Prototypical Region Proposal Networks for Few-shot Localization and Classification

Anonymous Author(s)
Affiliation
Address
email

Abstract

Recently proposed few-shot image classification methods have generally focused on use cases where the objects to be classified are the central subject of images. Despite success on benchmark vision datasets aligned with this use case, these methods typically fail on use cases involving densely-annotated, busy images: images common in the wild where objects of relevance are not the central subject, instead appearing potentially occluded, small, or among other incidental objects belonging to other classes of potential interest. To localize relevant objects, we employ a prototype-based few-shot segmentation model which compares the encoded features of unlabeled query images with support class centroids to produce region proposals indicating the presence and location of support set classes in a query image. These region proposals are then used as additional conditioning input to few-shot image classifiers. We develop a framework to unify the two stages (segmentation and classification) into an end-to-end classification model—PRoPnet—and empirically demonstrate that our methods improve accuracy on image datasets with natural scenes containing multiple object classes.

1 Introduction

Metric learning approaches for image classification have shown impressive accuracy in the few-shot setting, where models must learn to detect unobserved classes from few labeled examples. The most successful demonstrations have been on datasets such as mini-ImageNet [7], ImageNet [4], and Omniglot [15]—datasets comprised of images typically containing centered, non-occluded objects commonly appearing in the foreground. However, performance of state-of-the-art few-shot classification methods can severely degrade on naturally-occurring, so-called "busy image" scenes wherein the object to classify is small or occluded, or accompanied by several other incidental objects from other classes. In the common use case where query images have neither image-level annotations nor any indication of where objects of interest are in the image, the few-shot classification task becomes considerably more challenging. This work presents and assesses methods for addressing these limitations by conditioning model predictions on regions of interest for improved classification performance in use cases involving busy, unlabeled query images.

Our presented methods assume the use of an annotated support set—images that define each new few-shot class—wherein each support image is accompanied not only with labels indicating their corresponding class, but also with an annotation mask indicating the objects of interest that belong to the class in question. However, we also explore the use case in which query images are *entirely unlabeled*—no image-level annotations or annotation masks are available—and attempt to gain insight into how this can impact model performance. We address three basic research questions related to few-shot classification with busy images:

1. Can annotation information be leveraged for enhanced classification of busy images?

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

2. How crucial are query image localization annotations for few-shot busy image classification?
3. Can few-shot models be leveraged to generate localization proposals for busy, unlabeled query images?

To address the first question, we compare the accuracy of few-shot classifiers with and without localization, finding that including localization features is beneficial to performance, particularly on densely-annotated datasets. Then, in comparing performance with and without query images with region annotations, we find that for datasets containing busy images which benefit most from localization, few-shot classification methods which incorporate localization yield sub-par performance when annotations are not available for query images.

As a solution to localizing and classifying busy, unlabeled query images under the few-shot paradigm, we present an n-shot, one-way segmentation network which produces query region proposal masks conditioned on a set of annotated support images. We develop an end-to-end classifier we call Prototypical Region Proposal Networks (PRoPnet), whose first stage generates region proposals for query images given a set of support images with class labels and object annotations, and whose second stage augments the standard ResNet-50 architecture with a four-channel input (color channels plus annotation mask) for localization-conditioned classification.

We conduct localization conditioning experiments on seven datasets which span a range of classification difficulties—difficult datasets like Visual Genome which contain images of very busy scenes, as well as those which fall on the other end of the spectrum of difficulty, such as ImageNet, where most images depict only objects from a single class of interest in the foreground. Our experiments show that query image annotation information can make few-shot classification a tractable problem for difficult scenes characteristic of Visual Genome, increasing 5-shot, 5-way accuracy from 59 to 76%. We further demonstrate PRoPnet's improved 5-shot, 5-way classification for the FSS-1000 and PASCAL-5^i datasets. Together, we conclude that i) localization conditioning can be crucial for adequate performance on busy natural image scenes and ii) two-stage networks like PRoPnet which incorporate generated query region proposals as input to a Prototypical Network are a promising approach to providing this essential conditioning.

