Finding Four-Leaf Clovers: A Benchmark for Fine-Grained Object Localization

Laura Bravo* Alejandro Pardo* Gustavo Pérez* Pablo Arbeláez
Universidad de los Andes
{lm.bravo10, la.pardo2014, ga.perezs, pa.arbelaez}@uniandes.edu.co

Abstract

We present a new experimental framework with the objective of studying fine-grained object localization problems. The Four-Leaf Clover (FLC) dataset was built with the contribution of trained hobbyists, who were assigned the task of spotting four-leaf clovers on a fixed geographical location over two clover seasons, one season for the train set and another for the test set. Each object instance was then annotated for the tasks of object detection, semantic segmentation, instance segmentation, object parsing and semantic boundary detection. Our dataset is composed of more than 100,000 images, containing 2,151 carefully annotated clover instances of four, five or six leaves. The Four-Leaf Clover Dataset is extremely challenging due to its small inter-class variance and its very large intra-class variation. We perform extensive experiments with state-of-the-art methods in order to establish strong baselines for each of the tasks. The dataset and all its resources will be made publicly available.

1. Introduction

Fine-grained object recognition is a challenging open problem, in which the goal is to distinguish subordinate categories within entry-level categories. In other words, fine-grained recognition and, specifically, fine-grained localization is akin to looking for a needle in a haystack. Fine-grained image classification has attracted a growing interest from the computer vision community in recent years. Since 2011, a workshop for fine-grained visual categorization (FGVC) has been organized in conjunction with CVPR. In 2018, the fourth version of FGVC presented two main challenges: visual categorization of materials [29] and visual categorization of species of plants and animals [25]. The goal of these competitions is to push the state-of-the-art in automatic-image classification across fine-grained categories.

*Indicates equal contribution.

However, a large variety of applications require the study not only of fine-grained object categorization but also of fine-grained object localization. Fine-grained tasks are defined as having a small inter-class variation, that is distinguishing between extremely similar objects; and a large intra-class variation, namely identifying highly variable objects as pertaining to the same class. Anomaly detection is a major example of such applications, which consists in finding an unusual pattern that does not match an expected behavior or visual appearance. For instance, video surveillance of crowded scenes demands the ability to detect subtle deviations from normal crowd behaviors [15, 12, 13]. Another fine-grained localization example is the diagnosis of lung cancer [1, 23]. With the purpose of diagnosing this disease in early stages, lung nodules with a size of only 3mm need to be detected and separated from visually similar lung structures in 3D images from a human thorax.

To the best of our knowledge, the computer vision community currently lacks a unified benchmark to study the different tasks associated with fine-grained localization problems. On the one hand, datasets for fine-grained image categorization [11, 20, 26, 16, 27] have low inter-class variation, but the objects to be recognized are large and centered in the images, eliminating the object localization problem. On the other hand, datasets for object localization [5, 14, 22, 30, 2] exhibit high inter-class variation, making them inadequate for fine-grained problems. Our main contribution in this paper is to present a comprehensive new experimental framework, with highly detailed annotations, for studying fine-grained object localization tasks.

We introduce the Four-Leaf Clover (FLC) Dataset, that contains 5 different fine-grained localization tasks: object detection, semantic segmentation, instance segmentation, instance parsing, and semantic boundary detection. Fine-grained problems are characterized by being highly unbalanced and, most importantly, by their low inter-class and high intra-class variation. Clovers normally have three leaves, but through mutations four-leaf specimens can arise. These mutations are uncommon, with only about 1 in 10,000 clover plants having four leaves [24] in the wild.

Figure 1. (Left panel) Example of an image of the FLC dataset. (Right panel) Positive samples of four-leaf clovers on the image on the left.

This low mutation rate makes spotting four-leaf clovers an extremely difficult fine-grained recognition problem, often associated to good luck in popular culture. Figure 1 illustrates the difficulty of detecting four leaf clovers in the FLC dataset.

The low inter-class and high intra-class variation of object localization problems are intrinsic characteristics of the tasks within the FLC experimental framework. Three-leaf and four-leaf clovers are extremely similar (low inter-class variation), while several clover species are found in the dataset (high intra-class variation). Moreover, variability in factors such as different positions, coloring, and lighting are additional challenges of the dataset. We provide a comprehensive benchmark that allows for object detection, semantic segmentation, instance segmentation, instance parsing, and semantic boundary detection.

