# FICUS: FEW-SHOT IMAGE CLASSIFICATION WITH UNSUPERVISED SEGMENTATION

Jonathan Lys, Frédéric Lin,
Clément Béliveau, Jules Decaestecker
*IMT Atlantique*
Brest, France
name.surname@imt-atlantique.org

Yassir Bendou, Aymane Abdali,
Bastien Pasdeloup
*IMT Atlantique, Lab-STICC, UMR CNRS 6285*
Brest, France
name.surname@imt-atlantique.fr

***Abstract*—In the realm of image classification, annotations often describe a single category. However, images might contain multiple objects including spurious ones with respect to the annotation. In few-shot image classification, where data is scarce, the ambiguity of these labels can severely impact classification performance. This paper addresses this issue by localizing objects in test images before classification and providing a disambiguated image embedding. We first show that using ground truth localization information can significantly improve performance. Second, we propose a method that leverages unsupervised object segmentation to detect and segment objects in images, in a training-free manner. Through extensive experiments and evaluations, we illustrate the efficacy of our method, highlighting its capacity to improve state-of-the-art classifiers in few-shot classification.**

***Index Terms*—deep learning, machine learning, few-shot classification, disambiguation, segmentation**

## I. Introduction

Few-shot image classification is characterized by a small number of available annotated examples, typically 1–5 per class. Given the few available data, existing approaches often leverage the power of pre-trained image encoders, mostly due to their generalized classification capabilities [1], [2]. However, such models might struggle when faced with complex scenes involving multiple objects, as the feature representation of such an image integrates information not only from the object of interest but also from surrounding elements.

In this article, we address the challenge of ambiguous few-shot image classification. A possible solution to improve the classification of ambiguous images is to segment their content before computing the feature representations of corresponding parts using a pre-trained Vision Transformer (ViT) such as DINO [3]. Indeed, these parts may convey more specific information on areas of the images, and therefore allow one to separate concepts for few-shot classification.

In this work, we introduce a fully unsupervised methodology – FICUS – to decompose query images into meaningful crops in a few-shot image classification problem. Our method capitalizes on pre-trained segmentation models, in particular, the Segment Anything Model (SAM) [4].

This article is organized as follows: after introducing related work in Section II, we describe the few-shot classification pipeline used, and introduce our methodology in Section III for instance detection and integration in that pipeline. Section IV then presents our experiments[1] on three datasets. Section V is a conclusion. Main findings of this work are:

- We demonstrate that the ability to localize instances in images improves performance;
- We propose a methodology – FICUS – exploiting pre-trained image segmentation models, and evaluate its performance in few-shot image classification.

## II. Related Work

### A. Few-shot image classification

In few-shot classification [5], one aims to predict the labels of a query set $\mathcal{Q} = \{(\mathbf{x}_i, y_i)\}_i$, given a support set $\mathcal{S} = \left\{(\mathbf{x}_j^*, y_j^*)\right\}_j$, where $\mathbf{x}_i$ (resp. $\mathbf{x}_j^*$) denote data to classify (resp. annotated data) and $y_i \in \mathcal{Y}$ (resp. $y_j^* \in \mathcal{Y}$) the associated class labels. In few-shot problems, $y_i$ are the targets to predict, and $y_j^*$ are generally available in very low numbers. In this work, we consider that $\mathcal{S}$ contains $K$ examples of each class. We generally distinguish two settings: 1) inductive few-shot classification has access to points in $\mathcal{Q}$ one by one and should make predictions on the fly; 2) transductive few-shot classification has full access to $\mathcal{Q}$ and makes a posteriori predictions. We focus on the inductive setting, leaving transductive for future work.

There are several approaches to address few-shot classification problems. Nearest Class Mean (NCM), or prototypical networks [6], [5] rely on computing a prototype $\overline{\mathbf{z}}_y$ for each class $y \in \mathcal{Y}$, or by finding it by gradient descent [1], and associate a class to a query by assigning the label of the closest prototype. Methods based on meta-learning [7], [8] aim at training models on abstract tasks for rapid adaptation to new problems. Other methods rely on knowledge transferability of models and train a simple linear layer for classification [1]. Those methods leverage a pre-trained network. It is shown in [2] that using a pre-trained DINO model [3] is very effective, demonstrating that having a great semantic diversity in the training dataset – *e.g.*, ImageNet with its 21k classes – is a key factor for generalization properties, which is important in few-shot learning. The efficacy of feature extraction heavily relies on the model's ability to generalize across diverse visual concepts encountered during training.

