** far-fetched and fits tangible industrial use cases. We believe that these results should push us to take a step back and re-assess the way we currently think about Few-Shot Image Classification.

We used to define a few-shot classification task by its number of ways and its number of shots, addressing $n$-way $k$-shot classification as an indivisible problem. What we did here can be seen as a novel framework, in which the number of classes is not sufficient to define a task: we need to know what these classes are. To go further, we would need to go beyond defining tasks by their number of shots and consider which images are chosen for the support set. This is the source motivation for the perspectives drawn in Chapter 6.

In this work, we addressed what we believe to be a very limiting bias of current Few-Shot Learning benchmarks *i.e.*, a bias towards coarse tasks. We chose to tackle this particular shortcoming because we observed that it was the main difference between academic benchmarks and the industrial applications of Few-Shot Learning that we encountered. However, many more limitations of few-shot learning benchmarks are yet to address: the fixed shape of the tasks, the strict balance in both support and query sets, the empty overlap between large-scale classes (currently only used for base training) and few-shot classes, no prior in the choice of support instances, and many other of which we did not think yet. We believe that addressing these shortcomings must be considered a priority in our field, and we encourage any and all who agree to join us in this effort.

WHERE ARE THEY NOW? Very recently, the problem of sampling "hard" tasks to evaluate few-shot image classification was discussed by Basu et al. 2023 in building a new benchmark called *Hard-Meta-Dataset++*. They propose a model-based method to sample "difficult support sets", and find that the combination of their method with the semantic task sampling proposed in our contribution gives the most "difficult" tasks.

# 6 Perspective: Observations on Support Set Quality

In the previous chapters, we presented what we believe to be necessary steps to bridge the gap between Few-Shot Learning research and its application to industrial use cases. However, a lot remains to be done, especially in the way that we evaluate the methods. In Chapter 5, we limited our study to the relation between classes and did not consider the individual images forming the support set. We believe however that the selection of these images is crucial to achieving high performance in a Few-Shot setting. In this chapter, we report some experiments that we performed on this issue and draw some perspectives for future work that may focus on support set quality.

## 6.1 Motivation

In the last ten years, we counted thousands of iterations on Few-Shot Learning models: architectures (Koch et al. 2015), additional modules (Z. Jiang et al. 2020; Sung et al. 2018), comparison strategies (Snell et al. 2017), feature hallucination (Hariharan and Girshick 2017; Y.-X. Wang et al. 2018), transductive assumptions (Boudiaf, Ziko, et al. 2020; Dhillon et al. 2020; Y. Liu et al. 2019), learning schemes (Laenen and Bertinetto 2021; Ouali, Hudelot, et al. 2021), "meta-learning" schemes (Finn et al. 2017; Vinyals et al. 2016)... While this drove interesting theoretical discoveries and considerably improved performance on many specialized benchmarks, we argue that this race toward better models leaves an obvious blind spot: the quality of available data.

Data quality is universally known as a critical factor for performance in Machine Learning problems. Intuitively, the quality of individual samples is even more decisive when only a handful of labeled instances are available. Let us recall the context of Few-Shot Image Classification, with only one available sample per class (*i.e.*, *one-shot* classification). If the only example that we give an agent to define a class is of bad quality, we cannot hope for a good performance in recognizing this class.

We were able to verify this intuition through observations on the widely used *tiered*-ImageNet benchmark. Figure 6.1 shows two of the worst-performing tasks using a variety of Few-Shot Learning models. We can understand that our algorithms were struggling to accurately classify query images based only on the represented support examples.

Nonetheless, the issue of the quality of the support set has been understudied in the Few-Shot Learning literature (see Section 2.3.2). In fact, there is no definition or measure for the quality of a support set. This motivated us to explore this issue.


(a) A model would have to recognize a great white shark based on this small stain in the water, and would understand that a tub contains feet.


(b) A model would probably recognize a wine bottle based on the presence of a shirtless man smashing something by seaside.

Figure 6.1: The support sets for two 1-shot 5-way classification tasks sampled from *tiered*-ImageNet.

**Chapter's Summary**

In this chapter, we report our investigations and present what we deem to be interesting perspectives for future works on the support set's quality. More precisely, we focused on the two following questions:

1. How can we characterize the quality of an example in a support set?
2. What is the impact of the selection of support set instances on the performance of Few-Shot Learning models?

## 6.2 Assessing the quality of the support set

### 6.2.1 Problem statement

In order to measure the *quality* of a support set, we first need to define the purpose of a support set in the context of Few-Shot Classification: in this chapter, we will assume that a Few-Shot Learning model takes as argument both the query instance and the support set, to output a prediction. In other words, given a set of classes $\mathbb{C}$, we define a few-shot learning model $\psi_{\boldsymbol{\theta}} : (\boldsymbol{x}_i^q, \mathbb{S}) \mapsto \boldsymbol{p}_i^q = (\mathbb{P}(y_i^q = k|\boldsymbol{x}_i^q))_{k\in\mathbb{C}}$. For instance, when we are using prototypical classification (Snell et al. 2017) without any normalization and a given distance $||\cdot||$, the $k^{th}$ element of the prediction is obtained from the following, using a feature extractor $\phi_{\boldsymbol{\theta}}$ with parameters $\boldsymbol{\theta}$:

$$\psi_{\boldsymbol{\theta}}(\boldsymbol{x}_i^q, \mathbb{S})_k = \frac{e^{-||\phi_{\boldsymbol{\theta}}(\boldsymbol{x}_i^q)-\boldsymbol{\mu}_k||^2}}{\sum_{k'=1}^{|\mathbb{C}|} e^{-||\phi_{\boldsymbol{\theta}}(\boldsymbol{x}_i^q)-\boldsymbol{\mu}_{k'}||^2}} \text{ with } \boldsymbol{\mu}_k = \underset{\substack{(\boldsymbol{x}^s,y^s)\in\mathbb{S}\\ y^s=k}}{\text{mean}} \phi_{\boldsymbol{\theta}}(\boldsymbol{x}^s) \tag{6.1}$$

Then, the quality of the support set $\mathbb{S}$ can be considered a function of the correlation between $\psi_{\boldsymbol{\theta}}(\cdot, \mathbb{S})$ and the ground truth distribution. From this definition, it follows that the quality of

$\mathbb{S}$ depends on the Few-Shot Learning model $\psi_{\boldsymbol{\theta}}$. Therefore, in this work, we limit ourselves to definitions of the support set's quality inside of the feature space determined by the model's parameters $\boldsymbol{\theta}$. This choice and its consequences are discussed in Section 6.4.

### 6.2.2 Representativeness by distance to the class centroid

In the context of a dataset $\mathcal{D}_{\text{test}} = \{(\boldsymbol{x}_i, y_i) \mid \boldsymbol{x} \in \mathcal{X},\ y \in \mathcal{C}_{\text{test}}\}$, and a feature extractor $\phi_{\boldsymbol{\theta}} : \mathcal{X} \longrightarrow \mathcal{Z}$, we define the *representativeness* $\rho_{\boldsymbol{\theta}}(\boldsymbol{x})$ (or simply $\rho(\boldsymbol{x})$ when there is no ambiguity) of an instance $\boldsymbol{x}$ with label $y = k \in \mathcal{C}_{\text{test}}$ from its distance to its class centroid weighted by the class's standard deviation, in the feature space *i.e.*,

$$\rho_{\boldsymbol{\theta}}(\boldsymbol{x}) = ||\frac{\boldsymbol{\sigma}_k}{\phi_{\boldsymbol{\theta}}(\boldsymbol{x}) - \boldsymbol{\mu}_k}||_2^2 \tag{6.2}$$

$$\text{with } \boldsymbol{\sigma}_k = \sqrt{\sum_{y_i=k} \left(\phi_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - \boldsymbol{\mu}_k\right)^2}$$

From this, we define the representativeness of a support set $\mathbb{S}$ as the average representativeness of its instances *i.e.*,

$$\rho_{\boldsymbol{\theta}}(\mathbb{S}) = \underset{(\boldsymbol{x},y)\in\mathbb{S}}{\text{mean}}\ \rho_{\boldsymbol{\theta}}(\boldsymbol{x}) \tag{6.3}$$

Note that this definition of representativeness does not take into account the relationship between different instances and classes composing the support set. Once again, we refer to Section 6.4 for a discussion on this choice and its consequences.

We considered two other definitions of representativeness, that we report for completeness. They reported similar results as the representativeness by distance to the class centroid, but their computation time was several orders of magnitudes longer, which motivated our choice for the representativeness by distance to the class centroid.

- Representativity by average distance to other instances of the same class *i.e.*, $\rho_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{|k|-1} \sum_{y_i=k} \frac{||\phi_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - \phi_{\boldsymbol{\theta}}(\boldsymbol{x})||_2^2}{\boldsymbol{\sigma}_k 2}$, with $|k|$ the number of instances with label $k$ in $\mathcal{D}_{\text{test}}$.

- Representativity as the proportion of other instances of the same class lying in a neighborhood $\eta\boldsymbol{\sigma}_k$ of the considered instances.

### 6.2.3 Observations on *tiered*-ImageNet

We computed all representativeness for the test set of *tiered*-ImageNet, using the features extracted by a ResNet12 with checkpoints provided by the authors from Ye et al. 2020. Figure 6.2 shows all representativeness as a histogram to provide an idea of admissible values. In Figure 6.3 we provide some examples of strawberry-picked images with their representativeness, showing that images with low representativeness overall correspond to images that would be considered "bad examples" by human standards.