2. Related Work

Table 1 shows a summary of the comparison between different publicly available datasets and the FLC dataset. Below we present a brief review of the most relevant datasets. We divide it into two families: the first one studies fine-grained object recognition, focusing on image classification; while the second one studies object localization, but not at a fine-grained level.

Fine-grained categorization datasets. The Stanford Dogs dataset [11] contains 120 dog species with bounding box and class annotations, and was presented as a fine-grained benchmark due to its reduced inter-class and large intra-class variation. Analogous to Stanford Dogs, Caltech-UCSD Birds 200 (CUB-200) [26] and Cats and Dogs [20] are datasets that were created to address the problem of fine-grained image categorization. They contain additional annotations of bounding boxes and pixel-level segmentation of the image object. In order to have finer descriptions of images, another family of datasets to study this problem has also included object attribute annotations. CUB-200 includes 25 bird attributes per image, related to color and patterns of each bird species. Fine-Grained Visual Classification Aircraft (FGVC-A) [16] and CompCars [27], are datasets that introduce attributes of each aircraft/car related to its manufacturer, model, family, and other finer characteristics in order to create more accurate tasks and train finer classifiers. VegFru, another dataset for image categorization, was published at ICCV 2017 and contains images of vegetables and fruits [10]. Finally, Fine-Grained Car Detection for Visual Census Estimation (CDVCE) [6] is a dataset that combines the tasks of car detection with fine-grained categorization with more than 2000 car classes. However, these datasets [11, 26, 20, 16, 27, 10, 6] do not allow the study of all object recognition problems on a fine-grained level due to the fact that most of the images contain a single large centered object instance. This issue greatly restricts the semantic segmentation and object detection tasks, making these datasets useful mainly for the fine-grained image categorization task they were originally designed to address.

Object Localization datasets. Imagenet [3], Pascal [5], and MS-COCO [14] datasets include images with several instances and classes with annotations for detection problems and object segmentation. Pascal also includes boundaries and pixel level labels within the additional annotations of SBD [7] and Pascal Context [19] respectively. Additionally, the DAVIS challenge addresses the same problem as Pascal and MS-COCO, with semantic segmentation [21] and object segmentation [22] in videos. The ADE20K [30] dataset includes labels for semantic segmentation and scene parsing, focusing on the problem of scene understanding. The Cityscapes [2] dataset provides images to address urban scene understanding, with annotations for scene categorization, object detection, and semantic and instance segmentation. Another group of datasets incorporates finer annotations by including amodal detections [4] and segmentations [31]. These annotations include additional information of non visible pixels in the image by marking the full extension of the object. All of these datasets were created to address object recognition problems. However, none of them allows to study the problem at a fine-grained level.

Therefore, the community currently has fine-grained datasets, but only for image categorization; and datasets for object localization, but not on a fine-grained level. To our knowledge, the FLC is the first dataset for fine-grained object localization whose target objects represent a small portion of the image, in addition to including several object

Dataset	Detection	Semantic Sem.	Instances	Boundaries	Parsing	Classification
Imagenet [3]	△	×	×	×	×	×
Pascal [5, 19, 7]	△	△	△	△	×	×
COCO [14]	△	×	△	×	×	×
DAVIS [22, 21]	△	×	△	×	×	×
ADE20K [30]	△	△	△	×	△	×
CityScapes [2]	△	△	△	×	×	×
Stanford Dogs [11]	△	×	×	×	×	♣
Cats & Dogs [20]	△	△	×	×	×	♣
CUB-200 [26]	△	△	×	×	×	♣
CompCars [27]	△	×	×	×	×	♣
VegFru [10]	×	×	×	×	×	♣
FLC	♣	♣	♣	♣	♣	×

Table 1. Comparison of FLC to major visual recognition datasets. Clubsuit (♣) indicates that a dataset allows to study a recognition problem at a fine-grained level, triangle (△) indicates that the version of the problem is not fine-grained, and (×) indicates that a dataset does not allow to study a problem. The first six rows correspond to object recognition datasets that lack a fine-grained nature, while the next five rows present examples of fine-grained image classification datasets.

instances per image. The FLC dataset includes multiple object recognition tasks at a fine-grained level such as instance and semantic segmentation, object detection, object parsing, and semantic boundary detection.