[1]Codes are available at https://github.com/NewS0ul/FICUS.

### B. Ambiguous few-shot image classification

More specifically, we consider in this work ambiguous few-shot problems. In this setting, images in $\mathcal{S}$ and $\mathcal{Q}$ can feature multiple objects, but each image only has a single annotation. It may therefore be beneficial to restrict images to the crops that feature these objects only.

The exploration of ambiguity within few-shot problems has not received extensive attention in recent research. While prior studies [9], [10], [11] have tackled similar challenges by incorporating multiple local features to represent an image, instead of a single global representation, it is noteworthy that many of these approaches heavily rely on training convolutional neural networks for feature generation. In contrast, our approach stands out for being training-free, providing a distinctive perspective on addressing ambiguity in few-shot problems without the need for high performance devices.

Within the limited body of works addressing ambiguous few-shot problems, the authors of [12] introduce an optimization routine based on random crops to detect multiple objects in an image. FewTure [13] adjusts the model's patch weights to enhance the representation of the area of interest in an image. In this work, we take an approach that exploits pre-trained segmentation models to identify possible objects of interest.

### C. Object detection and image segmentation

SAM [4], is a class-agnostic segmentation model, having undergone training on an extensive dataset comprising 1 billion masks across 11 million images. This extensive training ensures minimal bias from the dataset and notable zero-shot capabilities. Notably, SAM's class-agnostic nature empowers the model with broad applicability across various classes, showcasing its strong generalization capabilities. SAM comes with an Automatic Mask Generator (AMG), which evenly distributes a grid of points on the image to use them as prompts. Each prompt results in three masks of various scales, and it frequently happens that objects are split into separate masks. This method therefore often results in numerous masks, as the grid of points holds no prior information about where the object is located in a scene.

Unsupervised methods for object detection and semantic segmentation models from ViT have been developed. The LOST method [14] exploits a graph representation of patches in the image to expand a seed patch until a coarse estimation of patches that contain the object of interest is produced. However, this method can only extract one object of interest, meaning it falls short when dealing with ambiguous situations with multiple instances. Deep Spectral Method (DSM) [15] also considers the output of a ViT as a graph. Then, a spectral decomposition is performed on the corresponding Laplacian matrix. The resulting eigenvectors represent several orthogonal concepts in the latent space. Those eigenvectors can be reshaped into images, each representing a different semantic concept due to the orthogonality property.

## III. Methodology

Our methodology is illustrated in Figure 1. Given a query image $\mathbf{x}_i$, FICUS consists of three steps:

### A. Instance extraction

We first proceed by localizing objects in the image. The rationale is to identify a set of crops in the image that may accurately describe distinct semantic concepts.

We segment the query image $\mathbf{x}_i$ using SAM. It takes as input an image and a prompt, and generates a segmentation in a zero-shot and class-agnostic way.

To determine interesting prompt points, we leverage $E$ eigenmaps $\{\mathbf{e}_{i,e}\}_{e\in\{1\dots E\}}$ extracted by the Deep Spectral Method (DSM) II-C, which we turn into multinomial spatial distributions by applying a softmax function. For each map $\mathbf{e}_{i,e}$, we thus obtain a distribution of the concept captured by the map across the image. We can then randomly sample $P$ points $\{\mathbf{p}_{i,e,p}\}_{p\in\{1\dots P\}}$ using this distribution, which can be used to prompt SAM.

By default, for eigenmap $\mathbf{e}_{i,e}$ and point $\mathbf{p}_{i,e,p}$, SAM outputs a triplet $\left(\mathbf{m}^{\text{small}}_{i,e,p}, \mathbf{m}^{\text{medium}}_{i,e,p}, \mathbf{m}^{\text{large}}_{i,e,p}\right)$ of masks, with distinct scales, around the prompt point.

