Supplementary Material of Visual Prompting for Generalized Few-shot Segmentation: A Multi-scale Approach

Mir Rayat Imtiaz Hossain[1,2] Mennatullah Siam[3,1] Leonid Sigal[1,2,4] James J. Little[1]
[1]University of British Columbia [2]Vector Institute for AI [3] Ontario Tech University
[4]Canada CIFAR AI Chair

Abstract

This document provides additional material that is supplemental to our main submission. Section A describes the computational efficiency results. Section B includes additional ablation studies. Section C provides split and class-wise performance results, followed by Section D that analyzes the performance on objects of varying size. Section E discusses the performance of our model in a cross-dataset scenario. Section F discusses additional qualitative results on COCO-20^i dataset. Finally, Section G details the societal impact of our work as standard practice in computer vision research.

A. Computational efficiency comparison

In Table 1 in the main submission, we demonstrated the superior performance of our approach, even in the inductive setting, outperforming transductive methods such as DIaM. Transductive approaches necessitate test-time optimization for each data example computing transductive losses, leading to higher inference times. As illustrated in Table A, there exists approximately 20× increase in inference speed of our inductive approach compared to transductive methods like DIaM [1]. Another important observation is that while our model has a higher number of parameters overall, it is computationally efficient as the number of FLOPs is approximately 57% lower.

Model	Learning	mean mIoU	Total Params	FLOPs	Inference Time
Ours	Inductive	**54.79**	69.19M	**55.16G**	**0.015s**
DIaM [1]	Transductive	53.00	**46.72M**	128.26G	0.32s

Table A. Parameter and 1-shot inference time comparison on PASCAL-5^i dataset. FLOPs calculation is done for forward pass only and is computed using the flopth library https://pypi.org/project/flopth/. To compute FLOP we used a image size 417×417.

B. Additional Ablation Studies

B.1. Number of Training Iterations

In Figure A, we present a graph of how our performance varies in the inductive setting for different number of training iterations. While the results reported in Table 1 of the main paper were evaluated after 100 training iterations, both inductive and transductive, the figure reveals that extending the optimization process to a higher number of iterations consistently enhances performance, peaking notably at 300 iterations, but at the expense of run-time.

B.2. Transduction Ablation

As mentioned in Section 4.4 of the original paper, our observations indicate that incorporating transductive losses from the initial iteration results in sub-optimal mIoU on novel classes. This is attributed to inaccurate estimation of the label marginal distribution that is used in the transductive losses. To illustrate this, Figure B showcases the performance of our model in 1-shot inference within a transductive setting, varying the number of iterations at which transduction is applied. Notably, it demonstrates that applying transduction either early or late results in performance degradation.

B.3. Shots Analysis

We conducted an ablation to evaluate the performance of our model in an inductive setting, considering various shot configurations. We compared it with the baseline lacking novel-to-base causal attention and inductive DIaM (without transduction) [1].

In Figure C, we demonstrate the variation in base class performance across different shots. As observed, our method maintains a consistently high base class performance. In contrast, the model without causal attention exhibits a gradual decline in base class performance as the number of shots increases. Our base class mIoU remains largely flat with increasing shots since it has already converged during large-scale training. Notably, inductive DIaM experiences a sharp decline in base performance with the growing number of shots. This observation underscores the

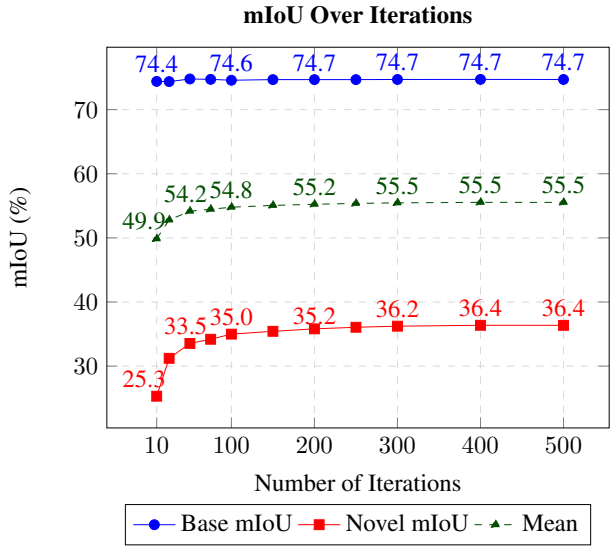

Figure A. 1-shot generalized few-shot segmentation performance for different numbers of training iterations (inductive setting) on Pascal-5^i dataset.