Related Work Given saturation of performance on the standard few-shot benchmarks, recent works have extended few-shot tasks to more difficult settings, e.g. object classes with confounding texture [1], cluttered scene segmentation [22; 9], long tailed distributions [32], and segmentation of homogeneous object clusters [33]. Enhancements to Prototypical Networks (ProtoNets) [29] in particular have been proposed for difficult few-shot tasks involving, for instance, inhomogeneous noisy datasets [8], domain adaptation [24], and relation classification in text [10]. Several works have incorporated localization conditioning into a few-shot classification architecture [12; 17; 18; 26; 32]. We include comparison with the Few-Shot Localization (FSL) method presented in [32] in our experimental evaluations. Numerous contemporaneous papers have presented few-shot semantic segmentation methods [18; 6; 20; 31; 35; 23; 28; 22; 21; 36; 37; 3], which differ from our approach along several axes of variation including comparison metric, support set aggregation, feature map backbone, and mask refinement.

2 Methods

In this section we formulate our region proposal and classification architecture, and the implementation, training regimen, and selected datasets. Before describing our methods for region proposal and localized classification under the few-shot paradigm, we begin with a review of the few-shot problem and the Prototypical Network approach to few-shot classification. For the few-shot classification task, we have a ***support set*** $\mathcal{S} = \bigcup_{k=1}^{c} \{\mathbf{x}_{k,1}^{(s)}, \ldots, \mathbf{x}_{k,n}^{(s)}\}$ containing n labeled examples for each of c classes—"n-shot, c-way" in few-shot parlance. Given some unlabeled ***query*** example $\mathbf{x}^{(q)}$ whose class is unknown, our task is to determine which of the c support classes the query example represents. We assume some parametric ***encoder*** function f used to map support and query examples to $\mathbb{R}^d$. Let $\mathbf{s}_{k,i} = f(\mathbf{x}_{k,i}^{(s)})$ be the encoded i-th example in $\mathcal{S}$ from the k-th class. The ***support*** for the k-th class, is the matrix of encoded examples for that class, $\mathbf{S}_k = [\mathbf{s}_{k,1} \quad \ldots \quad \mathbf{s}_{k,n}]^\mathsf{T} \in \mathbb{R}^{n \times d}$.

For all few-shot classification models in this work, we use a Prototypical Network [29] variant with a ResNet-50 [11] encoder function for classification. The ***centroid*** for the k-th class

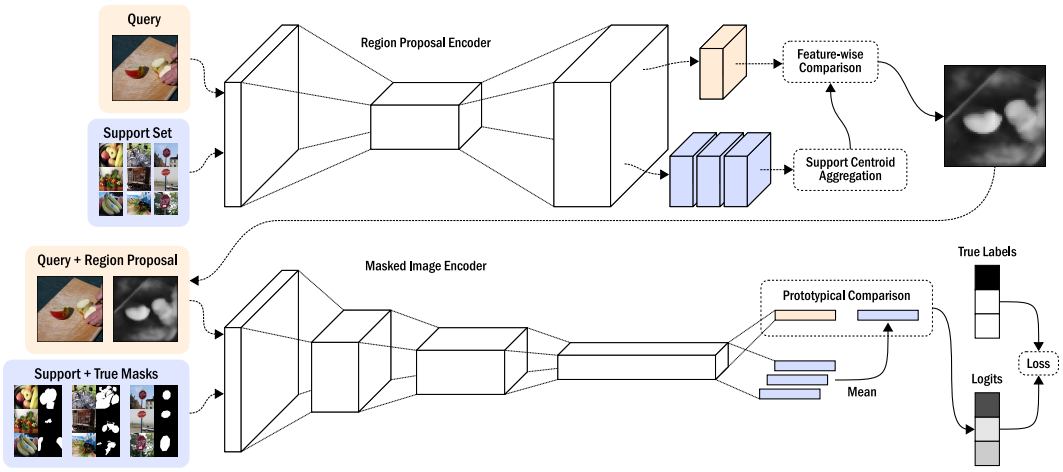

Figure 1: The full end-to-end PRoPnet model.

is the mean vector, $\bar{\mathbf{s}}_k = \frac{1}{n}\sum_{i=1}^{n} \mathbf{s}_{k,i}$. Squared Euclidean distances between the encoded query, $\mathbf{q} = f(\mathbf{x}^{(q)})$, and support centroids, $\mathbf{d}_k = ||\,\mathbf{q} - \bar{\mathbf{s}}_k\,||_2^2$, model a class membership probability distribution, $\hat{\mathbf{y}} = \text{Softmax}(-\mathbf{d})$. Models are trained by minimizing the predicted negative log likelihood of the query's ground-truth class: $-\log(\hat{\mathbf{y}}_k)$, when $\mathbf{q}$ belongs to the k-th class.