### B. Masks selection

To select the correct scale among the three candidate masks per eigenmap and point, we create an additional mask $\mathbf{m}^{\text{otsu}}_{i,e}$. It is determined using Otsu's method, a parameter-free automatic image thresholding method [16], on the eigenmap $\mathbf{e}_{i,e}$. We then keep the candidate mask that maximizes Intersection over Union (IoU) with $\mathbf{m}^{\text{otsu}}_{i,e}$.

After applying this procedure to all eigenmaps and points, we filter out some masks that are redundant, in the sense that they cover the same image areas. To do so, we employ a Non-Maximum Suppression (NMS) algorithm, based on masks intersection and the confidence score provided by SAM.

Masks left after this filtering are then used to extract corresponding crops. To these crops, the original image is added to account for any segmentation model failure. The resulting set is denoted $\{\mathbf{x}_{i,c}\}_c$, from image $\mathbf{x}_i$ .

### C. Classification

For each class $y \in \mathcal{Y}$, we compute its prototype:

$$\forall y \in \mathcal{Y} : \overline{\mathbf{z}}_y = \frac{1}{K} \sum_{j \text{ s.t. } y^*_j = y} \mathbf{z}^*_j \,, \tag{1}$$

where $\mathbf{z}^*_j$ is the feature representation of support image $\mathbf{x}^*_j$.

Finally, we attribute a label to $\mathbf{x}_i$ using Equation (2), an adapted NCM classifier with cosine similarity:

$$NCM(\mathbf{x}_i) = \arg\max_{y\in\mathcal{Y}} \left( \max_c \frac{\mathbf{z}_{i,c} \cdot \overline{\mathbf{z}}_y}{\|\mathbf{z}_{i,c}\| \, \|\overline{\mathbf{z}}_y\|} \right) \,, \tag{2}$$

where $\{\mathbf{z}_{i,c}\}_c$ are the feature representations of crops $\{\mathbf{x}_{i,c}\}_c$ from previous step, and $\cdot$ is the dot product.