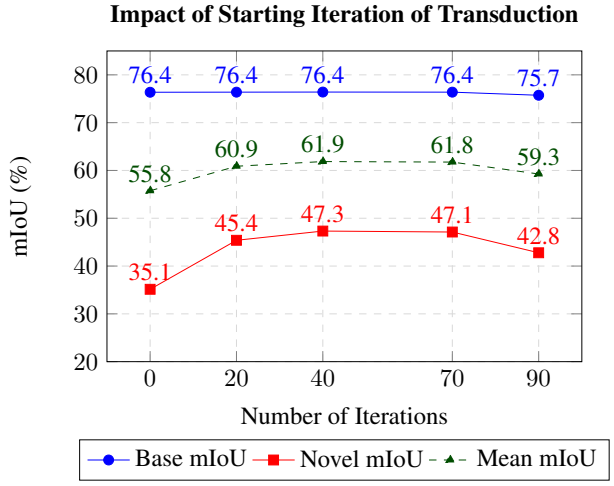

Figure B. 1-shot GFSS performance for the number of iterations at which we start applying transduction losses on Pascal-5^i dataset.

robustness of our model in retaining base class performance as it encounters more examples of novel classes.

Likewise, in Figure D, we illustrate the change of novel class performance across a different number of shots. Notably, our model demonstrates a consistent improvement in novel class performance as the number of shots increases. In contrast, both the baseline without causal attention and inductive DIaM [1] exhibit a plateau in novel class performance around the 5-shot mark. In our model, the leap in performance between different shots is notably more pronounced, demonstrating its capacity for significant improvement with increasing number of examples of the novel class.

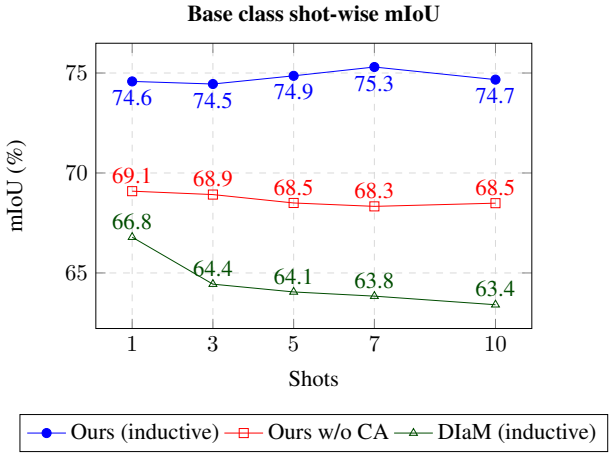

Figure C. Base Class IoU comparison in inductive setting for our approach against both; the baseline without causal attention and inductive DIaM [1], at various support set shots on PASCAL-5^i. w/o CA: without causal attention.

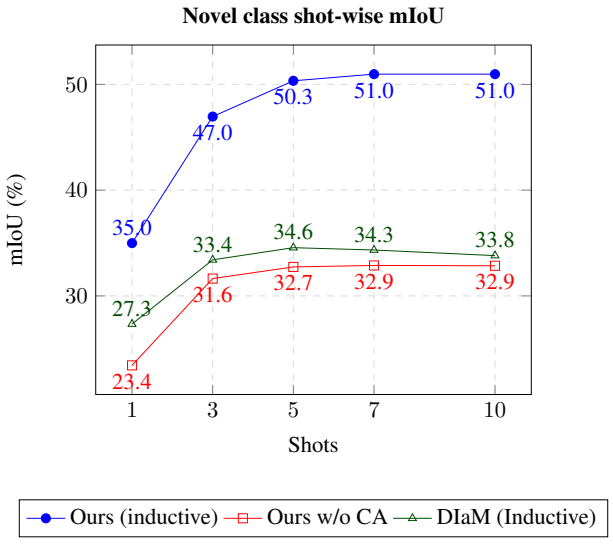

Figure D. Base Class performance comparison in inductive setting for our approach against the baseline without causal attention and inductive DIaM [1] at various support set shots on PASCAL-5^i. w/o CA: without causal attention.