Region Proposal Network Our region proposal network (RPN) is a logical extension of the prototypical network classification to few-shot segmentation. We define a feature map encoder $\hat{f} : \mathbb{R}^{h\times w\times 3} \rightarrow \mathbb{R}^{h'\times w'\times d}$ to be a convolutional network that maps an image $\mathbf{X}$ to a feature map with d channels and possibly downsampled resolution, where $h' \leq h$, $w' \leq w$ and $d > 3$. The RPN is defined as a function which maps a tuple of n support images $\mathcal{S} = (\mathbf{X}_1, ..., \mathbf{X}_n) \in \mathbb{R}^{n\times h\times w\times 3}$, along with their respective n support annotation masks $\mathcal{M} = (\mathbf{M}_1, ..., \mathbf{M}_n) \in \{0,1\}^{n\times h\times w}$, and a query image $\mathbf{Q} \in R^{h\times w\times 3}$ to a query mask $\mathbf{P} \in [0,1]^{h\times w}$. Specifically, with i, j indexing vertical and horizontal pixel positions, and k indexing the support set exemplars:

$$\text{RPN}(\mathcal{S}, \mathcal{M}, \mathbf{Q})_{i,j} = s_{\cos}\big(\mathbf{c}, \hat{f}(\mathbf{Q})_{i,j,:}\big) \tag{1a}$$

$$\mathbf{c} = \frac{1}{n}\sum_{k=1}^{n} \text{MAP}\big(\mathbf{M}_k, \hat{f}(\mathbf{X}_k)\big) \tag{1b}$$

$$\text{MAP}(\mathbf{M}, \hat{f}(\mathbf{X})) = \frac{\sum_{i=1}^{h}\sum_{j=1}^{w} m_{i,j}\hat{f}(\mathbf{X})_{i,j,:}}{\sum_{i=1}^{h}\sum_{j=1}^{w} m_{i,j}} \tag{1c}$$

where $\mathbf{c}$ is a support class centroid derived by averaging the Masked Average Pooled (MAP) feature representations for the support set class given masks $\mathbf{M} \in \{0,1\}^{h\times w}$ and feature maps $\hat{f}(\mathbf{X})$. Our query mask prediction is then derived by a pixel-wise cosine similarity, $s_{\cos}$, comparison of query pixel features with the class centroid vector. Figure 1 shows a computational graph of the region proposal network. We choose the feature map encoder, $\hat{f}$, as the portion of UperNet [34] preceding their final 1×1 convolution classification layer. RPN models are trained using the Lovász-Softmax loss [2] between the predicted and ground truth query masks.

Prototypical Region Proposal Network Our Prototypical Region Proposal Network (PRoPnet) for localized few-shot classification is a composition of the RPN and a Prototypical Network modified with early fusion localization conditioning. Query masks are generated by the RPN, then used in a four-channel ResNet-50 feature encoder for ProtoNet few-shot classification (Figure 1). Unlike standard ProtoNets, for the classification stage of our network, the query has a different feature representation corresponding to each support class.

Implementation For the ResNet-50 model used by our localized few-shot classifier, we adapt the reference implementation provided by Torchvision[1]. We add an option to their ResNet-50 to replace the input layer with a four-channel convolutional layer. CSAIL's UperNet implementation[2] was used for our region proposal model—we retain only the Object head of the UperNet for prototypical segmentation. For the FSL baseline [32], we integrate their few-shot localization code[3] into our test harness—we omit the covariance pooling and batch folding techniques from our study.