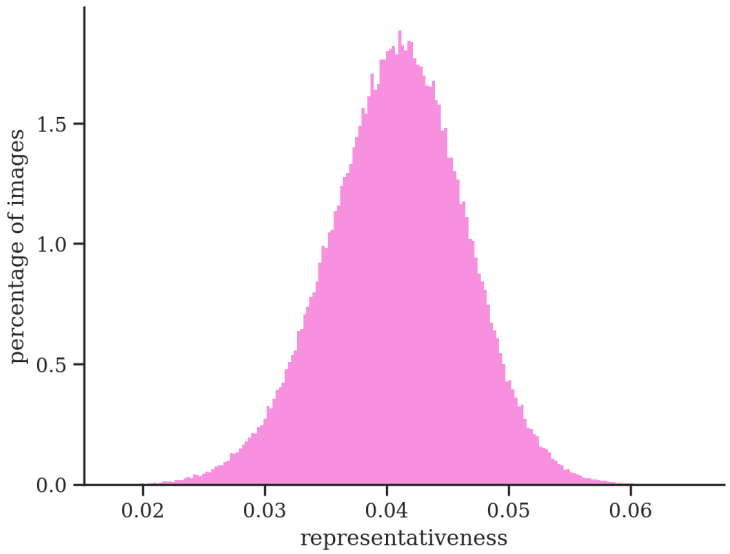



Figure 6.2: Histogram of representativeness in *tiered*-ImageNet's test set with features extracted by a ResNet12. Ordinate values correspond to the percentage of the dataset's population lying in each bin. The average representativeness is 0.0408. All values are contained in the range $[0.0175, 0.0653]$. 90% of images have a representativeness between 0.0313 and 0.0497.

(a) $\rho = 0.0230$ (b) $\rho = 0.0271$ (c) $\rho = 0.0360$ (d) $\rho = 0.0473$ (e) $\rho = 0.0578$

Figure 6.3: Examples of images of the class *strawberry* with increasing scores of representativeness $\rho$. Images with low representativeness may be considered as *worst* examples of strawberries than images with higher representativeness.

What are the causes for bad representativeness? In Figure 6.4, we display some of the images with the worst representativeness and expose four common causes for low representativeness:

1. A *distracting object* catches the "eye" (or attention of the model) at the expense of the object responsible for the image's label. This is a consequence of the *single-label* pattern of ImageNet: when labeling images, annotators are solely asked whether a label is present in the image, *not* whether it is the main object in the image. This, however, seems to be fixable with a simple re-cropping around the object of interest, as proposed in Bendou et al. 2022.

2. Re-cropping would not solve this issue in all instances, because some images suffer from the *sliding-door effect*: there is, as before, a distracting object, but it is "inside" of the object of interest.

3. The object, albeit being without question what is described in the label, is very different from what we would expect for an image with this label (*e.g.*, chocolate sauce is usually presented in a bowl or a plate, in the context of a desert).

4. The last cause is a very specific one. It is, in fact, most likely too specific to occur outside of ImageNet, but we report it for completeness: a fairly large number of images with low representativeness correspond to some kind of grid in the foreground partially hiding the object of interest.

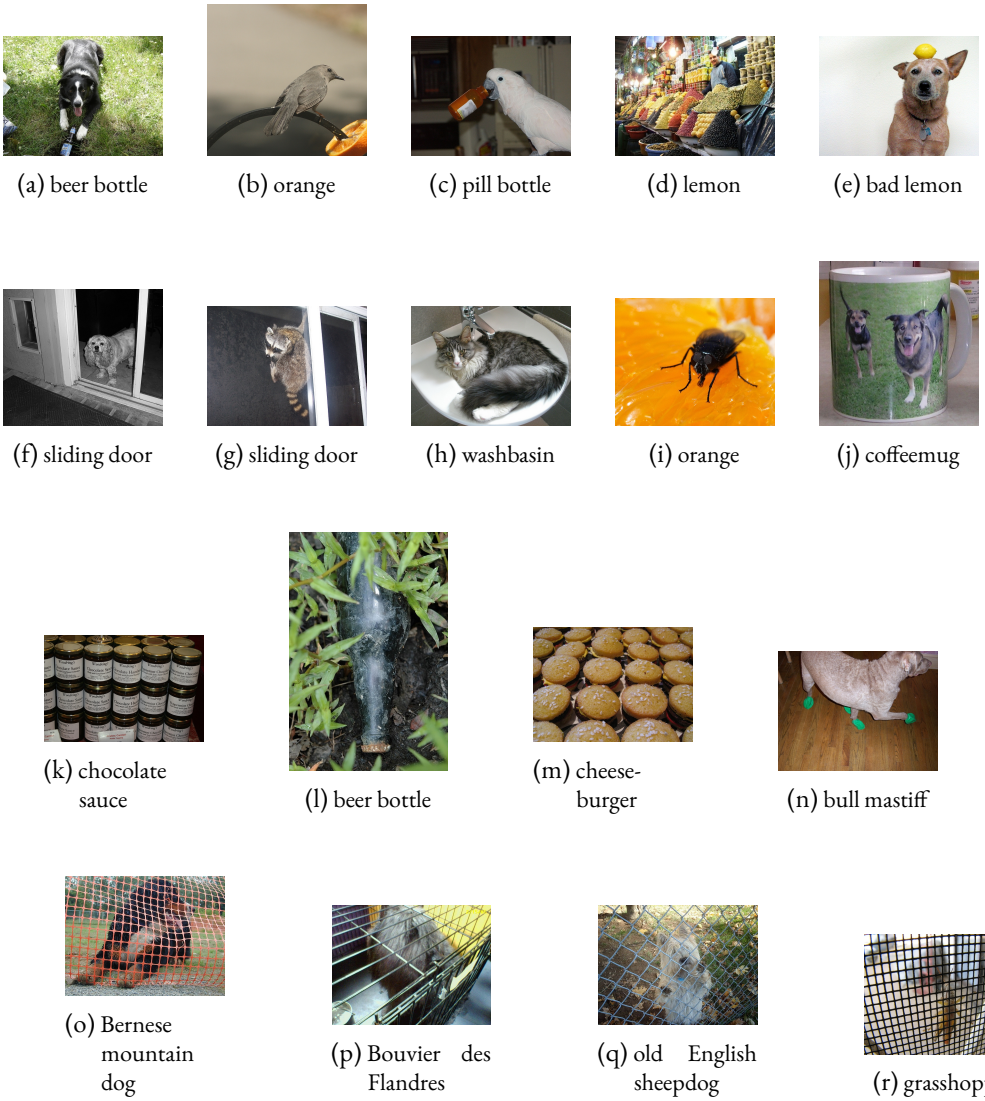
## 6.3 Correlation between support set quality and models' performance

The natural next step would be to exhibit a correlation between the quality of the support examples in a few-shot task and the performance of Few-Shot Learning models on this task.

Methodology We consider three standard Few-Shot Learning methods: Prototypical Networks (Snell et al. 2017), Transductive Information Maximization (TIM) (Boudiaf, Ziko, et al. 2020), and BD-CSPN (J. Liu et al. 2020). We apply these methods on $L2$-normalized features extracted with the ResNet12 described in Section 6.2.3, and observe the results on a testbed $\tilde{\mathcal{E}}_{\text{test}}(1)$ of 5000 5-way 1-shot tasks sampled uniformly at random from the *tiered*-ImageNet benchmark. This gives us a series $\{\mathcal{A}_i\}_1^{5000}$ of observed accuracies. On the other hand, we observe for each task $\mathbb{T}_i$ with a support set $\mathbb{S}_i$ its representativeness $\mathcal{R}_i = \text{mean}_{\boldsymbol{x}\in\mathbb{S}}\, \rho_{\boldsymbol{\theta}}(\boldsymbol{x})$, where $\boldsymbol{\theta}$ are the parameters of the aforementioned ResNet12. In the following, we exhibit a correlation between the series $\mathcal{A}$ and $\mathcal{R}$. To do so, we use Pearson's correlation which is defined as the following:

$$r(\mathcal{A}, \mathcal{R}) = \frac{\text{cov}(\mathcal{A}, \mathcal{R})}{\sigma(\mathcal{A})\sigma(\mathcal{R})}$$

with cov the covariance and $\sigma$ the standard deviation. Pearson's correlation coefficient takes values between $-1$ and $1$. $r = 0$ means that the two variables are independent, while $r = 1$ means full correlation between them.



(a) beer bottle (b) orange (c) pill bottle (d) lemon (e) bad lemon

(f) sliding door (g) sliding door (h) washbasin (i) orange (j) coffeemug

(k) chocolate sauce (l) beer bottle (m) cheese-burger (n) bull mastiff

(o) Bernese mountain dog (p) Bouvier des Flandres (q) old English sheepdog (r) grasshopper

Figure 6.4: Examples of images taken among the 100 images with worst representativeness (out of more than 200k images in *tiered*-ImageNet). Images are arranged by likely cause of their low representativeness, such that the first row (a-e) corresponds to images with a *distracting object* catching the eye at the expense of the object responsible for the image's label. The second row (f-j) corresponds to what we would call the *sliding-door effect*, as not only is there a distracting object, but also the image could not be cropped to focus on the object of interest. The third row (k-n) corresponds to a different likely cause for low representativeness: the object, albeit being without question what is described in the label, is very different from what we would expect for an image with this label (*e.g.*, chocolate sauce is usually presented in a bowl or a plate, in the context of a desert). Finally, the fourth row (o-r) corresponds to the very specific but oddly frequent issue of the object of interest being partially hidden behind a grid, which seems to distract the feature extractor.

| | $r(\mathcal{A}, \mathcal{R})$ | $r(\mathcal{A}, \{\kappa\})$ |
|---|---|---|
| ProtoNet (Snell et al. 2017) | **0.40** | 0.15 |
| BD-CSPN (J. Liu et al. 2020) | **0.34** | 0.20 |
| TIM (Boudiaf, Ziko, et al. 2020) | **0.39** | 0.18 |

Table 6.1: Pearson's correlation coefficient $r(\mathcal{A}, \mathcal{R})$ between accuracy and representativeness on 5000 5-way 1-shot tasks sampled uniformly at random from *tiered*-ImageNet. For comparison, we show the Pearson's correlation coefficient $r(\mathcal{A}, \{\kappa\})$ between accuracies and the series formed by the coarsities of the tasks, as defined in Equation 5.2.

Strong correlation between representativeness and accuracy Our experiments show that following the aforementioned methodology, we obtain a strong correlation between $\mathcal{A}$ and $\mathcal{R}$. This is exhibited in Table 6.1, as the correlation between accuracy and representativeness exceeds the correlation between accuracy and task coarsity which was proven in Section 5.6.2. To support these findings, we show in Figure 6.5 a visualization of the correlation between representativeness and accuracy.