B.4. Quality of Input Prompts

We conducted an ablation on how the quality of the input prompts affect the overall GFSS performance. For this, we corrupted the input support images, that is used to initialize the novel prompts, with different levels of additive Gaussian Noise. Figure E shows the 1-shot GFSS performance for split 0 of Pascal-5^i dataset for different degree of additive Gaussian Noise. As can be observed, there is minimal

Inductive Setting

Dataset	Split	1-shot Base	1-shot Novel	1-shot Mean	5-shot Base	5-shot Novel	5-shot Mean
COCO-20^i	0	50.80	13.45	32.13	50.82	25.54	38.18
	1	50.04	22.07	36.06	50.62	34.35	42.49
	2	53.50	18.50	35.77	53.29	31.41	42.17
	3	51.85	17.99	34.85	51.75	28.95	40.35
	mean	51.55	18.00	34.78	51.59	30.06	40.80
Pascal-5^i	0	75.95	29.86	52.91	76.70	46.72	61.71
	1	72.44	45.59	59.02	73.14	58.09	65.62
	2	72.16	34.08	53.12	72.66	56.09	64.38
	3	77.75	30.43	54.09	76.95	40.45	58.70
	mean	74.58	34.99	54.79	74.86	50.35	62.60

Transductive Setting

Dataset	Split	1-shot Base	1-shot Novel	1-shot Mean	5-shot Base	5-shot Novel	5-shot Mean
COCO-20^i	0	53.59	15.25	34.42	54.68	27.55	41.12
	1	52.37	22.67	37.52	52.39	36.08	44.24
	2	54.93	18.81	36.87	55.12	31.87	43.50
	3	54.31	16.48	35.40	53.05	29.05	41.05
	mean	53.80	18.30	36.05	53.81	31.14	42.48
Pascal-5^i	0	76.62	33.69	55.16	76.65	53.02	64.84
	1	75.46	50.95	63.21	75.47	63.30	69.39
	2	74.64	39.65	57.15	74.67	60.51	67.59
	3	78.82	35.02	56.92	78.89	47.65	63.27
	mean	76.39	39.83	58.11	76.42	56.12	66.27

Table B. GFSS performance for each split in inductive **(top)** and transductive **(bottom)** settings respectively.

degradation of performance, even in cases of higher distortions of additive Gaussian Noise with standard deviation of $\sigma = 30$.

C. Split and Class-wise Results

Split-wise Results. Adhering to the established few-shot segmentation protocol, our model undergoes evaluation across four distinct splits or folds for both the PASCAL-5^i and COCO-20^i datasets. Within each fold or split, a subset of the classes are reserved as novel, serving as the validation set. Results presented in the main manuscript represent an average across all four splits. For a detailed breakdown, Table B provides the performance of our model on each of the four splits individually for both datasets, in both inductive and transductive settings. As expected, the performance on novel classes improves substantially for 5-shot cases across all the splits. Additionally, transduction generally enhances both base and novel class accuracy, except for split-3 of COCO-20^i, where the novel class performance experiences a minor degradation, although the base class performance increases.

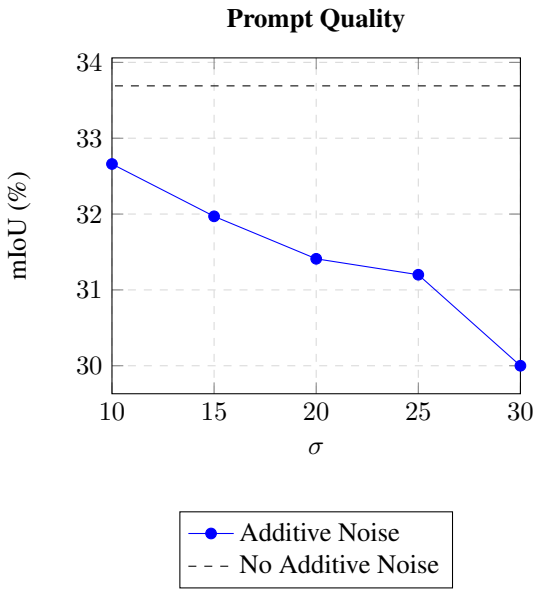

Figure E. 1-shot GFSS performance on split 0 for different additive noise variations (standard deviation σ of Gaussian noise) on Pascal-5^i dataset. The additive noise is applied on the support set images which affect the quality of the novel prompts after performing masked average pooling.