2.1 Training

For the end-to-end system, we employ a two-stage training regimen. In the first stage of training the RPN and early fusion classifiers are trained independently. The RPN is initially trained for few-shot segmentation on Open Images. The early fusion ResNet-50 classifier is independently trained with ground truth masks as a standard image classifier, and then trained on the few-shot classification task, once again with ground truth masks. For the second stage of training, the final layers of the early fusion classifier are fine-tuned to adapt to masks generated from the RPN. Each epoch of few-shot training runs a given number of episodes in which a random combination of classes is selected and split into support and query classes, then examples for these classes are sampled to form proper support and query sets. For our experiments, we perform 500 sampling episodes per epoch for 100 epochs during training, and validate our models with 100 sampling episodes. We use data augmentation during training by applying random horizontal flipping, rotations, translations, and uniform scaling to both images and corresponding masks.

2.2 Datasets

Seven publicly available image datasets were selected which meet the criteria of a large number of classes for few shot training and evaluations, and object class label and localization annotations. We selected ILSVRC 2012 (ImageNet) [25] and iNaturalist [30] as these results can be easily benchmarked against previous few-shot research. Visual Genome (VG) [13] was selected as an exemplary example of a difficult dataset containing many busy real world scenes. We select four semantic segmentation datasets: COCO [19], FSS-1000 [16], PASCAL VOC [5], and Open Images [14]. COCO, FSS-1000, and VOC have all been featured as experimental datasets in recent work on few-shot localization. Recently becoming the *de facto* standard benchmark for few-shot semantic segmentation, we use a version of VOC, PASCAL-5^i [27] which designates four folds for cross-validation, each containing 15 classes for training and 5 for testing.

We filter classes in VG, iNaturalist, and Open Images to those which contain at least 200 example images, and exclude object annotations which cover less than 0.2% of the image. We combine the standard alias classes listed for VG and remove annotations for spurious class definitions such as "air," "the," "a," etc. which do not correspond to a homogeneous class of physical objects. After filtering, we create random train/validation/test splits over classes. For VG, iNaturalist, ImageNet, and Open Images which have a large number of classes, we perform an 80/10/10 split over classes. Since few-shot training and validation involve sampling combinations of classes, for COCO which has only 80 classes, we split the classes 60/20/20 such that train contains 48 classes and validation and test each contain 16 classes. For FSS, we use the test split provided by the dataset authors, and derive train and validation classes by forming a validation set with as many classes as the test set, then forming the training set from leftover classes. For datasets with multiple classes per image, we remove all test images from the training and validation sets.

Table 1 shows dataset statistics which indicate relative difficulty of the respective classification tasks. Datasets with more classes and many images per class such as ImageNet have historically been demonstrated to be easier for few-shot tasks. We expect datasets with a high number of classes per image like VG and COCO to benefit most from localization as this statistic is a key indicator for how cluttered typical images are. Mean Area/Sample is the average number of pixels for an annotation, with objects covering less area tending to be more difficult to localize and classify.

[1] `https://github.com/pytorch/vision/blob/master/torchvision/models/resnet.py`

[2] `https://github.com/CSAILVision/unifiedparsing`

[3] `https://github.com/daviswer/fewshotlocal`

	Dataset	Samples	Classes	Imgs/Class	Classes/Img	Mean Area/Sample
B-Box	**ImageNet**	593,233	1,000	593	1	0.49
	iNaturalist	282,898	530	593	1	0.23
	COCO	251,855	80	3,148	2.13	0.193
	Visual Genome	1,027,725	841	1,222	9.69	0.189
Segment	**FSS-1000**	10,000	1,000	10	1	0.234
	COCO	251,855	80	3,148	2.13	0.151
	OpenImages	1,261,147	345	3,656	1.41	0.172
	PASCAL VOC	4,318	20	1,222	1.48	0.17

Table 1: Summary statistics for each dataset following preprocessing and before splitting. Samples are counted as image-annotation pairs.

3 Experiments

Although it may be a fair assumption that incorporating localization features should improve a few-shot classifier's accuracy, we first want to verify this empirically and discover whether localization always benefits few-shot models under a variety of conditions. We introduce a systematic analysis of localized few-shot image classifiers combining two axes of comparison.

Importance of Localized Queries The paradigm few-shot classification task assumes that only support examples are labeled, and query examples are totally unlabeled. Consequently, we assume that typical use cases for few-shot image classifiers with localization will have region annotations for support set images only. However, observing how the presence or absence of annotated query images affects model accuracy can provide insight into whether generating region proposals for query images is a worthwhile endeavor. To evaluate this directly, we train and evaluate few-shot classifiers under two major experimental configurations:

- *Oracle*: use ground-truth annotations for both support and query images,
- *Support*: use ground-truth annotations for support images only.