3. The Four-Leaf Clover Dataset

3.1. Dataset description

The FLC dataset was created as a benchmark to study fine-grained object localization. All images were collected by trained hobbyists with smart-phones that capture image of approximately 8 to 10 Mega-pixels. The collection of the images was made into two different clover seasons, one season for train and validation sets, and the other one for test set. The current version of the dataset is composed of more than 100,000 images, both positive, containing at least one four-leaf clover (for simplicity we refer to clovers with four or more leaves as four-leaf clovers), and negative, containing only common three-leaf clovers. The authentic nature of each four-leaf clover was verified *in situ*, and negative images were captured after verifying that the patch did not contain any four-leaf clovers. The images were captured from approximately the same distance and viewpoint as that of the trained hobbyists who spotted the clovers. The high class imbalance shown in Table 2 was intentionally made to approximate the natural distribution between four-leaf and three-leaf clovers (1 in 10,000).

Figure 2. Examples of detail in segmentation annotations.

Table 2. FLC dataset statistics. 4-leaf clover pixels and 4-leaf clover boundary pixels refer to the rate of the total of positive pixels over the total of pixels in the FLC dataset.

General statistics	Trainval set	Test set
Total positive images	1,000	500
Total negative images	51,637	51,670
Total images	52,637	52,170
4-leaf clover instances	1,412	739
4-leaf clover leaves	5,858	3,094
4-leaf clover pixels	0.0445%	0.0588%
4-leaf clover boundary pxls.	0.0026%	0.0030%

The purpose of the FLC dataset is the study of fine-grained tasks simultaneously with increasing level of detail. Figure 2 shows examples of the level of detail in the annotations of the FLC dataset. Each four-leaf clover instance is annotated with a segmentation mask and a bounding box. In accordance with the fine-grained nature of the problem, the dataset also provides segmentation masks of the individual leaves and their boundaries. These rich annotations allow for the study of five different but complementary fine-grained tasks within the experimental framework of the FLC dataset: object detection, semantic segmentation, instance segmentation, object parsing and semantic boundary detection.

3.2. Human consistency

In order to assess the intrinsic difficulty of the fine-grained localization problem within the FLC experimental framework, and to provide an upper-bound for algorithmic performance, four trained subjects, with different levels of training, annotated independently all the images of the FLC dataset. We then evaluated the subjects' predictions against the ground-truth labels collected *in situ*. For a human, the

Table 3. *Left*. Human consistency results against ground-truth. *Right*. Second human consistency experiment results in which every subject's predictions were compared against the remaining subjects as ground-truth. Final values are mean and deviation per experiment metric.

Subject	Precision	Recall	F1	Av. Time
1	0.81	0.71	0.76	1'04"±0.36 "
2	0.78	0.73	0.75	1'09"±0.04 "
3	0.80	0.84	0.82	1'06"±0.05 "
4	0.74	0.60	0.66	1'10"±0.10 "
Mean	**0.78**	**0.72**	**0.74**	**1'07"**

Precision*	Recall*	F1*
0.77 ±0.1	0.62 ±0.01	0.69 ±0.05
0.73 ±0.11	0.65 ±0.02	0.69 ±0.06
0.69 ±0.08	0.66 ±0.08	0.67 ±0.06
0.74 ±0.04	0.55 ±0.02	0.63 ±0.003
0.73	**0.62**	**0.67**

challenge lies in detecting the four-leaf clovers; once this task is accomplished, the remaining ones are straightforward and, given enough time, can be accomplished with arbitrary level of detail. For this reason, the human predictions were bounding boxes and we evaluated them as a detection task. Additionally, the annotators timed themselves, providing a proxy for the difficulty of the task in individual images.

Table 3 (*Left*) depicts the results of the human consistency experiment with respect to the *in situ* gold standard. Overall, the annotators were highly consistent in their detections, in both their average time and ability to detect the four-leaf clovers. We observed a correlation between the F1 score and the annotators' experience. Nevertheless, the recall shows that humans miss approximately 30% of the complete set of four-leaf clovers.

To further assess human consistency in this task, we compared the predictions provided by the annotators to each other and averaged over candidate annotator. This experiment was performed by setting the detections of a given annotator as the prediction and the others as the ground-truth. This process was repeated for every annotator following the precedent set by the BSDS500 dataset [18].