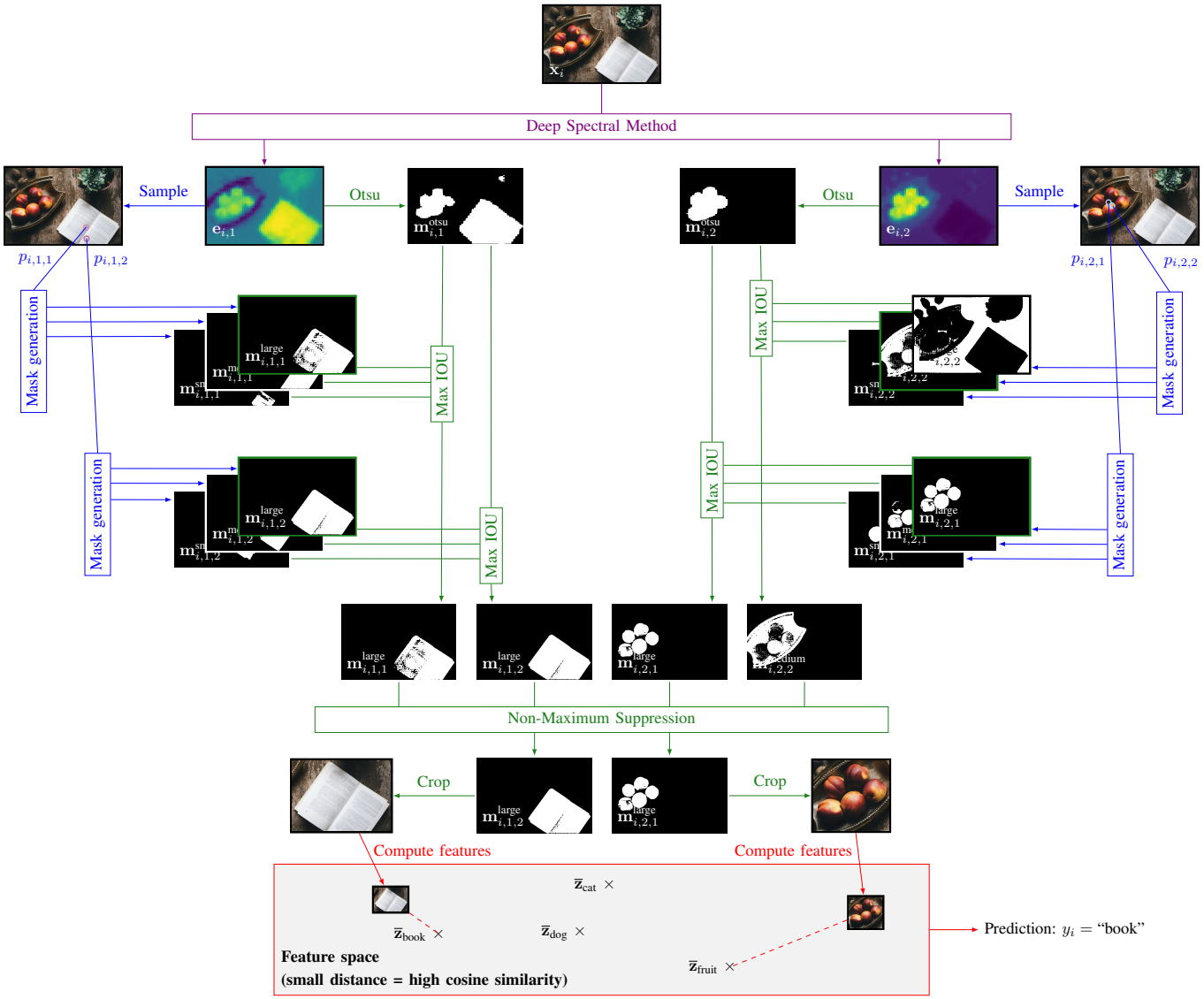



Fig. 1: Overview of FICUS. (violet) First, eigenmaps are produced using DSM (here 2). Each map is treated separately. (blue) Using the maps, random points (here 2) are sampled and used to prompt SAM. For each point, we obtain 3 candidate masks. (green) Out of each group of 3 candidate masks, we keep the one that maximizes IOU with an Otsu thresholding of the map. Redundant masks are then filtered out using NMS. (red) Finally, kept masks are used to compute feature representations of associated crops. A NCM is then used to return a label for the image.

## IV. Experiments

### A. Experimental settings

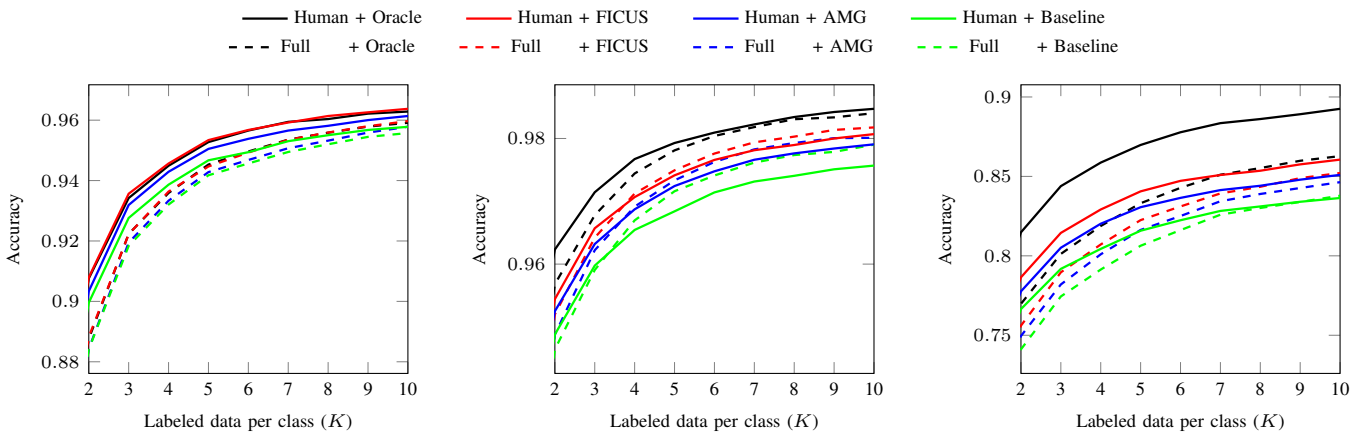
We consider two distinct settings for support images, corresponding to possible use cases:

- **Full.** Support images are used in their entirety to compute prototypes in Equation (1);
- **Human.** For each pair $(\mathbf{x}_j^*, y_j^*) \in \mathcal{S}$, we ask a human annotator to extract a set of crops $\{\mathbf{x}_{j,c}^*\}_c$ of objects of class $y_j^*$ in that image. This leads to a better estimation of class prototypes, by replacing $\mathbf{z}_j^*$ with $\frac{1}{C_i}\mathbf{z}_{j,c}^*$ in Equation (1), where $\{\mathbf{x}_{j,c}^*\}_c$ are the feature representations of these $C_i$ crops.