*Ideal* support sets decisively improve performance Going further, we wanted to measure the upper bound of the improvement potential of improving the quality of the support set. To do so, we constituted an *ideal testbed* $\tilde{\mathcal{E}}^{I}_{\text{test}}(1)$ replicating the same tasks of the original 1-shot testbed $\tilde{\mathcal{E}}_{\text{test}}(1)$, but in each support set, for each class, the original support image is replaced by the image with the highest representativeness, among all images in the dataset with the same label and that do not appear in the corresponding query set. In doing so, we ensure that for each task, we provide, according to our representativeness criteria, the best available support set. In Figure 6.6, we show the consistent and decisive gain in accuracy provided by these *ideal* support sets.

## 6.4 Conclusion and Limitations

Of course, there is a tremendous bias in this study of the correlation between accuracy and representativeness, as both are based on the same feature extractor. A deeper study of these phenomena will need to consider measures of representativeness that are agnostic of the feature extractor. This is expected to be a very hard task since we are terribly dependent on learned representation models for computer vision tasks.

Another drawback of the chosen measure of representativeness is that it does not take into account the relationship between the instances and classes composing the support set. In Chapter 5, we exhibited that the similarity between the classes composing a task is of prior importance to the performance of Few-Shot Learning models. Intuitively, it would also factor in the optimal choice for support examples. In Figure 6.1a, in addition to the terrible example for the *great white shark* class, we notice that there are two very similar classes: *bathtub* and *tub*. It is understandable that based on the two available examples for these classes, it would be hard to properly define the frontier between them (it would probably be strongly biased towards the presence of feet). In this situation, the best examples would not necessarily be those that lie at the center of their respective class's cluster, but rather examples that draw the decision frontier that is closest to the ground

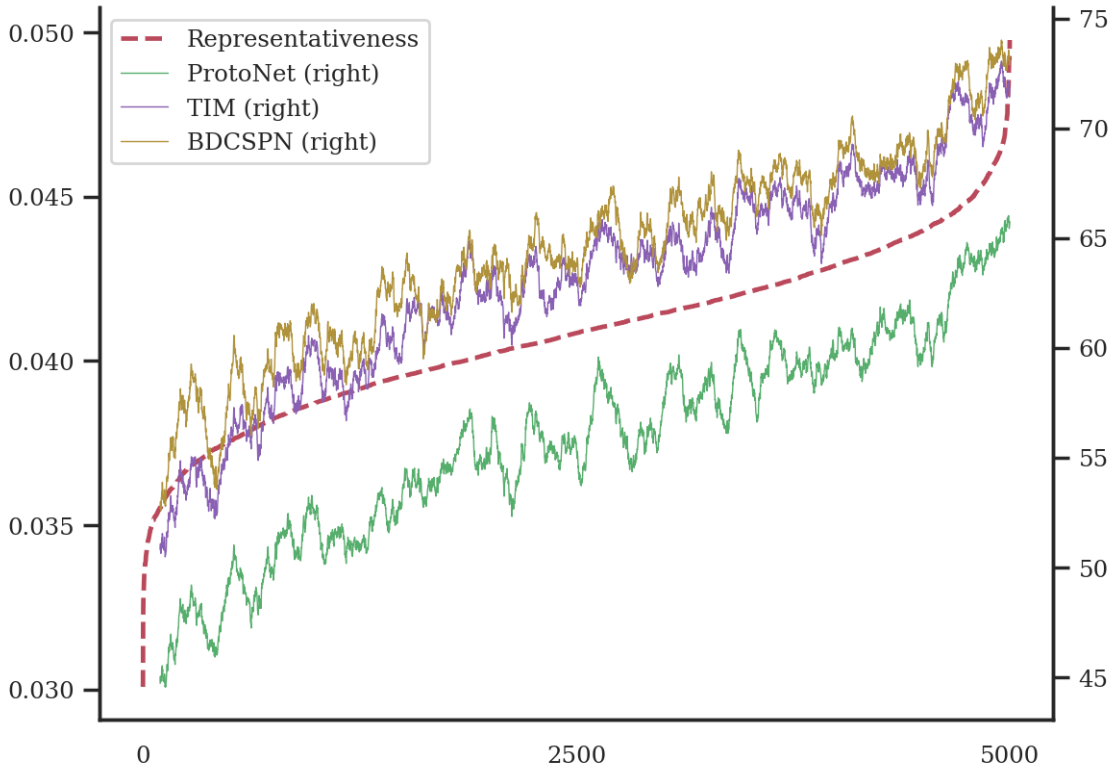
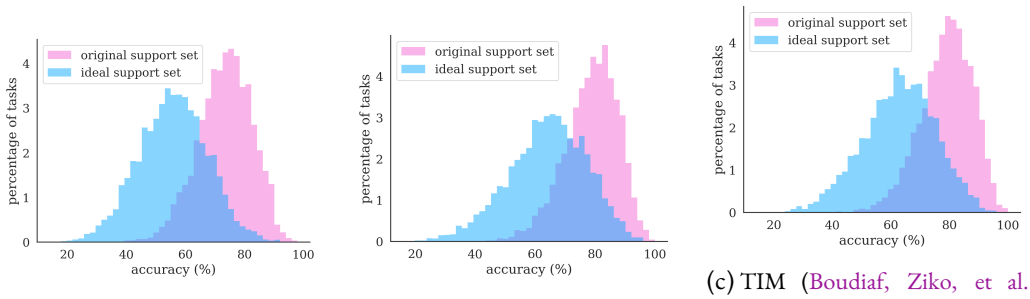



Figure 6.5: Representativeness (left axis) versus accuracies of three Few-Shot Learning models in percents (right axis). The 5000 tasks of the benchmarks are sorted by order of increasing representativeness. The accuracies are smoothed over a rolling window of 100 tasks.




(a) ProtoNet (Snell et al. 2017). Average accuracy gain: 18%

(b) BD-CSPN (J. Liu et al. 2020). Average accuracy gain: 15%

(c) TIM (Boudiaf, Ziko, et al. 2020). Average accuracy gain: 17%

Figure 6.6: Histogram of accuracies on the original testbed $\tilde{\mathcal{E}}_{\text{test}}(1)$ and the *ideal* testbed $\tilde{\mathcal{E}}^{I}_{\text{test}}(1)$ for three distinct methods, using a ResNet12. For all methods, switching to the ideal testbed mitigates the error rate by more than 40%.

truth. Therefore, we postulate that future works focusing on the quality of the support set will need to provide a measure of representativeness *in the context of a given support set*.

Nevertheless, this study shows that the selection of support examples is a decisive factor in the performance of Few-Shot Learning models. Even though the *ideal* support examples were selected through a highly illegal process, we cannot ignore that selecting the right examples can mitigate the error rate of Few-Shot Learning models by more than 40% in the 1-shot scenario.

# 7 Conclusion and perspectives

## 7.1 Looking back at the contributions

In this thesis, we identified limitations in the current formulation and evaluation processes in Few-Shot Image Classification and proposed countermeasures to mitigate these limitations. Let us review these contributions chronologically:

1. In 2021, we introduced the problem of Few-Shot Learning under Support-Query Shift. We proposed `FewShiftBed`, a novel benchmark to evaluate Few-Shot Learning methods under Support-Query Shift. We offered a first solution to the problem by combining Prototypical Networks with Optimal Transport. This contribution was reported in Chapter 3. Note that `FewShiftBed` was built for episodic training. Indeed, we put a lot of effort into building benchmarks that offered a wide variety of classes *and* domains both in the test set, training set, and validation set, which is necessary to *learn to learn on new domains* on the training set. However, we did not find any evidence that episodic training improved the models' performance compared to standard empirical risk minimization.

2. Then in 2022, we provided a quantitative and qualitative study of the test tasks generated from ImageNet-derived benchmarks and showed that these benchmarks are heavily biased towards tasks composed of semantically unrelated classes. We harnessed the WordNet taxonomy to enforce the sampling of fine-grained tasks from *tiered*ImageNet in order to re-balance this benchmark. This contribution was reported in Chapter 5. It is worth mentioning that the strategy we exposed in this chapter to sample semantically fine-grained tasks was originally meant to be used during episodic training, with the intuition that training on finer-grained tasks would improve predictions. However, the performance only improved marginally compared to either standard episodic training or standard empirical risk minimization, and therefore this did not lead to a contribution.

3. In 2023, we tackled the recent Few-Shot Open-Set Recognition problem using pre-trained models and transductive inference. Both our solutions set a new milestone in model-agnostic Few-Shot Learning. Indeed, we did not make any assumption on the backbone, its architecture, or its training process, and focus on the *inference strategy*. In doing so, we were able to report a large and steady improvement over the baselines on a wide variety of base models, without any re-training or re-tuning of the hyperparameters. This last published contribution was reported in Chapter 4.

4. Finally, in the winter of 2022-23, we pursued investigations on what makes the quality of a support set. In this study, we assume representations extracted by a given pre-trained model, without any assumption on its architecture and training procedure. These investigations are detailed in Chapter 6.

As such, the story of this thesis can be seen as the story of the transition, from a Few-Shot Learning research limited to the scope of "meta-learning", towards a Few-Shot Learning research focusing on the inference, leaving the base training to the neighboring field of representation learning. This corresponds to an emerging trend, following an increasing number of studies that raise concern about the theoretical foundations of episodic training for classification tasks and question its empirical validation (Antoniou, Edwards, et al. 2019; Bennequin 2019; Laenen and Bertinetto 2021).

## 7.2 What we did not do

Mix between large-scale and few-shot classes Looking back at the use cases presented in Section 1.2, we addressed, be it partially, all the gaps that we observed between those use cases and the standardized academic setting for Few-Shot Learning. Except for one: the mix between large-scale and few-shot classes *i.e.*, the possibility that some classes may have a large number of labeled examples. This is close to a problem addressed, with different perspectives and hypotheses, in Generalized Few-Shot Learning and Long-Tail Recognition.

- Generalized Few-Shot Learning (Ye et al. 2020) is, quite transparently, a generalization of the Few-Shot Learning setting. In this generalization, we break the assumption that the classes of a few-shot task are entirely disjoint from the classes in the base set: some base classes may *come back*. Depending on the chosen setting, we may or may not assume that we still have access to the entire base set.

- Long-Tail Recognition (Y. Zhang, Kang, et al. 2023) studies a specification of the standard visual recognition problems in which the number of images in the classes follows a long-tailed distribution *i.e.*, some classes have many examples and many classes have very few examples.