Table 3 (*Right*) shows the results of the second human consistency experiment. We can observe that the subjects are more consistent among them than when compared against the *in situ* ground-truth, particularly in their recall. Thus, the gap between the human F-measure and the perfect prediction can be attributed to domain-specific difficulties.

3.3. Dataset challenges

Although all the images in the FLC dataset were acquired in the same geographical location, the dataset includes several complexities inherent to the environment and the application domain. Figure 3 shows examples of these challenges: changes in clover size, light variation due to the time of day of acquisition and weather conditions, shadows and occlusions created by other objects or clovers, leaf 3D orientation, and shape. All these challenges increase the difficulty of the proposed fine-grained tasks due to the augmented intra-class variation.

(a) (b) (c) (d) (e) (f)

Figure 3. Challenges of the FLC dataset. **(a)** Lighting and color variations. **(b)** Cast shadows. **(c)** Occlusion. **(d)** Leaf shape. **(e)** Leaf 3D orientation. **(f)** Different clover species.

3.4. Fine-grained object localization tasks

The FLC dataset contains five fine-grained tasks with increasing level of detail. In this section, we describe each of these tasks and the evaluation methodology to assess them. The FLC annotation format is compatible with MS-COCO and PASCAL, in the interest of allowing a transparent use of the metrics and evaluation code proposed by reference benchmarks.

3.4.1 Object Detection.

In this task, the purpose is to determine the location of a four-leaf clover in an image. The annotations are the coordinates of the lower left end and the upper right end of the bounding box surrounding each four-leaf clover. For the evaluation of an algorithm, we use the mean average precision (mAP) at a 0.5 overlap (intersection over union (IoU) between annotation and prediction bounding boxes) as used in PASCAL VOC [5] (denoted as mAP@.5); and the averaged mAP with overlap of IoU $\in$ [0.50: 0.05: 0.95] which is MS-COCO's [14] standard metric for detection (denoted as mAP@[.5, .95]). Figure 4a shows an example of detection labels of four-leaf clovers in an FLC image.

3.4.2 Semantic Segmentation.

When semantically segmenting an image, the goal is to classify all pixels corresponding to a specific class or classes. In

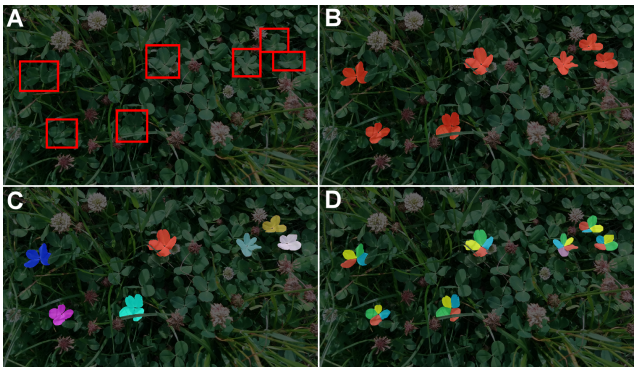

Figure 4. Annotation examples within the same sample image of Figure 1. **(a)** Annotations of four-leaf clover detections. **(b)** Annotations of four-leaf clover semantic segmentation. **(c)** Annotations of four-leaf clover instance segmentation. **(d)** Annotations of four-leaf clover parsing.

our case, the objective is to classify all four-leaf clover pixels. Our annotations are masks for all four-leaf clovers in each image. We can see an example of the annotations in Figure 4b. For the evaluation of an algorithm, we use the mean intersection over union (mIoU) as used in PASCAL segmentation [5] and ADE20k [30].

3.4.3 Instance Segmentation.

In this task, the main goal is simultaneous detection and segmentation [8] of four-leaf clover instances. The annotations are specific masks for each of the instances in the dataset, we show an example in Figure 4c. We evaluate the performance on our dataset using the same evaluation metrics of MS-COCO and PASCAL [8] for instance segmentation. MS-COCO uses Average precision (AP) at mask overlap (IoU) $\in$ [0.50: 0.05: 0.95]. While PASCAL reports only AP at IoU at 0.5.

3.4.4 Object Parsing.

In object parsing, we are interested not only on the object instances but also on their parts. We want to detect each of the parts and be able to analyze how they interact at the whole object scale. The difficulty of this finer task increases considerably. On the FLC dataset, we want to segment and detect each leaf of a four-leaf clover as a separate instance. An annotation example for this task is shown in Figure 4d. We use the same evaluation metrics proposed in the Instance Segmentation task, setting as annotations the individual leaves of four-leaf clovers.