Under the latter configuration, query images are given localizations that encompass the entire image (*i.e.* masks comprised entirely of ones).

End-to-End Model Finally, we evaluate whether few-shot classification performance can be improved upon by incorporating query image localizations generated by a few-shot segmentation model. We train and test our proposed PRoPnet model on a host of vision datasets, testing our architecture on datasets which include either bounding box or segmentation localization features. We compare our approach with the Oracle and Support localization, as well as the method proposed in [32].

3.1 Results and Analysis

We report 5-shot, 5-way few-shot classification accuracy on the test set. Accuracy is averaged over 11000 episodes, each with 5 query image predictions. The first three rows of Table 2 give few-shot classification results for models incorporating different degrees of localization information: no localization, localized support and query images (Oracle), or localized support images only (Support). Comparing models without localization to those with Oracle localization, for all datasets, we observe accuracy gains via early conditioning on full localization information (Oracle). For datasets with one class per image (ImageNet, iNat, FSS) Oracle localization provides marginal gains of less than 2% accuracy. However, datasets with multiple classes per image show greater improvement with full localization conditioning. The three datasets with highest average number of classes per image (VG, COCO, P-5^i) show drastic improvement with over 10% gains in accuracy. For these three datasets which contain many crowded scenes, without query localization to indicate which subjects in an image are relevant, the model must learn to localize *and* classify, a more difficult learning problem. Comparing Oracle to Support localization, model performance drops without localized

	Bounding Box				Segmentation			
Model	**VG**	**COCO**	**iNat**	**ImageNet**	**OpenImg**	**COCO**	**FSS**	**P-5^i**
No Localization	59.27	72.31	93.79	91.50	88.73	72.44	96.21	60.79
EF Support	60.22	69.71	93.09	90.59	87.67	68.84	97.02	60.15
EF Oracle	76.24	82.90	94.05	93.75	92.67	84.51	98.17	73.40
PRoPnet	59.66	71.44	89.42	88.78	88.77	71.84	97.80	67.79
FSL [32]	59.87	73.74	94.61	93.08	90.00	75.09	96.93	62.12

Table 2: 5-shot, 5-way test set accuracy (average over 11000 5-query test set episodes) for few-shot classifiers without localization (**No Localization**), ground truth early fusion localization using support only (**EF Support**), and using both support and query (**EF Oracle**), and two methods with localization by region proposal—our proposed method (**PRoPnet**) and Few-shot Localization (**FSL**) [32].

query images, *especially* for densely-annotated datasets. This indicates the importance of approaches for localization of query images in the busy image few-shot setting.

The last two rows of of Table 2 give results for our proposed early fusion annotation conditioning, PRoPnet, and experiments using a late fusion localization conditioning method, FSL, proposed by [32]. Blue colored table entries indicate improvement of a localization method from the No Localization baseline, with the single red colored entry (P-5^i) indicating the greatest overall performance improvement. The first four columns show results of the respective localization methods using bounding box annotations. The FSL method is able to leverage this information for more accurate classification whereas PRoPnet does not successfully leverage coarse bounding box support localization, having been designed for pixel-level annotations. The last four columns of Table 2 give results on the datasets with fine-grained segmentation annotations. Segmentation datasets containing fewer classes and with the largest average classes per image (P-5^i and COCO), demonstrate the largest gains from the localization conditioning methods. FSL improves accuracy on COCO by ~2.5% and PRoPnet improves accuracy on P-5^i by ~7%.

4 Conclusion and Future Work

In this work we performed a systematic analysis on few-shot classification with localization using several vision datasets with varying statistical properties. A comparison of classification performance with and without ground-truth localized queries reveals that accuracy declines without localized queries for densely-annotated datasets. However, we find that datasets with one class per image can, to a lesser extent, also benefit from ground-truth localization, likely because less confusion about which objects in an image are relevant obviates any explicit need for localization.