The prison of transductivity The advertised goal of this thesis was to move Few-Shot Learning research towards more realistic settings, in order to make it more relevant to industrial applications. However, in doing so, we had to resolve to a very strong assumption: *transduction*. In both Chapters 3 and 4, our proposed methods rely on transduction *i.e.*, performing prediction on batches of queries at a time. The reader might notice that this assumption holds in exactly zero of the use cases for few-shot image classification presented in Section 1.2. Revisiting our contributions, we motivated this choice by the increased complexity resulting from the relaxation of other assumptions. Indeed, in the context of Support-Query Shift, it seems delusive to try and adapt to a target distribution empirically defined by one point. The same problem arises in Few-Shot Open-Set Recognition: it would be very hard (and has been confirmed as such by previous studies) to detect that a single point deviates from class distributions empirically defined by only one example each. We were evidently unable to solve these difficult problems in an inductive fashion. We believe that this motivates very challenging future research. In Section 4.2, we linked the difficulty of Few-Shot Open-Set Recognition to the poor "quality" of the class distributions in the representation space (considered through the proxy of the integrity of class clusters) in a few-shot scenario. We believe that improving this representation space is of prior importance to

the success of Few-Shot Learning models, especially in the more challenging scenarios considered in this thesis.

Thorough study of relevant metrics In Chapter 5, we started to reflect on the metrics used to evaluate Few-Shot Learning models, departing from the current idea that the top-1 accuracy on uniformly sampled 1-shot and 5-shot 5-way tasks gives a good enough measure of the value of a model. In addition to a new way of sampling test tasks, we propose a first benchmark including 100-way tasks, for which we report both the top-1 and top-5 accuracy. The results shown in Table 5.2 raise strong concerns about the ability of current Few-Shot Learning models to solve tasks that are harder than discriminating between 5 classes. However, we believe this study is insufficient, and that a lot remains to be done to standardize more relevant benchmarks and metrics for Few-Shot Learning. Indeed, not only is this study incomplete (why 100 classes? why still 1 and 5 shots? couldn't we replace these arbitrary settings with aggregated benchmarks?), it failed to arouse the interest of the community and convince our fellow researchers of the need for more realistic benchmarks.

Breaking the walls between sub-domains A persistent theme of this thesis is the study of settings made by the combination of constraints, such as Few-Shot Learning *under Support-Query Shift*, or Few-Shot *Open-Set* Recognition. Our next study could very well be about Few-Shot Open-Set Recognition under Support-Query Shift. It would be relevant to the e-shopping use case described in Section 1.2. Real-world use cases are not limited to one such constraint, nor are they limited to two. All the combinations of such constraints can be thought of, and it is likely that there currently is a real industrial use case for many of them. Should it motivate us to build a specific area of research for each of them, with a siloed research community, specific benchmarks, and convoluted methods? Should there be a thesis titled *Semi-Supervised Fine-Grained Incremental Few-Shot Open-Set Recognition under Three Out of Four Possible Kinds of Distribution Shift*?

In this thesis, we studied interesting openings of the Few-Shot Learning assumptions but mostly stayed inside of the Few-Shot Learning landscape, comparing to its specific methods on its homemade benchmarks. Yet our main finding seems to be that the convoluted methods and training schemes designed specifically for Few-Shot Learning do not provide much value, and we ended up focusing on the best way to harness a pre-trained feature extractor. In the end, we consider Few-Shot Learning, as well as Few-Shot Open-Set Recognition, as a downstream task of representation learning. This being said it would seem more relevant to join our efforts with the communities that currently study the neighboring *few-data* and *few-label* problems.

## 7.3 The future of Few-Shot Learning

As we reach the end of this thesis, our strong intuition is that future works in Few-Shot Learning need to study more general problems than Few-Shot Learning. Our findings in Chapter 4 left us with the impression that the specific methods designed for Few-Shot Learning brought little value, compared to the performance that can be achieved by plugging simple, principled models on top of the best foundation models obtained through the latest advances in representation learning.

We believe that the natural next step would be a thorough empirical study evaluating the respective importance of the feature extractor compared to the chosen Few-Shot Learning method. This study would surely be a very difficult one, as for fair comparison one would need to scale up all Few-Shot Learning methods, including those based on episodic training, to combine them with very large models, even though we know episodic training to be very hard to tune, even on small models. We still believe that this study is necessary to rid the Few-Shot Learning research community of its misconception and allow us to move forward.

We would then be able to leverage all the latest advances in representation learning to improve performance on a wide variety of small-data tasks, including the current Few-Shot Learning setting. Free from the burden of large-scale training of foundations models, we could focus on the development of simple, principled methods to solve some or all of the small-data tasks, given a pre-trained feature extractor.

# Bibliography


Achille, A., M. Lam, R. Tewari, A. Ravichandran, S. Maji, C. C. Fowlkes, S. Soatto, and P. Perona (2019). "Task2vec: Task embedding for meta-learning". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6430–6439.

Aimen, A., B. Ladrecha, and N. C. Krishnan (2023). "Adversarial Projections to Tackle Support-Query Shifts in Few-Shot Meta-Learning". In: *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2022, Grenoble, France, September 19–23, 2022, Proceedings, Part III*. Springer, pp. 615–630.

Alves, G., M. Couceiro, and A. Napoli (2020). "Sélection de mesures de similarité pour les données catégorielles". In: *EGC 2020-20ème édition de la conférence Extraction et Gestion des Connaissances*.

Amodei, D., C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané (2016). "Concrete problems in AI safety". In: *arXiv preprint arXiv:1606.06565*.

Antoniou, A., H. Edwards, and A. Storkey (2019). "How to train your MAML". In: *International Conference on Learning Representations (ICLR)*.

Antoniou, A., A. Storkey, and H. Edwards (2017). "Data augmentation generative adversarial networks". In: *arXiv preprint arXiv:1711.04340*.

Antoniou, A. and A. J. Storkey (2019). "Learning to learn by self-critique". In: *Advances in Neural Information Processing Systems*, pp. 9940–9950.

Basu, S., M. Stanley, J. F. Bronskill, S. Feizi, and D. Massiceti (2023). "Hard-Meta-Dataset++: Towards Understanding Few-Shot Performance on Difficult Tasks". In: *The Eleventh International Conference on Learning Representations*.

Ben-David, S., J. Blitzer, K. Crammer, and F. Pereira (2007). "Analysis of representations for domain adaptation". In: *Advances in neural information processing systems*, pp. 137–144.

Bendale, A. and T. E. Boult (2016). "Towards open set deep networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1563–1572.

Bendou, Y., L. Drumetz, V. Gripon, G. Lioi, and B. Pasdeloup (2022). "Le manchot, la banane et la bibliothèque...(de la désambiguisation d'une tâche de classification avec un exemple)". In: *GRETSI 2022*.

Bengio, Y., A. Courville, and P. Vincent (2013). "Representation learning: A review and new perspectives". In: *IEEE transactions on pattern analysis and machine intelligence* 35.8, pp. 1798–1828.

Bengio, Y., J. Louradour, R. Collobert, and J. Weston (2009). "Curriculum learning". In: *Proceedings of the 26th annual international conference on machine learning*, pp. 41–48.

Bennequin, E. (2019). "Meta-learning algorithms for few-shot computer vision". In: *arXiv*.

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

Bennequin, E., M. Tami, A. Toubhans, and C. Hudelot (2022). "Few-Shot Image Classification Benchmarks are Too Far From Reality: Build Back Better with Semantic Task Sampling". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4767–4776.

Bertinetto, L., J. F. Henriques, P. Torr, and A. Vedaldi (2019). "Meta-learning with differentiable closed-form solvers". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=HyxnZh0ct7.

Bhushan Damodaran, B., B. Kellenberger, R. Flamary, D. Tuia, and N. Courty (2018). "Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation". In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 447–463.

Boudiaf, M., E. Bennequin, M. Tami, C. Hudelot, A. Toubhans, P. Piantanida, and I. B. Ayed (2022). "Model-Agnostic Few-Shot Open-Set Recognition". In: *arXiv preprint arXiv:2206.09236*.

Boudiaf, M., E. Bennequin, M. Tami, A. Toubhans, P. Piantanida, C. Hudelot, and I. Ben Ayed (2023). "Open-Set Likelihood Maximization for Few-Shot Learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24007–24016.

Boudiaf, M., I. Ziko, J. Rony, J. Dolz, P. Piantanida, and I. Ben Ayed (2020). "Information maximization for few-shot learning". In: *Advances in Neural Information Processing Systems* 33, pp. 2445–2457.

Bouvier, V., P. Very, C. Chastagnol, M. Tami, and C. Hudelot (2020). "Robust Domain Adaptation: Representations, Weights and Inductive Bias". In: *ECML*.

Bronskill, J., J. Gordon, J. Requeima, S. Nowozin, and R. Turner (2020). "Tasknorm: Rethinking batch normalization for meta-learning". In: *ICML*. PMLR, pp. 1153–1164.

Caldas, S., S. M. K. Duddu, P. Wu, T. Li, J. Konečnỳ, H. B. McMahan, V. Smith, and A. Talwalkar (2018). "Leaf: A benchmark for federated settings". In: *arXiv preprint arXiv:1812.01097*.

Cao, Z., L. Ma, M. Long, and J. Wang (2018). "Partial adversarial domain adaptation". In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 135–150.

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

Chapelle, O., B. Scholkopf, and A. Zien (2009). "Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]". In: *IEEE Transactions on Neural Networks* 20.3, pp. 542–542.

Cheek, M., E. Nic Lughadha, P. Kirk, H. Lindon, J. Carretero, B. Looney, B. Douglas, D. Haelewaters, E. Gaya, T. Llewellyn, et al. (2020). "New discoveries: plants and fungi. Plants". In: *People Planet*.

Chen, T., S. Kornblith, M. Norouzi, and G. Hinton (2020). "A simple framework for contrastive learning of visual representations". In: *International conference on machine learning*. PMLR, pp. 1597–1607.

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

Chen, X., C.-J. Hsieh, and B. Gong (2022). "When vision transformers outperform ResNets without pre-training or strong data augmentations". In: *International Conference on Learning Representations (ICLR)*.