Class-wise Results. In Table C, we present the base mIoU and novel mIoU, calculated as the mean across all base and novel classes, respectively. The overall mean mIoU of the Generalized Few-Shot Segmentation (GFSS) is determined by averaging these values, in accordance with the evaluation protocol outlined in [1]. Table C provides insights into the 1-shot performance of our model within an inductive setting for each class (both base and novel) on the PASCAL-5^i dataset, covering all four splits.

To emphasize the significance of our model's novel-to-base causal attention, we conduct a comparative analysis against the baseline without causal attention. The results, as depicted in the table, underscore a notable enhancement in performance for both novel and base classes across the various splits. Notably, on split-1, our base class performance exhibits a substantial improvement across most base classes.

D. Multi-scale Analysis

One of the key components of our approach involves prompting at multiple scales of the image. Our hypothesis posits that this strategy contributes to improved segmentation accuracy across objects of varying sizes. To substantiate this claim, Table D presents an analysis of our model's performance in a transductive setting, evaluating objects of different sizes across various splits of Pascal-5^i, and comparing it with the transductive DIaM [1]. We categorize

	Split 0			Split 1			Split 2			Split 3		
	Class Name	With Causal Attn	No Causal Attn	Class Name	With Causal Attn	No Causal Attn	Class Name	With Causal Attn	No Causal Attn	Class Name	With Causal Attn	No Causal Attn
Base Classes	**Bus**	93.85	93.80	**Airplane**	88.05	82.85	**Airplane**	87.48	87.70	**Airplane**	87.81	88.26
	Car	84.27	84.10	**Bicycle**	45.10	35.21	**Bicycle**	36.84	37.93	**Bicycle**	43.33	43.34
	Cat	92.15	92.08	**Bird**	85.55	77.46	**Bird**	87.51	86.39	**Bird**	86.67	87.43
	Chair	36.67	30.13	**Boat**	70.93	71.28	**Boat**	72.23	72.08	**Boat**	75.08	75.46
	Cow	90.67	89.80	**Bottle**	81.54	59.00	**Bottle**	81.02	68.70	**Bottle**	80.06	81.31
	Dining Table	55.20	46.43	**Dining Table**	57.99	33.90	**Bus**	93.56	93.68	**Bus**	92.95	89.85
	Dog	88.74	89.03	**Dog**	89.15	70.01	**Car**	85.96	85.56	**Car**	87.18	82.98
	Horse	84.95	83.81	**Horse**	72.94	26.94	**Cat**	90.11	88.58	**Cat**	92.04	92.49
	Motorcycle	79.45	82.50	**Motorcycle**	82.42	29.27	**Chair**	30.09	30.22	**Chair**	34.20	18.14
	Person	86.56	86.61	**Person**	85.81	71.86	**Cow**	71.70	74.77	**Cow**	88.82	88.33
	Potted Plant	55.47	38.44	**Potted Plant**	56.11	33.14	**Potted Plant**	53.67	49.07	**Dining Table**	57.93	58.00
	Sheep	86.30	85.63	**Sheep**	86.19	77.59	**Sheep**	84.16	64.65	**Dog**	88.57	89.62
	Sofa	43.86	42.64	**Sofa**	42.85	43.99	**Sofa**	45.54	47.11	**Horse**	81.44	85.95
	Train	86.93	86.63	**Train**	75.77	57.45	**Train**	87.16	87.32	**Motorcycle**	84.95	82.70
	TV	75.13	71.26	**TV**	73.05	72.63	**TV**	75.63	75.40	**Person**	84.83	86.49
Novel Classes	**Airplane**	36.96	17.40	**Bus**	51.52	35.29	**Dining Table**	13.08	13.97	**Potted Plant**	23.69	16.37
	Bicycle	26.95	18.02	**Car**	28.98	24.65	**Dog**	35.71	33.19	**Sheep**	56.30	20.32
	Bird	41.56	13.33	**Cat**	74.20	43.03	**Horse**	36.77	31.11	**Sofa**	12.91	13.33
	Boat	10.15	7.25	**Chair**	9.54	8.98	**Motorcycle**	48.42	49.72	**Train**	38.73	33.31
	Bottle	33.67	17.68	**Cow**	63.55	38.30	**Person**	36.40	31.76	**TV**	20.50	16.03

Table C. 1-shot GFSS performance for each class in each split of PASCAL-5^i dataset in inductive setting with and without causal attention.