For each of these settings, we consider the following four approaches to classify the query images:

- **Baseline.** The whole image is used;
- **FICUS.** Our methodology introduced in Section III;
- **AMG.** SAM's automatic mask generator;
- **Oracle.** We use crops obtained from ground truth bounding boxes. This aims at estimating the gap between previous approaches and the best achievable accuracy when cropping.

We report results on PascalVOC [17], CUB [18] and ImageNet [19], complemented with bounding boxes from [20]. Bounding boxes associated with each dataset are used for the Human support setting and the Oracle query setting. Few-shot




(a) CUB. (b) ImageNet. (c) PascalVOC.

Fig. 2: Comparison of accuracies obtained in Human (continuous lines) versus Full (dashed lines) support settings.

problems created on PascalVOC are in general ambiguous as it is a detection dataset. This is also partially the case for ImageNet, but not for CUB, which allows us to evaluate our method on that setting too.

To create few-shot problems, we randomly select 5 classes, from which we sample $K \in \{1 \dots 10\}$ support images and 15 query images per class. In our experiments, for the FICUS methodology, we use $E = 5$ eigenmaps as described in [15] with the softmax temperature set to 0.1, and sample $P = 10$ points per eigenmap.

To compute the feature representation of images/crops, we used the `dino_vits16` ViT encoder [3]. Note that the image encoder used was trained on the Imagenet dataset. Consequently, the image encoder is exposed to the dataset's distribution, which may affect its classification performance.

For each dataset, we perform the following experiments: for multiple values of $K$, we compare – both for the Full and Human support settings – all query settings previously described. We report the classification accuracy using the NCM described in Equation (2), averaging over 500 random independent few-shot runs for each $K$.

### B. Results

**Support localization importance.** We first examine the potential benefits of incorporating localization information to improve accuracy compared to full images. Specifically, we study the improvements observed in the Human setting compared to the Full one. As shown in Figure 2, with the exception of the ImageNet dataset, the accuracy is always higher in the Human setting compared to the Full setting, for both query settings. We hypothesize that the ImageNet gap is related to the encoder knowing the dataset's distribution, and therefore slightly overfitting on background data. Interestingly, results on PascalVOC highlight the high benefits of support localization in ambiguous situations. We also observe a small improvement on CUB, showing interest of cropping even in non-ambiguous scenarios.

**Performance evaluation.** Quantitatively speaking, results in Figure 3 and Table I show a slight improvement of accuracy with FICUS over AMG, in nearly all cases. In all settings, both approaches improve performance over baseline, highlighting the benefits of localization in the query set, in conjunction with the support, on ambiguous datasets or not. Note that there is still room for improvement, as the Oracle is still way above achieved performance in most settings, notably on PascalVOC. Thus, better localization of objects of interest is still an interesting direction for future research.

## V. Conclusions

This work demonstrated the positive impact of image disambiguation in the context of few-shot image classification. Instance localization improved classification accuracy for all tested datasets, whether applied to support or query images. We introduced FICUS, a non-supervised approach leveraging segmentation models to crop query images. FICUS exhibits higher performance compared to methods that rely on segmentation tools only, or do not address ambiguity at all.

Future work includes extension of the approach to the transductive setting. Additionally, to obtain a fully non-supervised process for both support and query sets, we should investigate methods for support disambiguation.