Chen, X., S. Wang, M. Long, and J. Wang (2019). "Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation". In: *International Conference on Machine Learning*, pp. 1081–1090.

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

Cohen, G., S. Afshar, J. Tapson, and A. Van Schaik (2017). "EMNIST: Extending MNIST to handwritten letters". In: *IJCNN*. IEEE.

Courty, N., R. Flamary, A. Habrard, and A. Rakotomamonjy (2017). "Joint distribution optimal transportation for domain adaptation". In: *Advances in Neural Information Processing Systems*, pp. 3730–3739.

Courty, N., R. Flamary, D. Tuia, and A. Rakotomamonjy (2016). "Optimal transport for domain adaptation". In: *IEEE transactions on pattern analysis and machine intelligence* 39.9, pp. 1853–1865.

Cuturi, M. (2013). "Sinkhorn distances: Lightspeed computation of optimal transport". In: *Advances in neural information processing systems* 26, pp. 2292–2300.

Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei (2009). "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee, pp. 248–255.

Deselaers, T. and V. Ferrari (2011). "Visual and semantic similarity in imagenet". In: *CVPR 2011*. IEEE, pp. 1777–1784.

Dhillon, G. S., P. Chaudhari, A. Ravichandran, and S. Soatto (2020). "A Baseline for Few-Shot Image Classification". In: *ICLR*.

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

Du, Y., X. Zhen, L. Shao, and C. G. M. Snoek (2021). "MetaNorm: Learning to Normalize Few-Shot Batches Across Domains". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=9z_dNsC4B5t.

Dumoulin, V., N. Houlsby, U. Evci, X. Zhai, R. Goroshin, S. Gelly, and H. Larochelle (2021). "A unified few-shot classification benchmark to compare transfer and meta learning approaches". In: *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Fei, N., Z. Lu, T. Xiang, and S. Huang (2020). "Melr: Meta-learning via modeling episode-level relationships for few-shot learning". In: *International Conference on Learning Representations*.

Finn, C., P. Abbeel, and S. Levine (2017). "Model-agnostic meta-learning for fast adaptation of deep networks". In: *International conference on machine learning*. PMLR, pp. 1126–1135.

Fukushima, K. (1980). "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position". In: *Biological cybernetics* 36.4, pp. 193–202.

Ganin, Y. and V. Lempitsky (2015). "Unsupervised Domain Adaptation by Backpropagation". In: *International Conference on Machine Learning*, pp. 1180–1189.

Ge, Z., S. Demyanov, Z. Chen, and R. Garnavi (2017). "Generative openmax for multi-class open set classification". In: *arXiv preprint arXiv:1707.07418*.

Goldblum, M., L. Fowl, and T. Goldstein (2020). "Adversarially robust few-shot learning: A meta-learning approach". In: *Advances in Neural Information Processing Systems* 33, pp. 17886–17895.

Goldblum, M., S. Reich, L. Fowl, R. Ni, V. Cherepanova, and T. Goldstein (2020). "Unraveling meta-learning: Understanding feature representations for few-shot tasks". In: *International Conference on Machine Learning*. PMLR, pp. 3607–3616.

Grandvalet, Y. and Y. Bengio (2005). "Semi-supervised learning by entropy minimization". In: *Advances in neural information processing systems*, pp. 529–536.

Gulrajani, I. and D. Lopez-Paz (2021). "In Search of Lost Domain Generalization". In: *International Conference on Learning Representations*.

Hadsell, R., S. Chopra, and Y. LeCun (2006). "Dimensionality reduction by learning an invariant mapping". In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*. Vol. 2. IEEE, pp. 1735–1742.

Hariharan, B. and R. Girshick (2017). "Low-shot visual recognition by shrinking and hallucinating features". In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3018–3027.

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

Hendrycks, D. and T. Dietterich (2019). "Benchmarking Neural Network Robustness to Common Corruptions and Perturbations". In: *ICLR*.

Hiller, M., R. Ma, M. Harandi, and T. Drummond (2022). "Rethinking generalization in few-shot classification". In: *Advances in Neural Information Processing Systems* 35, pp. 3582–3595.

Hu, Y., V. Gripon, and S. Pateux (2021). "Leveraging the feature distribution in transfer-based few-shot learning". In: *International Conference on Artificial Neural Networks*. Springer, pp. 487–499.

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

Ioffe, S. and C. Szegedy (2015). "Batch normalization: Accelerating deep network training by reducing internal covariate shift". In: *ICML*. PMLR.

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

Jiang, J. J. and D. W. Conrath (1997). "Semantic similarity based on corpus statistics and lexical taxonomy". In: *arXiv preprint cmp-lg/9709008*.

Jiang, S., W. Ding, H.-W. Chen, and M.-S. Chen (2022). "Pgada: perturbation-guided adversarial alignment for few-shot learning under the support-query shift". In: *Advances in Knowledge Discovery and Data Mining: 26th Pacific-Asia Conference, PAKDD 2022, Chengdu, China, May 16–19, 2022, Proceedings, Part I*. Springer, pp. 3–15.

Jiang, Z., B. Kang, K. Zhou, and J. Feng (2020). "Few-shot classification via adaptive attention". In: *arXiv preprint arXiv:2008.02465*.

Kaddour, J., S. Sæmundsson, et al. (2020). "Probabilistic active meta-learning". In: *Advances in Neural Information Processing Systems* 33, pp. 20813–20822.

Killamsetty, K., X. Zhao, F. Chen, and R. Iyer (2021). "RETRIEVE: Coreset Selection for Efficient and Robust Semi-Supervised Learning". In: *Neural Information Processing Systems (NeurIPS)* 34.

Kingma, D. P., S. Mohamed, D. Jimenez Rezende, and M. Welling (2014). "Semi-supervised learning with deep generative models". In: *Advances in neural information processing systems* 27.

Knight, P. A. (2008). "The Sinkhorn–Knopp algorithm: convergence and applications". In: *SIAM Journal on Matrix Analysis and Applications* 30.1, pp. 261–275.

Koch, G., R. Zemel, and R. Salakhutdinov (2015). "Siamese neural networks for one-shot image recognition". In: *ICML deep learning workshop*. Vol. 2.

Krizhevsky, A., G. Hinton, et al. (2009). "Learning multiple layers of features from tiny images". In.

Krizhevsky, A., I. Sutskever, and G. E. Hinton (2012). "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems* 25.

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

Laine, S. and T. Aila (2017). "Temporal Ensembling for Semi-Supervised Learning". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=BJ6oOfqge.

Lake, B., R. Salakhutdinov, J. Gross, and J. Tenenbaum (2011). "One shot learning of simple visual concepts". In: *Proceedings of the annual meeting of the cognitive science society*. Vol. 33.

Lake, B. M., R. Salakhutdinov, and J. B. Tenenbaum (2015). "Human-level concept learning through probabilistic program induction". In: *Science* 350.6266, pp. 1332–1338.

Lee, D.-H. et al. (2013). "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks". In: *Workshop on challenges in representation learning, ICML*. Vol. 3. 2, p. 896.

Li, J., Z. Wang, and X. Hu (2021). "Learning intact features by erasing-inpainting for few-shot classification". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 35. 9, pp. 8401–8409.

Li, Z., Y. Zhao, N. Botta, C. Ionescu, and X. Hu (2020). "COPOD: copula-based outlier detection". In: *International Conference on Data Mining (ICDM)*.

Liang, J., D. Hu, and J. Feng (2020). "Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation". In: *International conference on machine learning*. PMLR, pp. 6028–6039.

Liu, B., H. Kang, H. Li, G. Hua, and N. Vasconcelos (2020). "Few-shot open-set recognition using meta-learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8798–8807.

Liu, C., Z. Wang, D. Sahoo, Y. Fang, K. Zhang, and S. C. Hoi (2020). "Adaptive Task Sampling for Meta-learning". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*. Springer, pp. 752–769.

Liu, F. T., K. M. Ting, and Z.-H. Zhou (2008). "Isolation forest". In: *International Conference on Data Mining (ICMD)*.

Liu, H., M. Long, J. Wang, and M. Jordan (2019). "Transferable Adversarial Training: A General Approach to Adapting Deep Classifiers". In: *International Conference on Machine Learning*, pp. 4013–4022.

Liu, J., L. Song, and Y. Qin (2020). "Prototype rectification for few-shot learning". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, pp. 741–756.

Liu, L., T. Zhou, G. Long, J. Jiang, L. Yao, and C. Zhang (Aug. 2019). "Prototype Propagation Networks (PPN) for Weakly-supervised Few-shot Learning on Category Graph". In: pp. 3015–3022. DOI: 10.24963/ijcai.2019/418.

Liu, Y., J. Lee, M. Park, S. Kim, E. Yang, S. J. Hwang, and Y. Yang (2019). "Learning to propagate labels: Transductive propagation network for few-shot learning". In: *ICLR*.

Long, M., Y. Cao, J. Wang, and M. I. Jordan (2015). "Learning transferable features with deep adaptation networks". In: *Proceedings of the 32nd International Conference on International Conference on Machine Learning-Volume 37*. JMLR. org, pp. 97–105.

Long, M., Z. Cao, J. Wang, and M. I. Jordan (2018). "Conditional adversarial domain adaptation". In: *Advances in Neural Information Processing Systems*, pp. 1640–1650.

Lu, J., S. Jin, J. Liang, and C. Zhang (2020). "Robust few-shot learning for user-provided data". In: *IEEE transactions on neural networks and learning systems* 32.4, pp. 1433–1447.

Luo, X., L. Wei, L. Wen, J. Yang, L. Xie, Z. Xu, and Q. Tian (2021). "Rectifying the shortcut learning of background for few-shot learning". In: *Advances in Neural Information Processing Systems* 34, pp. 13073–13085.

Maji, S., J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi (2013). *Fine-Grained Visual Classification of Aircraft*. Tech. rep. arXiv: 1306.5151 [cs-cv].

Miller, G. A. (1995). "WordNet: a lexical database for English". In: *Communications of the ACM* 38.11, pp. 39–41.

Miyato, T., S.-i. Maeda, M. Koyama, and S. Ishii (2018). "Virtual adversarial training: a regularization method for supervised and semi-supervised learning". In: *IEEE transactions on pattern analysis and machine intelligence* 41.8, pp. 1979–1993.