We present PRoPnet, an end-to-end model for joint few-shot region proposal and classification to improve accuracy on the difficult few-shot task of classifying objects in busy natural scenes. For segmentation datasets, PRoPnet shows improvement over classification with no localization and with ground-truth localized support images for datasets where cluttered natural scenes are commonplace. In addition, datasets with fewer classes benefit most from conditioning on region proposals despite being more challenging in the few-shot context, suggesting that these additional features are more beneficial to highly-constrained few-shot classification tasks. Overall, our results suggest that full localization information is essential for classifying objects in cluttered natural scenes, and two-stage networks like PRoPnets and FSL are promising approaches to generate query region proposals which can provide decisive context.

For future work we plan to extend our two-stage approach to effectively utilize coarser grained annotation information in the form of bounding boxes by replacing the first stage few-shot segmentation architecture with a few-shot object detection architecture.

References

[1] AZAD, R., FAYJIE, A. R., KAUFFMAN, C., AYED, I. B., PEDERSOLI, M., AND DOLZ, J. On

the texture bias for few-shot CNN segmentation. *arXiv preprint arXiv:2003.04052* (2020).

[2] BERMAN, M., RANNEN TRIKI, A., AND BLASCHKO, M. B. The Lovász-Softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 4413–4421.

[3] BHAT, G., LAWIN, F. J., DANELLJAN, M., ROBINSON, A., FELSBERG, M., VAN GOOL, L., AND TIMOFTE, R. Learning what to learn for video object segmentation. *arXiv preprint arXiv:2003.11540* (2020).

[4] DENG, J., DONG, W., SOCHER, R., LI, L.-J., LI, K., AND FEI-FEI, L. ImageNet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (2009), Ieee, pp. 248–255.

[5] EVERINGHAM, M., VAN GOOL, L., WILLIAMS, C. K., WINN, J., AND ZISSERMAN, A. The PASCAL Visual Object Classes (VOC) challenge. *International journal of computer vision 88*, 2 (2010), 303–338.

[6] FEYJIE, A. R., AZAD, R., PEDERSOLI, M., KAUFFMAN, C., AYED, I. B., AND DOLZ, J. Semi-supervised few-shot learning for medical image segmentation. *arXiv preprint arXiv:2003.08462* (2020).

[7] FINN, C., ABBEEL, P., AND LEVINE, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70* (2017), JMLR. org, pp. 1126–1135.

[8] FORT, S. Gaussian prototypical networks for few-shot learning on Omniglot. *arXiv preprint arXiv:1708.02735* (2017).

[9] FORTIN, M. P., AND CHAIB-DRAA, B. Few-shot learning with contextual cueing for object recognition in complex scenes. *arXiv preprint arXiv:1912.06679* (2019).

[10] GAO, T., HAN, X., LIU, Z., AND SUN, M. Hybrid attention-based prototypical networks for noisy few-shot relation classification. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2019), vol. 33, pp. 6407–6414.

[11] HE, K., ZHANG, X., REN, S., AND SUN, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, year=2016.

[12] KARLINSKY, L., SHTOK, J., ALFASSY, A., LICHTENSTEIN, M., HARARY, S., SCHWARTZ, E., DOVEH, S., SATTIGERI, P., FERIS, R., BRONSTEIN, A., ET AL. StarNet: Towards weakly supervised few-shot detection and explainable few-shot classification. *arXiv preprint arXiv:2003.06798* (2020).

[13] KRISHNA, R., ZHU, Y., GROTH, O., JOHNSON, J., HATA, K., KRAVITZ, J., CHEN, S., KALANTIDIS, Y., LI, L.-J., SHAMMA, D. A., ET AL. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision 123*, 1 (2017), 32–73.

[14] KUZNETSOVA, A., ROM, H., ALLDRIN, N., UIJLINGS, J., KRASIN, I., PONT-TUSET, J., KAMALI, S., POPOV, S., MALLOCI, M., KOLESNIKOV, A., ET AL. The Open Images dataset V4. *International Journal of Computer Vision 128* (2020), 1956–1981.

[15] LAKE, B. M., SALAKHUTDINOV, R., AND TENENBAUM, J. B. The Omniglot challenge: a 3-year progress report. *Current Opinion in Behavioral Sciences 29* (2019), 97–104.

[16] LI, X., WEI, T., CHEN, Y. P., TAI, Y.-W., AND TANG, C.-K. FSS-1000: A 1000-class dataset for few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 2869–2878.