3.4.5 Semantic Boundary Detection.

The aim of this task is to find boundaries defined by their semantic meaning, in particular only the boundaries of four-leaf clovers. To get a sense of the scale of the problem, the proportion of positive pixels vs. negative pixels in the FLC, shown in Table 2, indicates that there are 100,000 negative pixels per every 31 positive pixels. Accordingly, this is one of the most difficult fine-grained tasks available in the FLC dataset. The evaluation of the proposed methods for this task is a Precision-Recall (PR) curve, the area it encompasses (Average Precision (AP)) and the maximal F-measure (F1), introduced in the BSDS [18] and the SBD dataset [7]. The result of a given method is calculated from the soft boundaries produced on the test set, which are then thresholded to obtain individual points of the PR curve.

3.5. Baselines

Existing object localization datasets are not adequately structured to study fine-grained tasks. To quantitatively assess the need for a new dataset in which better fine-grained object localization methods can arise, and to set strong baselines for future reference, we evaluated state-of-the-art methods on every task available in the FLC dataset. For every method we report results of the complete test set and a subset of the test set composed of only positive images.

4. Experiments

4.1. Object Detection

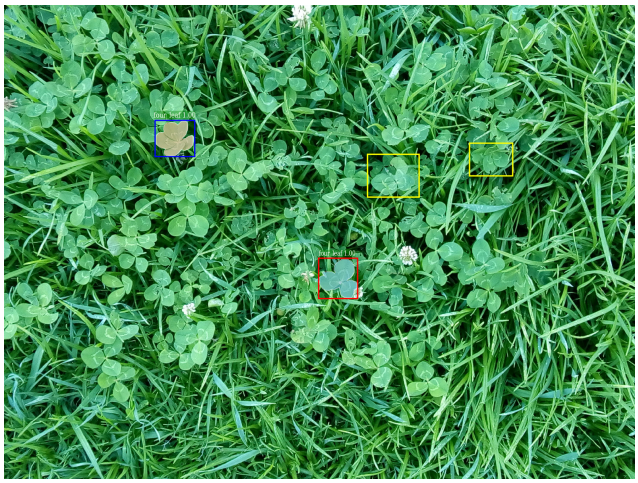

Figure 5. Qualitative results of the detection and instance segmentation task in the test set. This example shows one false positive (red), two false negatives (yellow) and a true positive (blue).

We evaluated the detection task using the Mask R-CNN network [9]. We retrained starting from ImageNet weights, using Resnet-50 with Feature Pyramid Network (FPN) as the backbone for the Mask R-CNN. We trained the model for 10 epochs using the suggested training settings.

Table 4 presents results on the test set. We see that the performance of Mask R-CNN in the FLC dataset is substantially lower in terms of mAP@.5 and mAP@[.5,.95]

Table 4. Comparison of results using Mask R-CNN for the task of detection of four-leaf clovers

Dataset	Method	mAP@.5(%)	mAP@[.5, .95] (%) all	mAP@[.5, .95] (%) small	mAP@[.5, .95] (%) medium	mAP@[.5, .95] (%) large
FLC test (Pos. only)	Mask R-CNN ResNet50+FPN	**44.9**	**36.4**	**0.0**	**24.5**	**36.7**
FLC test	Mask R-CNN ResNet50+FPN	**9.50**	**7.90**	**0.0**	**0.10**	**10.7**

than those reported by [9] in MS-COCO. However, when considering only the positive images in the test set both mAPs, while still lower, are comparable to those stated in MS-COCO. These results can be explained by the high false positive rate of Mask R-CNN when detecting four-leaf clovers, which soars when including the negative images found in the test set. Additionally, the results show that Mask R-CNN's efficacy is tied to the size of a four-leaf clover: the method accurately identifies large clovers, but completely misses small clovers. This result is consistent with the intrinsic difficulty of the FLC dataset, smaller clovers tend to be occluded and easily lost in the crowd. Figure 5 shows Mask R-CNN's prediction on an image from the test set. The method manages to find a five-leaf clover correctly, but produces two false negatives that are not occluded nor in a particularly crowded area. The method also detects an archetypal three-leaf clover as false positive with a high probability.