Munkhdalai, T. and H. Yu (2017). "Meta networks". In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, pp. 2554–2563.

Nado, Z., S. Padhy, D. Sculley, A. D'Amour, B. Lakshminarayanan, and J. Snoek (2020). "Evaluating prediction-time batch normalization for robustness under covariate shift". In: *arXiv preprint arXiv:2006.10963*.

Neal, L., M. Olson, X. Fern, W.-K. Wong, and F. Li (2018). "Open set learning with counterfactual images". In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 613–628.

Nguyen, C., T.-T. Do, and G. Carneiro (2021). "Similarity of Classification Tasks". In: *arXiv preprint arXiv:2101.11201*.

Nichol, A., J. Achiam, and J. Schulman (2018). *On First-Order Meta-Learning Algorithms*. arXiv: 1803.02999 [cs.LG].

Odena, A. (2016). "Semi-supervised learning with generative adversarial networks". In: *arXiv preprint arXiv:1606.01583*.

Oord, A. v. d., Y. Li, and O. Vinyals (2018). "Representation learning with contrastive predictive coding". In: *arXiv preprint arXiv:1807.03748*.

Ouali, Y., V. Bouvier, M. Tami, and C. Hudelot (2020). "Target Consistency for Domain Adaptation: when Robustness meets Transferability". In: *arXiv preprint arXiv:2006.14263*.

Ouali, Y., C. Hudelot, and M. Tami (2021). "Spatial contrastive learning for few-shot classification". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, pp. 671–686.

Pan, S. J. and Q. Yang (2009). "A survey on transfer learning". In: *IEEE Transactions on knowledge and data engineering* 22.10, pp. 1345–1359.

Peyré, G. et al. (2019). "Computational Optimal Transport: With Applications to Data Science". In: *Foundations and Trends® in Machine Learning* 11.5-6, pp. 355–607.

Picek, L., M. Šulc, J. Matas, T. S. Jeppesen, J. Heilmann-Clausen, T. Læssøe, and T. Frøslev (2022). "Danish Fungi 2020-Not Just Another Image Recognition Dataset". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1525–1535.

Quionero-Candela, J., M. Sugiyama, A. Schwaighofer, and N. D. Lawrence (2009). *Dataset shift in machine learning*. The MIT Press.

Ramalho, T. and M. Garnelo (2019). "Adaptive Posterior Learning: few-shot learning with a surprise-based memory module". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=ByeSdsC9Km.

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

Ravi, S. and H. Larochelle (2017). "Optimization as a Model for Few-Shot Learning". In: *5th International Conference on Learning Representations*.

Ren, M., E. Triantafillou, S. Ravi, J. Snell, K. Swersky, J. B. Tenenbaum, H. Larochelle, and R. S. Zemel (2019). "Meta-learning for semi-supervised few-shot classification". In: *ICLR*.

Ridnik, T., E. Ben-Baruch, A. Noy, and L. Zelnik-Manor (2021). "Imagenet-21k pretraining for the masses". In: *Neural Information Processing Systems (NeurIPS)*.

Roy, V., Y. Xu, Y.-X. Wang, K. Kitani, R. Salakhutdinov, and M. Hebert (2020). "Few-shot learning with intra-class knowledge transfer". In: *arXiv preprint arXiv:2008.09892*.

Ruan, X., G. Lin, C. Long, and S. Lu (2021). "Few-shot fine-grained classification with Spatial Attentive Comparison". In: *Knowledge-Based Systems* 218, p. 106840.

Russakovsky, O., J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. (2015). "Imagenet large scale visual recognition challenge". In: *International journal of computer vision* 115.3, pp. 211–252.

Sahoo, D., H. Le, C. Liu, and S. C. H. Hoi (2019). *Meta-Learning with Domain Adaptation for Few-Shot Learning under Domain Shift*.

Saito, K., D. Kim, and K. Saenko (2021). "OpenMatch: Open-Set Semi-supervised Learning with Open-set Consistency Regularization". In: *Neural Information Processing Systems (NeurIPS)*.

Sbai, O., C. Couprie, and M. Aubry (2020). "Impact of base dataset design on few-shot image classification". In: *European Conference on Computer Vision*. Springer, pp. 597–613.

Scheirer, W. J., A. de Rezende Rocha, A. Sapkota, and T. E. Boult (2012). "Toward open set recognition". In: *PAMI*.

Schneider, S., E. Rusak, L. Eck, O. Bringmann, W. Brendel, and M. Bethge (2020). "Improving robustness against common corruptions by covariate shift adaptation". In: *Advances in Neural Information Processing Systems* 33.

Schölkopf, B., J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson (2001). "Estimating the support of a high-dimensional distribution". In: *Neural computation*.

Schroeder, B. and Y. Cui (2018). "Fgvcx fungi classification challenge 2018". In: *Available online: github. com/visipedia/fgvcx_fungi_comp (accessed on 14 July 2021)*.

Shyu, M.-L., S.-C. Chen, K. Sarinnapakorn, and L. Chang (2003). *A novel anomaly detection scheme based on principal component classifier*. Tech. rep. Miami Univ Coral Gables Fl Dept of Electrical and Computer Engineering.

Snell, J., K. Swersky, and R. Zemel (2017). "Prototypical networks for few-shot learning". In: *Advances in neural information processing systems* 30.

Sohn, K., D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li (2020). "Fixmatch: Simplifying semi-supervised learning with consistency and confidence". In: *Advances in neural information processing systems* 33, pp. 596–608.

Sun, Q., Y. Liu, T.-S. Chua, and B. Schiele (2019). "Meta-transfer learning for few-shot learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 403–412.

Sun, Y., X. Wang, Z. Liu, J. Miller, A. Efros, and M. Hardt (2020). "Test-time training with self-supervision for generalization under distribution shifts". In: *ICML*.

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

Tachet des Combes, R., H. Zhao, Y.-X. Wang, and G. J. Gordon (2020). "Domain adaptation with conditional distribution matching and generalized label shift". In: *Advances in Neural Information Processing Systems* 33, pp. 19276–19289.

Tang, L., D. Wertheimer, and B. Hariharan (2020). "Revisiting pose-normalization for fine-grained few-shot recognition". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14352–14361.

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

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

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

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

Veilleux, O., M. Boudiaf, P. Piantanida, and I. Ben Ayed (2021). "Realistic evaluation of transductive few-shot learning". In: *Neural Information Processing Systems (NeurIPS)*.

Vinyals, O., C. Blundell, T. Lillicrap, D. Wierstra, et al. (2016). "Matching networks for one shot learning". In: *Advances in neural information processing systems* 29, pp. 3630–3638.

Wang, D., E. Shelhamer, S. Liu, B. Olshausen, and T. Darrell (2021). "Fully test-time adaptation by entropy minimization". In: *ICLR*.

Wang, Y.-X., R. Girshick, M. Hebert, and B. Hariharan (2018). "Low-shot learning from imaginary data". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7278–7286.

Wang, Y., W.-L. Chao, K. Q. Weinberger, and L. van der Maaten (2019). "Simpleshot: Revisiting nearest-neighbor classification for few-shot learning". In: *arXiv*.

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

Welinder, P., S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona (2010). *Caltech-UCSD Birds 200*. Tech. rep. CNS-TR-2010-001. California Institute of Technology.

Wightman, R. (2019). *PyTorch Image Models*. https://github.com/rwightman/pytorch-image-models. DOI: 10.5281/zenodo.4414861.

Wu, Z., Y. Xiong, S. X. Yu, and D. Lin (2018). "Unsupervised feature learning via non-parametric instance discrimination". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3733–3742.

Xu, J., H. Le, M. Huang, S. Athar, and D. Samaras (2021). "Variational Feature Disentangling for Fine-Grained Few-Shot Classification". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8812–8821.

Yalniz, I. Z., H. Jégou, K. Chen, M. Paluri, and D. Mahajan (2019). "Billion-scale semi-supervised learning for image classification". In: *arXiv*.

Yang, B., C. Liu, B. Li, J. Jiao, and Q. Ye (2020). "Prototype mixture models for few-shot semantic segmentation". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*. Springer, pp. 763–778.

Yang, S., L. Liu, and M. Xu (2021). "Free Lunch for Few-shot Learning: Distribution Calibration". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=JWOiYxMG92s.

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

Yeh, H.-W., B. Yang, P. C. Yuen, and T. Harada (2021). "SoFA: Source-Data-Free Feature Alignment for Unsupervised Domain Adaptation". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 474–483.

You, K., M. Long, Z. Cao, J. Wang, and M. I. Jordan (2019). "Universal domain adaptation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2720–2729.

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

Zhang, M., H. Marklund, A. Gupta, S. Levine, and C. Finn (2021). "Adaptive Risk Minimization: A Meta-Learning Approach for Tackling Group Shift". In: *ICLR*.

Zhang, Y., B. Kang, B. Hooi, S. Yan, and J. Feng (2023). "Deep long-tailed learning: A survey". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhang, Y., T. Liu, M. Long, and M. Jordan (2019). "Bridging theory and algorithm for domain adaptation". In: *International conference on machine learning*. PMLR, pp. 7404–7413.

Zhao, A., M. Ding, Z. Lu, T. Xiang, Y. Niu, J. Guan, and J.-R. Wen (2021). "Domain-adaptive few-shot learning". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1390–1399.

Zhao, Y., Z. Nasrullah, and Z. Li (2019). "PyOD: A Python Toolbox for Scalable Outlier Detection". In: *JMLR*.

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

Zhu, Y., C. Liu, and S. Jiang (2020). "Multi-attention Meta Learning for Few-shot Fine-grained Image Recognition." In: *IJCAI*, pp. 1090–1096.

Zieliński, B., A. Plichta, K. Misztal, P. Spurek, M. Brzychczy-Włoch, and D. Ochońska (2017). "Deep learning approach to bacterial colony classification". In: *PloS One* 12.9, e0184554.

Ziko, I., J. Dolz, E. Granger, and I. B. Ayed (2020). "Laplacian regularized few-shot learning". In: *International Conference on Machine Learning*. PMLR, pp. 11660–11670.