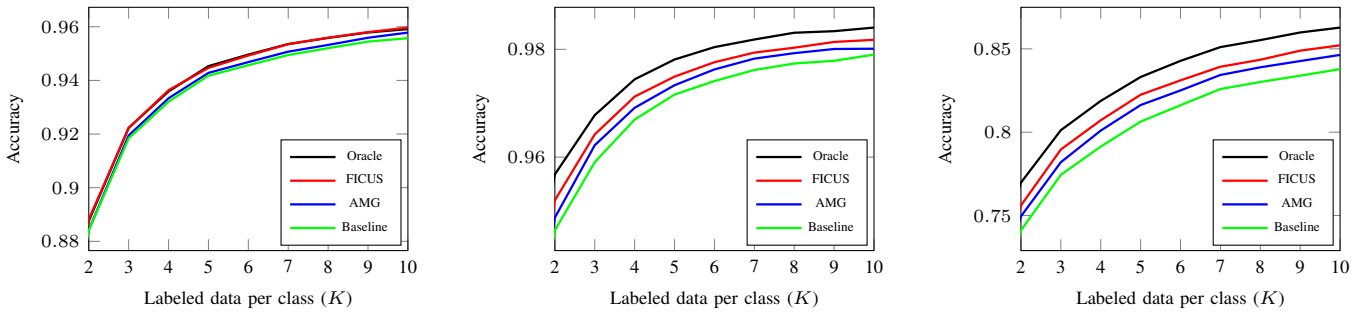
## References

[1] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C. F. Wang, and J.-B. Huang, "A closer look at few-shot classification," in *International Conference on Learning Representations*, 2018.

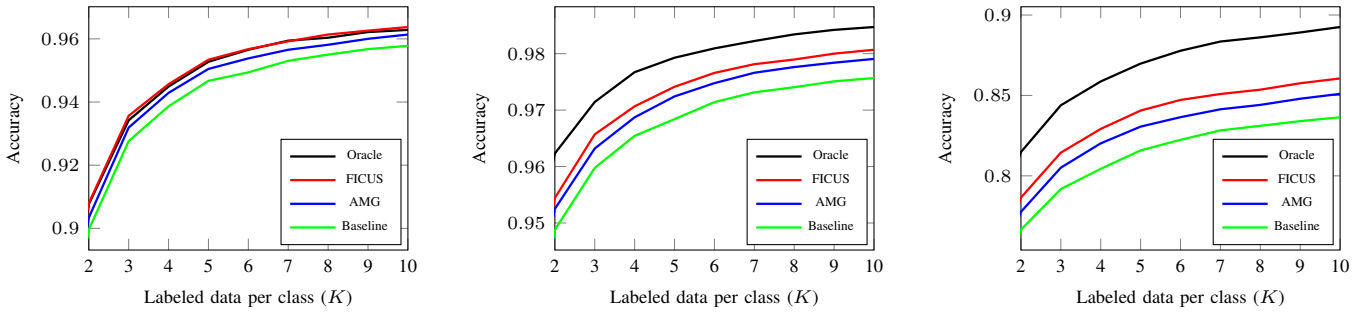
[2] X. Luo, H. Wu, J. Zhang, L. Gao, J. Xu, and J. Song, "A closer look at few-shot classification again," in *International Conference on Machine Learning*. PMLR, 2023.

[3] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021.

[4] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.




(a) CUB, Full support. (b) ImageNet, Full support. (c) PascalVOC, Full support.




(d) CUB, Human support. (e) ImageNet, Human support. (f) PascalVOC, Human support.

Fig. 3: Evaluation of methods for both settings on all datasets. Performance for $K = 1$ is given in Table I for readability.

TABLE I: Accuracies obtained for $K = 1$ and $K = 5$ per dataset per support setting. These settings are the commonly found "5-way 1-shot" and "5-way 5-shot" in the literature. We highlight the best non-oracle performance in bold.