	Split 0		Split 1		Split 2		Split 3		Mean	
Object Size	Ours	DIaM [1]	Ours	DIaM [1]	Ours	DIaM [1]	Ours	DIaM [1]	Ours	DIaM [1]
Small	49.98	40.58	54.42	45.42	44.75	33.50	52.75	46.57	50.48	41.52
Medium	77.47	71.29	76.63	69.00	69.66	54.07	76.39	72.01	75.03	66.59
Large	86.78	82.53	81.68	74.91	71.68	54.98	84.78	80.34	81.23	73.19

Table D. Performance analysis of our model in a transductive setting on 1-shot GFSS for **different object sizes** on PASCAL-5^i compared against DIaM (transductive) [1]. The object sizes are grouped based on the proportion of the image they occupy. Objects occupying more than 30% of the image are categorized as *large* objects; the objects occupying 10-30% of the image are classified as *medium*; and rest as *small*.

Cross Dataset	Base	Novel	Mean
Ours (w/o transd.+ causal att.)	63.8	23.9	43.9
Ours (w/o transd.)	63.3	29.9	46.6
DiaM [1]	72.3	27.6	50.2
Ours	**72.7**	**32.1**	**52.2**

Table E. Cross dataset evaluating mIoU (COCO2PASCAL).

objects into three size categories: *small*, *medium*, and *large*, based on the proportion of the image each object occupies. Specifically, we classify objects that occupy more than 30% of the image as *large*, those occupying between 10-30% as *medium*, and the remainder as *small*. As depicted in the table, our approach consistently outperforms transductive DIaM [1] across objects of all sizes, notably excelling on small-sized objects.

E. Cross-Dataset Evaluation

To evaluate the performance of our proposed approach across different domains, we conduct an experiment where the base training is done on split-0 of COCO-20^i and the few-shot inference is performed on Pascal-5^i on classes not overlapping with the base classes of split-0. This non-overlapping categories are following six classes: airplane, boat, chair, dining table, dog and person. Table E shows the performance of our approach on this cross-domain experiment compared to transductive DIaM [1], our model in inductive setting, and our model without novel-to-base causal attention. As observed in the table, we substantially outperform DIaM [1] in this cross-dataset experiment, particularly obtaining superior performance on novel categories. Even our inductive setting obtains better novel mIoU than DIaM [1]. Additionally, using our proposed novel-to-base causal attention mechanism leads to a better disentanglement of novel prompts to base prompts, leading to significant improvement in novel mIoU.

F. Additional Qualitative Results

In Figure 5 of the main manuscript, we presented the qualitative results for 1-shot Generalized Few-Shot Segmentation (GFSS) using our model on PASCAL-5^i in both inductive and transductive settings. We compared our outcomes against the baseline lacking causal attention and transductive DIAM [1]. However, no results were shown for COCO-20^i. Therefore, in Figure F, a similar qualitative analysis is presented for 1-shot GFSS on COCO-20^i.

As observed, in the first image, the baseline without novel-to-base causal attention fails to appropriately segment the novel class *bus*. The model with causal attention can segment the *bus* with higher accuracy, further improved by transduction. In comparison, the segmentation quality of *bus* in transductive DIAM [1] is considerably worse. In the