# Extended experimental results on FewShiftBed

In this section we present the extended results of our experiments. Prototypical Networks, Matching Networks and Transductive Propagation Networks have been declined in 10 distinct versions:

- Original algorithms: **episodic training**, with Conventional Batch-Normalization (**CBN**) and not Optimal Transport (**Vanilla**);
- **Episodic training** and **CBN**, with Optimal Transport applied at test time (**OT-TT**);
- **Episodic training** and **CBN**, with Optimal Transport integrated into the algorithm both during training and testing (**OT**);
- **Episodic training**, with Transductive Batch-Normalization (**TBN**) and not Optimal Transport (**Vanilla**);
- **Episodic training** and **TBN**, with **OT-TT**;
- **Episodic training** and **TBN**, with **OT**;
- Standard Empirical Risk Minimization (**ERM**) instead of episodic training, with **CBN** and not Optimal Transport (**Vanilla**);
- **ERM** with **CBN** and **OT**;
- **ERM** with **TBN** and no Optimal Transport (**Vanilla**);
- **ERM** with **TBN** and **OT**.

Transductive Fine-Tuning (FTNet) is not compatible with episodic training. Also the integration of Optimal Transport into this algorithm is non trivial. Therefore we only applied FTNet with ERM and without OT.

Every result presented in the following tables is the average over three runs with three random seeds (1, 2 and 3). For clarity, we do not report the 95% confidence interval for each result. Keep in mind that this interval is different for each result, but we found that it is always greater than $\pm$ 0.2% and smaller than $\pm$ 0.8%.

Details of the experiments and instructions to reproduce them are available in the code.

| | Episodic training | | | | | | Standard ERM | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | CBN | | | TBN | | | CBN | | TBN | |
| | Vanilla | w. OT-TT | w. OT | Vanilla | w. OT-TT | w. OT | Vanilla | w. OT | Vanilla | w. OT |
| **Meta-CIFAR100-C 1-shot 8-target** | | | | | | | | | | |
| ProtoNet (Snell et al. 2017) | 30.02 | 32.11 | 33.74 | 32.47 | 32.81 | 34.00 | 29.10 | 35.48 | 29.79 | 35.4 |
| MatchingNet (Vinyals et al. 2016) | 30.71 | 32.85 | 34.48 | 32.97 | 32.78 | 35.11 | 33.50 | **36.13** | 33.67 | **35.87** |
| TransPropNet (Y. Liu et al. 2019) | 30.26 | 28.70 | 26.87 | 34.15 | 29.48 | 27.68 | 23.33 | 31.08 | 22.55 | 31.20 |
| FTNet (Dhillon et al. 2020) | - | - | - | - | - | - | 28.91 | - | 28.75 | - |
| **Meta-CIFAR100-C 1-shot 16-target** | | | | | | | | | | |
| ProtoNet (Snell et al. 2017) | 29.98 | 32.24 | 35.63 | 32.52 | 31.72 | **36.20** | 29.02 | 35.89 | 29.61 | 35.94 |
| MatchingNet (Vinyals et al. 2016) | 31.1 | 30.94 | 35.53 | 33.08 | 33.28 | **36.36** | 33.49 | **36.61** | 33.64 | **36.54** |
| TransPropNet (Y. Liu et al. 2019) | 30.82 | 32.39 | 31.15 | 34.83 | 33.53 | 31.33 | 26.81 | 33.9 | 27.92 | 34.10 |
| FTNet (Dhillon et al. 2020) | - | - | - | - | - | - | 29.01 | - | 28.86 | - |
| **Meta-CIFAR100-C 5-shot 8-target** | | | | | | | | | | |
| ProtoNet (Snell et al. 2017) | 42.77 | 47.54 | 48.37 | 48.00 | 48.62 | **49.71** | 44.89 | 48.61 | 46.59 | 48.66 |
| MatchingNet (Vinyals et al. 2016) | 41.15 | 43.90 | 44.55 | 45.05 | 44.86 | 45.78 | 43.00 | 45.35 | 43.51 | 45.10 |
| TransPropNet (Y. Liu et al. 2019) | 39.13 | 40.60 | 25.68 | 47.39 | 40.47 | 27.29 | 29.32 | 39.82 | 29.50 | 29.82 |
| FTNet (Dhillon et al. 2020) | - | - | - | - | - | - | 37.28 | - | 37.40 | - |
| **Meta-CIFAR100-C 5-shot 16-target** | | | | | | | | | | |
| ProtoNet (Snell et al. 2017) | 42.07 | 48.26 | 48.25 | 46.49 | 48.71 | **49.94** | 44.67 | 48.61 | 46.48 | 48.89 |
| MatchingNet (Vinyals et al. 2016) | 41.74 | 44.51 | 45.71 | 44.91 | 44.71 | 47.37 | 42.97 | 46.06 | 46.22 | 46.37 |
| TransPropNet (Y. Liu et al. 2019) | 38.73 | 39.25 | 37.22 | 43.91 | 40.62 | 40.02 | 33.06 | 40.03 | 33.93 | 40.03 |
| FTNet (Dhillon et al. 2020) | - | - | - | - | - | - | 37.51 | - | 37.66 | - |
| **FEMNIST-FewShot 1-shot 1-target** | | | | | | | | | | |
| ProtoNet (Snell et al. 2017) | 84.31 | 94.00 | 92.31 | 90.36 | **94.92** | 93.63 | 80.20 | 94.30 | 86.22 | 94.22 |
| MatchingNet (Vinyals et al. 2016) | 84.25 | 93.66 | 92.73 | 91.05 | **95.37** | 93.62 | 85.04 | 94.34 | 87.19 | 94.26 |
| TransPropNet (Y. Liu et al. 2019) | 31.30 | 40.60 | 79.30 | 86.42 | 93.08 | 87.52 | 45.36 | 73.64 | 47.34 | 79.50 |
| FTNet (Dhillon et al. 2020) | - | - | - | - | - | - | 86.13 | - | 85.92 | - |

Table 1: Ablation study for all compared state-of-the-art methods on MC100-C and FEMNIST.

# Extended results for Open-Set Few-Shot Image Classification

## 1 Normalizing centroids

Because we work with normalized features, we state in our implementation details that we found normalizing $||\boldsymbol{\mu}||$ after each update helps. Here we show that this "projected step" is actually the exact solution to the optimization problem Eq. (4.10) when adding the constraint $\boldsymbol{\mu} \in \mathcal{B}_2$, where $\mathcal{B}_2 = \{\boldsymbol{x} : ||\boldsymbol{x}||_2 = 1\}$ is the unit hypersphere.

Specifically, adding the constraint $\boldsymbol{\mu} \in \mathcal{B}_2$ modifies the Lagrangian by infinitely penalizing $\boldsymbol{\mu}_k$ for being outside the unit hypersphere. Without loss of generality, we only consider the part of the Lagrangian pertaining to $\boldsymbol{\mu}_k$ for some $k \in [1, K]$, which we refer to as $\mathcal{L}_k$:

$$\mathcal{L}_k(\boldsymbol{\mu}_k) = \sum_{i=1}^{|\mathbb{S}|+|\mathbb{Q}|} \xi_i z_{ik} ||\boldsymbol{\mu}_k - \phi_\theta(\boldsymbol{x}_i)||^2 + \mathcal{L}_{\mathcal{B}_2}(\boldsymbol{\mu}_k)$$

where $\mathcal{L}_{\mathcal{B}_2}(\boldsymbol{\mu}_k)$ equals 0 if $\boldsymbol{\mu}_k \in \mathcal{B}_2$ and $\infty$ otherwise. Because $\mathcal{L}_k$ is no longer differentiable, we introduce the subdifferential operator $\partial_\cdot(\cdot)$, which generalizes the standard notion of differentiability. Akin to the standard case, we look for $\boldsymbol{\mu}_k$ such that:

$$0 \in \partial_{\boldsymbol{\mu}_k} \mathcal{L}_k(\boldsymbol{\mu}_k),$$

which amounts to:

$$
\begin{aligned}
\Leftrightarrow \quad & 0 \in \{\nabla_{\boldsymbol{\mu}_k} \sum_{i=1}^{|\mathbb{S}|+|\mathbb{Q}|} \xi_i z_{ik} ||\boldsymbol{\mu}_k - \phi_\theta(\boldsymbol{x}_i)||^2\} + \partial_{\boldsymbol{\mu}_k} \mathcal{L}_{\mathcal{B}_2}(\boldsymbol{\mu}_k) \\
\Leftrightarrow \quad & \sum_{i=1}^{|\mathbb{S}|+|\mathbb{Q}|} \xi_i z_{ik}\, \phi_\theta(\boldsymbol{x}_i) - \boldsymbol{\mu}_k \Big( \sum_{i=1}^{|\mathbb{S}|+|\mathbb{Q}|} \xi_i z_{ik} \Big) \in \partial_{\boldsymbol{\mu}_k} \mathcal{L}_{\mathcal{B}_2}(\boldsymbol{\mu}_k) \\
\Leftrightarrow \quad & \frac{\sum_{i=1}^{|\mathbb{S}|+|\mathbb{Q}|} \xi_i z_{ik}\, \phi_\theta(\boldsymbol{x}_i)}{\sum_{i=1}^{|\mathbb{S}|+|\mathbb{Q}|} \xi_i z_{ik}} - \boldsymbol{\mu}_k \in \partial_{\boldsymbol{\mu}_k} \frac{1}{\sum_{i=1}^{|\mathbb{S}|+|\mathbb{Q}|} \xi_i z_{ik}} \mathcal{L}_{\mathcal{B}_2}(\boldsymbol{\mu}_k) \\
\Leftrightarrow \quad & \frac{\sum_{i=1}^{|\mathbb{S}|+|\mathbb{Q}|} \xi_i z_{ik}\, \phi_\theta(\boldsymbol{x}_i)}{\sum_{i=1}^{|\mathbb{S}|+|\mathbb{Q}|} \xi_i z_{ik}} - \boldsymbol{\mu}_k \in \partial_{\boldsymbol{\mu}_k} \mathcal{L}_{\mathcal{B}_2}(\boldsymbol{\mu}_k) \\
\Leftrightarrow \quad & \boldsymbol{\mu}_k = \text{Proj}_{\mathcal{B}_2}\Big(\frac{\sum_{i=1}^{|\mathbb{S}|+|\mathbb{Q}|} \xi_i z_{ik}\, \phi_\theta(\boldsymbol{x}_i)}{\sum_{i=1}^{|\mathbb{S}|+|\mathbb{Q}|} \xi_i z_{ik}}\Big)
\end{aligned}
$$

where the penultimate step holds because $\lambda \mathcal{L}_{\mathcal{B}_2}(\boldsymbol{\mu}_k) = \mathcal{L}_{\mathcal{B}_2}(\boldsymbol{\mu}_k)$ by definition of $\mathcal{L}_{\mathcal{B}_2}(\boldsymbol{\mu}_k)$, and the last step holds because the projection operator $\text{Proj}_{\mathcal{B}_2}(\boldsymbol{\mu}_k) = \frac{\boldsymbol{\mu}_k}{||\boldsymbol{\mu}_k||}$ is the proximity operator of the constraint function $\mathcal{L}_{\mathcal{B}_2}(\boldsymbol{\mu}_k)$.