| Support | | Full | | | | Human | | | |
|---|---|---|---|---|---|---|---|---|---|
| Query | | Baseline | AMG | FICUS | Oracle | Baseline | AMG | FICUS | Oracle |
| ImageNet | $K = 1$ | 90.22 ± 0.67 | 90.53 ± 0.67 | **90.83 ± 0.68** | 91.50 ± 0.66 | 91.20 ± 0.61 | 91.71 ± 0.61 | **92.06 ± 0.61** | 92.90 ± 0.58 |
| | $K = 5$ | 97.16 ± 0.29 | 97.33 ± 0.30 | **97.49 ± 0.28** | 97.81 ± 0.27 | 96.84 ± 0.31 | 97.24 ± 0.29 | **97.41 ± 0.29** | 97.93 ± 0.26 |
| CUB | $K = 1$ | 80.43 ± 1.00 | 80.18 ± 1.02 | **80.51 ± 1.05** | 80.44 ± 1.06 | 82.53 ± 0.94 | 83.12 ± 0.97 | **83.18 ± 0.98** | 83.12 ± 1.00 |
| | $K = 5$ | 94.17 ± 0.47 | 94.38 ± 0.46 | **94.40 ± 0.49** | 94.45 ± 0.49 | 94.73 ± 0.43 | 94.98 ± 0.46 | **95.26 ± 0.46** | 95.27 ± 0.45 |
| PascalVOC | $K = 1$ | 66.43 ± 1.07 | 67.22 ± 1.07 | **67.86 ± 1.08** | 69.39 ± 1.11 | 69.95 ± 0.98 | 70.95 ± 0.99 | **71.64 ± 1.00** | 74.36 ± 0.98 |
| | $K = 5$ | 80.63 ± 0.79 | 81.63 ± 0.80 | **82.25 ± 0.80** | 83.31 ± 0.80 | 81.58 ± 0.77 | 83.06 ± 0.75 | **84.06 ± 0.73** | 86.98 ± 0.66 |


[5] E. Triantafillou, T. Zhu, V. Dumoulin, P. Lamblin, U. Evci, K. Xu, R. Goroshin, C. Gelada, K. Swersky, P.-A. Manzagol *et al.*, "Meta-dataset: A dataset of datasets for learning to learn from few examples," in *International Conference on Learning Representations*, 2019.

[6] Y. Bendou, Y. Hu, R. Lafargue, G. Lioi, B. Pasdeloup, S. Pateux, and V. Gripon, "Easy—ensemble augmented-shot-y-shaped learning: State-of-the-art few-shot classification with simple components," *Journal of Imaging*, 2022.

[7] A. Nichol, J. Achiam, and J. Schulman, "On first-order meta-learning algorithms," *arXiv:1803.02999*, 2018.

[8] C. Finn, K. Xu, and S. Levine, "Probabilistic model-agnostic meta-learning," *Advances in neural information processing systems*, 2018.

[9] J. Cheng, F. Hao, L. Liu, and D. Tao, "Imposing semantic consistency of local descriptors for few-shot learning," *IEEE Transactions on Image Processing*, 2022.

[10] Y. Liu, W. Zhang, C. Xiang, T. Zheng, D. Cai, and X. He, "Learning to affiliate: Mutual centralized learning for few-shot classification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022.

[11] J. Cheng, F. Hao, F. He, L. Liu, and Q. Zhang, "Mixer-based semantic spread for few-shot learning," *IEEE Transactions on Multimedia*, 2021.

[12] Y. Bendou, L. Drumetz, V. Gripon, G. Lioi, and B. Pasdeloup, "Disambiguation of one-shot visual classification tasks: A simplex-based approach," *arXiv:2301.06372*, 2023.

[13] M. Hiller, R. Ma, M. Harandi, and T. Drummond, "Rethinking generalization in few-shot classification," *Advances in Neural Information Processing Systems*, 2022.

[14] O. Siméoni, G. Puy, H. V. Vo, S. Roburin, S. Gidaris, A. Bursuc, P. Pérez, R. Marlet, and J. Ponce, "Localizing objects with self-supervised transformers and no labels," in *BMVC 2021-32nd British Machine Vision Conference*, 2021.

[15] L. Melas-Kyriazi, C. Rupprecht, I. Laina, and A. Vedaldi, "Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[16] N. Otsu *et al.*, "A threshold selection method from gray-level histograms," *Automatica*, 1975.

[17] M. Everingham and J. Winn, "The pascal visual object classes challenge 2012 development kit," *Pattern Anal. Stat. Model. Comput. Learn., Tech. Rep*, 2012.

[18] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," 2011.

[19] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, 2015.

[20] W. K. Addison Howard, Eunbyung Park, "Imagenet object localization challenge," 2018. [Online]. Available: https://kaggle.com/competitions/imagenet-object-localization-challenge