4.2. Instance Segmentation

To address the problem of instance segmentation, we used the masks produced by Mask R-CNN [9]. We employed ResNet-50 as backbone starting with the pre-trained ImageNet weights and trained for 10 epochs. We used the same parameters for training and for the region proposal networks used in the original implementation of MS-COCO.

Figure 6. Qualitative results of Mask R-CNN's high false positive rate in the instance segmentation task. We show examples of segmentations on negative images from the test set. We can observe the high false positive ratio that Mask R-CNN has, detecting four four-leaf clover instances in a negative image.

Table 5 shows the results of the instance segmentation task on the FLC dataset. As expected, we observe that the performance of the Mask R-CNN method for the instance segmentation task is highly consistent with the object detection task. Again, when we included negative images performance decreased substantiially. Figure 6 shows an example of the method's performance in the complete test set. We can see a true positive and a false positive in the prediction, illustrating the difficulty of the task of differentiating between a three-leaf clover and a four-leaf clover.

Table 5. Comparison between results in test set of Mask R-CNN for the instance segmentation task. We show results within only positive (Pos.) samples of the set (Test-Positives) and within the full set (FLC Test).

Dataset	Method	mAP@[.5,.95](%)
FLC(Pos.)	MaskR-CNN R50	**35.0**
FLC	MaskR-CNN R50	**8.0**

4.3. Semantic Segmentation

Table 6. Results using Mask R-CNN for the task of semantic segmentation of four-leaf clovers.

Dataset	Method	mIoU (%)
FLC (Positives)	MaskR-CNN R50	**32.71**
FLC	MaskR-CNN R50	**7.71**

We provide a baseline for the semantic segmentation task by using the outputs of our retrained Mask R-CNN models for instance segmentation as semantic segmentation masks. We report the results on the test set in Table 6. As expected, the results when the negatives samples are included are much lower than the results with only positive samples. The reason might be that models like Mask-RCNN are trained and designed to always detect objects in the images and thus produce a large amount of false positive detections when faced with negative images.

4.4. Object Parsing

For this task, we used the same pipeline as the one used for instance segmentation. Since object parsing is a more detailed task and Mask R-CNN was not originally designed for this, we trained the ResNet-50 model for 20 epochs instead of 10 as the original implementation of Mask R-CNN

for MS-COCO. For consistency, we used the same learning rate and configuration for the region-proposal network as the one used for MS-COCO in the original implementation.

Table 7 shows the results for the object parsing task. We obtain a lower mAP than for clover segmentation. In this task, a positive and a negative candidate have relatively similar visual cues; the differentiating factor lies at the clover scale, not at the instance-level configuration of leaves. So, the low performance of a method created to detect and segment large objects that have a lot of variation in their visual aspect is not surprising. We can see in Figure 8 one of the best results obtained within the test set with the model trained by 90 epochs. We can observe that Mask R-CNN finds two of the leaves of the four-leaf clover. The results show the complexity of this task, in which a leaf of a three-leaf clover or a leaf of a four-leaf clover have the same visual aspect and the only way to differentiate them is by using information of their spatial context.

Figure 7. Qualitative results of the semantic segmentation task in the test set. **Upper.** Label. **Lower.** Prediction with Mask R-CNN.

Table 7. Comparison between results in test set of Mask R-CNN for the parsing task within only positive (Pos.) samples of the test set (Test-Positives) and within the full test set (FLC Test).

Dataset	Method	mAP@[.5,.95] (%)
FLC(Pos.)	MaskR-CNN R50	**14.0**
FLC	MaskR-CNN R50	**3.1**

Figure 8. Qualitative results of the object parsing task in the test set. **Left.** Label. **Right.** Prediction.

4.5. Semantic Boundary Detection

We retrained two semantic boundary detection models: one without semantic label understanding, COB [17], and another with it, CASENet [28]. We retrained the COB models for 1000 epochs starting from the BSDS500 [18] and the PASCAL VOC12 Segmentation database (SegVOC12) [5] weights. For the CASENet model we started from the SBD weights [7] and trained for 200 epochs. Also, for CASENet we adapted the architectures for one class by removing splice and shared concatenation. Additionally, we performed an experiment in which we compare the effect of fine-tuning the COB and CASENet models by using the original weights on our test set.