## 2 Detailed 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 $\gamma$, 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.

## 3 Additional results

We provide a more complete version of Fig. 4.5 in Fig. 1 and 2, showing the additional Prec@0.9 metric, along with the results on the WRN2810 provided by Ye et al. 2020.

In Figure 3, we propose additional results on *mini*ImageNet to show the impact of the size of the query set on the performance of various methods.

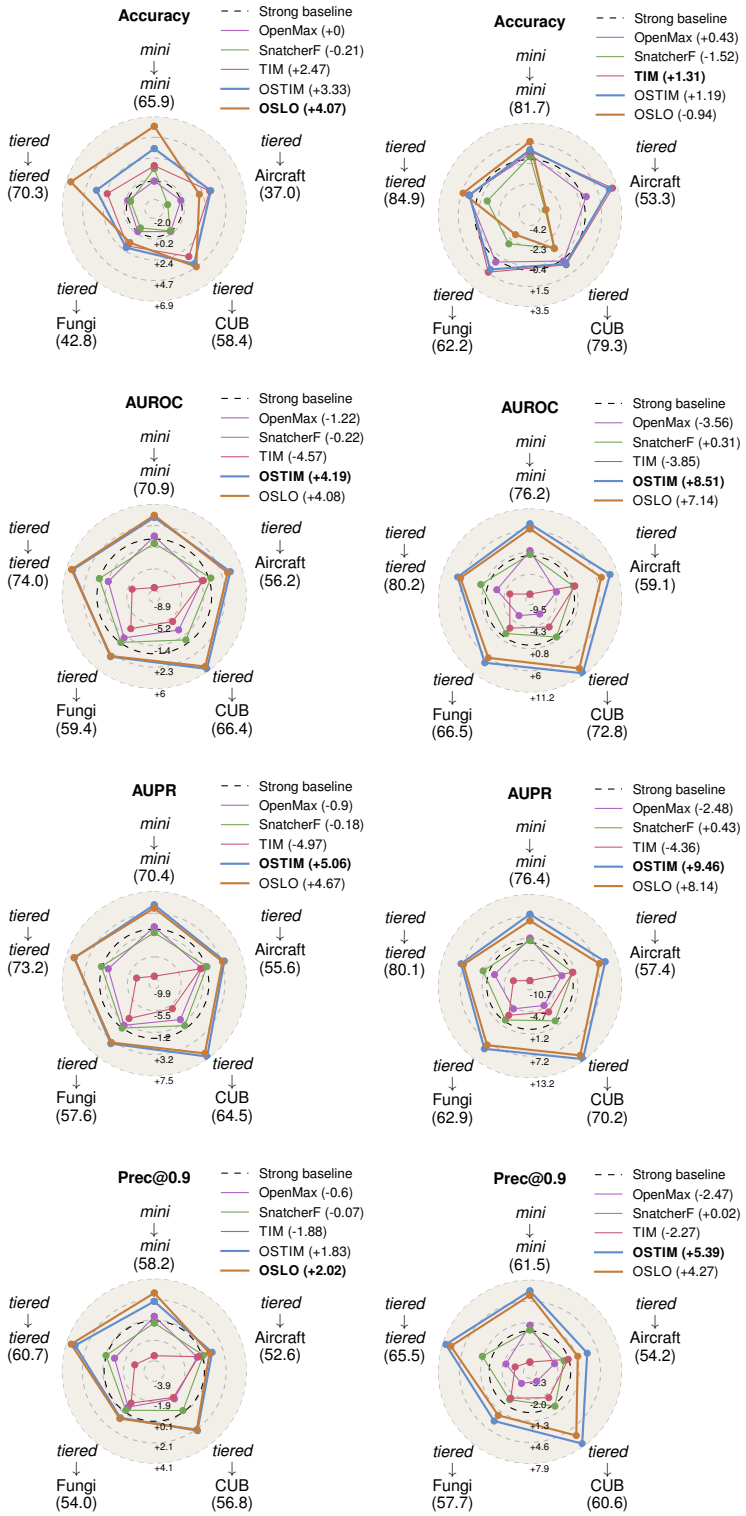



Figure 1: Complete version of Fig. 4.5 with a ResNet-12. (Left column): 1-shot. (Right column): 5-shot.

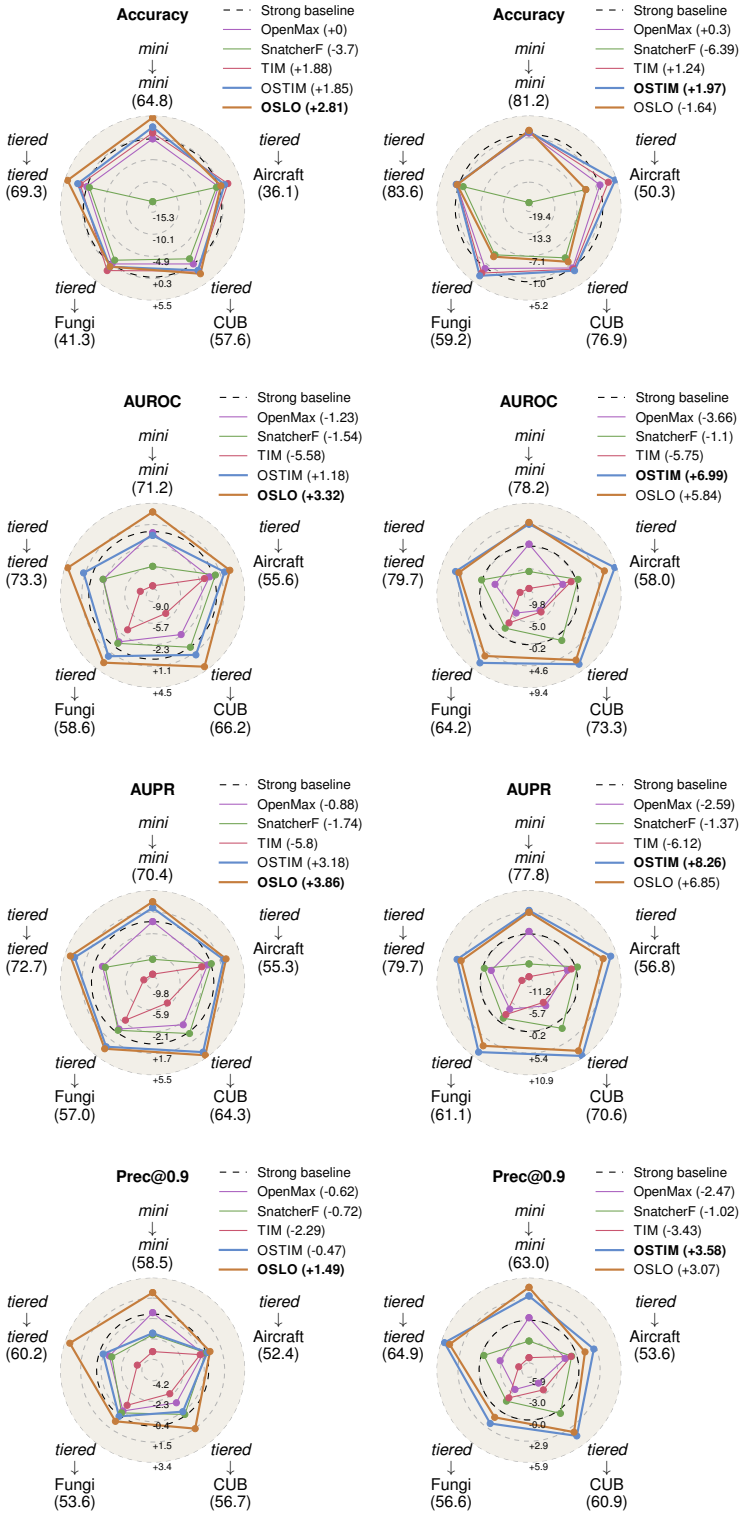



Figure 2: Complete version of Fig. 4.5 with a WideResNet 28-10. (Left column): 1-shot. (Right column): 5-shot.

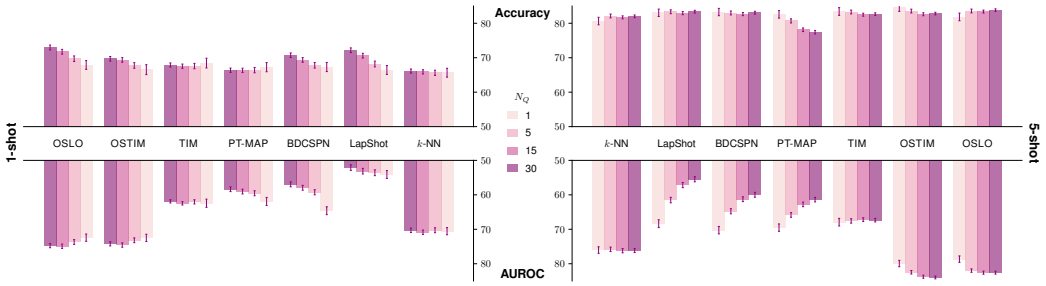



Figure 3: **OSLO improves performance even with few queries.** We study the closed-set (accuracy) and open-set (AUROC) performance of transductive methods depending on the size of the query set on *mini*-ImageNet in the 1-shot and 5-shot settings. The total size $|Q|$ of the query set is obtained by 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. This extends the results of Figure 4.7.