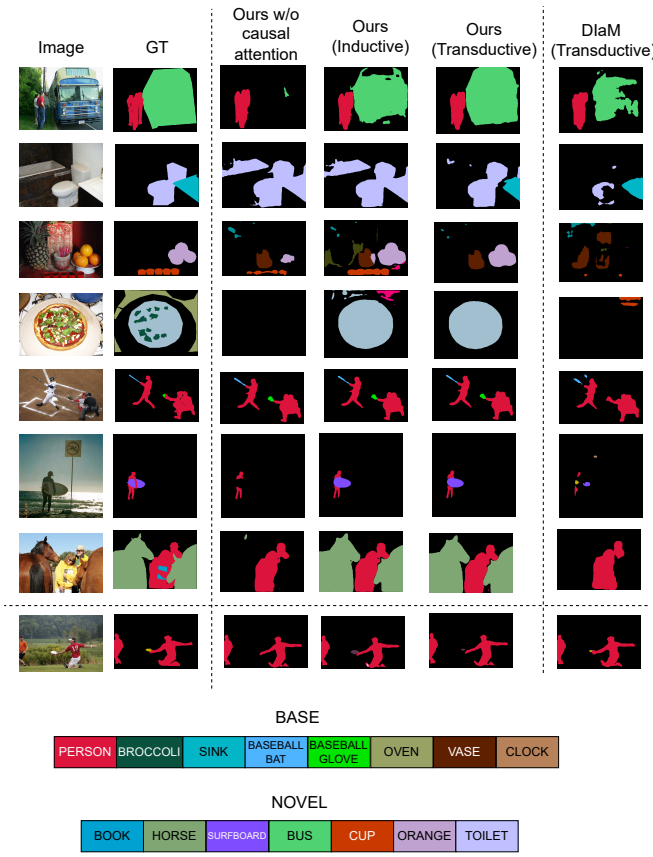

Figure F. **Qualitative Results** for 1-shot GFSS on COCO-20^i. The leftomost two columns show image and ground truth mask; (Third) Baseline without causal attention; (Fourth); Ours in inductive setting; (Fifth) Ours in transductive setting; (Last) DIaM [1]. Last row illustrates a *failure*.

second image, the baseline without causal attention can detect the novel category *toilet* but misclassifies *sink* and other unlabeled pixels (the tub) as a *toilet*. Adding novel-to-base attention slightly improves segmentation quality but is still unable to correctly classify the *sink*. However, adding transduction loss helps our model correctly classify it. Although DIAM can correctly classify *sink*, it mostly misses the novel category *toilet*. Similarly, in the third image, DIAM completely misses novel classes *orange* and *cup*. Our baseline without causal attention can identify both but with poor accuracy. Our model in the inductive setting can segment both the novel classes *orange* and *cup* with a higher degree of accuracy. Our model in the transductive setting, however, performs worse as it misses *cup*. In the fourth image, both DIAM and the baseline without causal attention miss the novel class *pizza*, and base classes *broccoli* and *oven*. Our models in both inductive and transductive settings can segment *pizza* but fail to identify *broccoli* and *oven*. The fifth image showcases the importance and strength of multi-scale prompting. All our models can detect the small object *baseball gloves* (base class) which DIAM misses. Moreover, DIAM performs worse in segmenting the base class *baseball bat*. Similarly, DIAM and the baseline without causal attention miss novel classes *surfingboard* and *horse* completely in the sixth and seventh images, respectively. Our models (in inductive and transductive settings) can correctly segment them with good accuracy. However, in the seventh image, all methods fail to identify the novel class *book*. Finally, we show a case where DIAM outperforms us in correctly segmenting the novel category *frisbee*. Our models (inductive and transductive) incorrectly classify *frisbee* as *kite*.

Overall, these qualitative results demonstrate the strong performance of our model on both base and novel classes in both inductive and transductive settings. They also underscore the importance of the novel-to-base causal attention module and the multi-scale prompting approach.

G. Societal Impact

Few-shot object segmentation has multiple positive societal impacts as it can be used for a variety of useful applications, *e.g.*, robot manipulation, augmented/virtual reality and assistive technologies (*e.g.*, aid to the blind and low-level vision) [2]. It can also help in democratizing computer vision research by enabling low resourced communities to use the technology, *e.g.*, Africa to develop their own techniques with the limited labelled data available.

However, as with many AI abilities, few-shot object segmentation also can have negative societal impacts. Nonetheless, we strongly believe these misuses are available in both few-shot and non few-shot methods and are not tied to the specific few-shot case. On the contrary, we argue that empowering developing countries towards decolonizing artificial intelligence is critical towards a decentralized and ethical AI approach.

References

[1] S. Hajimiri, M. Boudiaf, I. B. Ayed, and J. Dolz. A strong baseline for generalized few-shot semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11269–11278, 2023. 1, 2, 3, 4, 5

[2] Daniela Massiceti, Luisa Zintgraf, John Bronskill, Lida Theodorou, Matthew Tobias Harris, Edward Cutrell, Cecily Morrison, Katja Hofmann, and Simone Stumpf. Orbit: A real-world few-shot dataset for teachable object recognition. In *International Conference on Computer Vision (ICCV)*, pages 10818–10828, 2021. 5