Figure 9 shows the results of the semantic boundary detection task. These experiments evidence that at high levels of confidence the performance of both methods is similar, but COB has higher recall at lower levels of confidence. The difficulty of the task is highlighted again by the low AP scores of all methods. However, the points gained with the trained COB models over the untrained models suggest that the task can be learned, and that existing methods would benefit from an empirical framework such as FLC to increase their performance.

Figure 10-right evidences the high precision of the CASENet model, presenting a strong response where the annotation points to a four-leaf clover. Contrastingly, Figure 10-center evidences the high recall of the trained COB models, which respond to objects other than four-leaf clovers, but with finer boundaries. Regarding the source

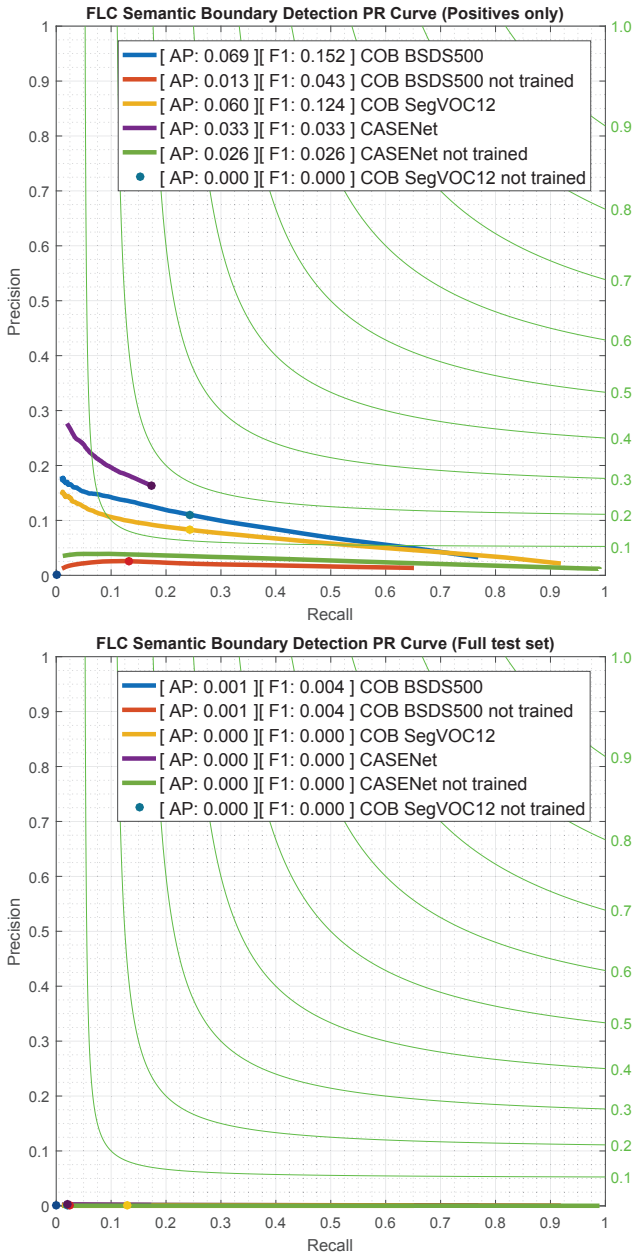

Figure 9. Semantic Boundary detection task PR curves in the test set of 5 models: untrained and retrained (using the BSDS500 and SegVOC12 weights) COB on the FLC, and retrained CASENet. **Top** PR curves evaluated over positive images only. **Bottom** PR curves evaluated over the entire test set.

of the pre-trained weights, the COB BSDS500 model was better than the SegVOC12 model initially. However, after training, the results point to the model benefiting more from the information previously learned on the SegVOC12 database.

The results shown for each of the tasks confirm that our dataset presents a novel set of interesting problems that can

Figure 10. Qualitative results of the semantic boundary detection task in the test set. **Left.** boundary annotation of a four-leaf clover. **Center.** Result of the trained COB model. **Right.** Result of the CASENet model.

lead the computer vision community to achieve significant advances in object localization on a much finer level. We present additional qualitative results for every task in the supplementary material.