[17] LIFCHITZ, Y., AVRITHIS, Y., AND PICARD, S. Few-shot few-shot learning and the role of spatial attention. *arXiv preprint arXiv:2002.07522* (2020).

[18] LIN, J., AND HE, X. Few-shot learning with weakly-supervised object localization. *arXiv preprint arXiv:2003.00874* (2020).

[19] LIN, T.-Y., MAIRE, M., BELONGIE, S., HAYS, J., PERONA, P., RAMANAN, D., DOLLÁR, P., AND ZITNICK, C. L. Microsoft COCO: Common objects in context. In *European conference on computer vision* (2014), Springer, pp. 740–755.

[20] LIU, J., AND QIN, Y. Prototype refinement network for few-shot segmentation. *arXiv preprint arXiv:2002.03579* (2020).

[21] LIU, W., ZHANG, C., LIN, G., AND LIU, F. CRNet: Cross-reference networks for few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 4165–4173.

[22] MICHAELIS, C., BETHGE, M., AND ECKER, A. One-shot segmentation in clutter. In *International Conference on Machine Learning* (2018), pp. 3549–3558.

[23] NGUYEN, K., AND TODOROVIC, S. Feature weighting and boosting for few-shot segmentation. In *Proceedings of the IEEE International Conference on Computer Vision* (2019), pp. 622–631.

[24] PAN, Y., YAO, T., LI, Y., WANG, Y., NGO, C.-W., AND MEI, T. Transferrable prototypical networks for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2019).

[25] RUSSAKOVSKY, O., DENG, J., SU, H., KRAUSE, J., SATHEESH, S., MA, S., HUANG, Z., KARPATHY, A., KHOSLA, A., BERNSTEIN, M., ET AL. ImageNet large scale visual recognition challenge. *International journal of computer vision 115*, 3 (2015), 211–252.

[26] SELVARAJU, R. R., COGSWELL, M., DAS, A., VEDANTAM, R., PARIKH, D., AND BATRA, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (2017), pp. 618–626.

[27] SHABAN, A., BANSAL, S., LIU, Z., ESSA, I., AND BOOTS, B. One-shot learning for semantic segmentation. *arXiv preprint arXiv:1709.03410* (2017).

[28] SIAM, M., ORESHKIN, B. N., AND JAGERSAND, M. AMP: Adaptive masked proxies for few-shot segmentation. In *Proceedings of the IEEE International Conference on Computer Vision* (2019), pp. 5249–5258.

[29] SNELL, J., SWERSKY, K., AND ZEMEL, R. Prototypical networks for few-shot learning. In *Advances in neural information processing systems* (2017), pp. 4077–4087.

[30] VAN HORN, G., MAC AODHA, O., SONG, Y., CUI, Y., SUN, C., SHEPARD, A., ADAM, H., PERONA, P., AND BELONGIE, S. The iNaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 8769–8778.

[31] WANG, K., LIEW, J. H., ZOU, Y., ZHOU, D., AND FENG, J. PANet: Few-shot image semantic segmentation with prototype alignment. In *Proceedings of the IEEE International Conference on Computer Vision* (2019), pp. 9197–9206.

[32] WERTHEIMER, D., AND HARIHARAN, B. Few-shot learning with localization in realistic settings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2019), pp. 6558–6567.

[33] WU, Z., CHANG, R., MA, J., LU, C., AND TANG, C. K. Annotation-free and one-shot learning for instance segmentation of homogeneous object clusters. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence* (2018), pp. 1036–1042.

[34] XIAO, T., LIU, Y., ZHOU, B., JIANG, Y., AND SUN, J. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 418–434.

[35] YANG, Y., MENG, F., LI, H., WU, Q., XU, X., AND CHEN, S. A new local transformation module for few-shot segmentation. In *International Conference on Multimedia Modeling* (2020), Springer, pp. 76–87.

[36] ZHANG, C., LIN, G., LIU, F., YAO, R., AND SHEN, C. CANet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2019), pp. 5217–5226.

[37] ZHANG, X., WEI, Y., YANG, Y., AND HUANG, T. S. SG-One: Similarity guidance network for one-shot semantic segmentation. *IEEE Transactions on Cybernetics* (2020).