5. Conclusions

We introduce the Four-Leaf Clover Dataset, a new benchmark for studying fine-grained object localization problems. We evaluated the human performance on the dataset and found evidence of the difficulty of the problem, as well as the consistency of human predictions. An important characteristic of the FLC dataset is the availability of increasingly precise fine-grained localization tasks: semantic segmentation, object detection, instance segmentation, object parsing, and semantic boundary detection. We propose an experimental framework for each task, testing state-of-the-art algorithms on each of them. Our results strongly indicate that straightforward domain adaptation from related recognition tasks is insufficient to solve fine-grained localization. Thus, we hope that the availability of our dataset will spur the appearance of new ideas and methods for this largely unexplored and yet critical aspect of visual recognition. We will make the FLC and all of its resources (images, annotations with COCO and PASCAL format, evaluation code, pre-computed results, and models) public to guarantee that it serves as a test-bed for the generalization of emerging ideas in other fine-grained application domains.

References

[1] Kaggle Data Science Bowl 2017. https://www.kaggle.com/c/data-science-bowl-2017#description, 2017. 1

[2] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 2, 3

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database.

In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 2, 3

[4] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 2

[5] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html. 1, 2, 3, 4, 5, 7

[6] T. Gebru, J. Krause, Y. Wang, D. Chen, J. Deng, and L. Fei-Fei. Fine-grained car detection for visual census estimation. In *AAAI*, volume 2, page 6, 2017. 2

[7] B. Hariharan, P. Arbelaez, L. Bourdev, S. Maji, and J. Malik. Semantic contours from inverse detectors. In *The IEEE International Conference on Computer Vision (ICCV)*, 2011. 2, 3, 5, 7

[8] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In *European Conference on Computer Vision (ECCV)*, 2014. 5

[9] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2980–2988. IEEE, 2017. 5, 6

[10] S. Hou, Y. Feng, and Z. Wang. Vegfru: A domain-specific dataset for fine-grained visual categorization. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 2, 3

[11] A. Khosla, N. Jayadevaprakash, B. Yao, and F.-F. Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC)*, 2011. 1, 2, 3

[12] L. Kratz and K. Nishino. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 1

[13] W. Li, V. Mahadevan, and N. Vasconcelos. Anomaly detection and localization in crowded scenes. *IEEE transactions on pattern analysis and machine intelligence*, 36(1):18–32, 2014. 1

[14] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common Objects in Context. In *European conference on computer vision (ECCV)*, 2014. 1, 2, 3, 4

[15] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos. Anomaly detection in crowded scenes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010. 1

[16] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013. 1, 2

[17] K.-K. Maninis, J. Pont-Tuset, P. Arbeláez, and L. Van Gool. Convolutional oriented boundaries. In *European Conference on Computer Vision (ECCV)*, 2016. 7

[18] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *The IEEE International Conference on Computer Vision (ICCV)*, 2001. 4, 5, 7

[19] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille. The role of context for object detection and semantic segmentation in the wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 2, 3

[20] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar. Cats and dogs. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 1, 2, 3

[21] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 3

[22] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, and L. Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 1, 2, 3

[23] A. Setio, A. Traverso, T. De Bel, M. S. Berens, C. van den Bogaard, P. Cerello, H. Chen, Q. Dou, M. E. Fantacci, B. Geurts, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge. *Medical Image Analysis*, 42:1–13, 2017. 1

[24] M. Shipman. That four-leaf clover you found may not be a four-leaf clover, 2013. https://phys.org/news/2013-06-four-leaf-clover.html. 1

[25] G. Van Horn, O. Mac Aodha, Y. Song, A. Shepard, H. Adam, P. Perona, and S. Belongie. The iNaturalist Challenge 2017 Dataset. *arXiv preprint arXiv:1707.06642*, 2017. 1

[26] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010. 1, 2, 3

[27] L. Yang, P. Luo, C. Change Loy, and X. Tang. A large-scale car dataset for fine-grained categorization and verification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1, 2, 3

[28] Z. Yu, C. Feng, M.-Y. Liu, and S. Ramalingam. Casenet: Deep category-aware semantic edge detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 7

[29] X. Zhang, Y. Cui, Y. Song, Y. Li, H. Adam, and S. Belongie. The iMaterialist Challenge 2017 Dataset. *Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC)*, 2017. 1

[30] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ade20k dataset. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1, 2, 3, 5

[31] Y. Zhu, Y. Tian, D. Mexatas, and P. Dollár. Semantic amodal segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2