ADAPTIVELY-LABELED VISION DATASETS VIA INSTANCE-LEVEL RETRIEVAL

Brandon Trabucco[1*], Rishav Mukherji[1*], Yutong Bai[2], Ruslan Salakhutdinov[1]
[1] Carnegie Mellon University, [2] UC Berkeley, [*] Equal Contribution

ABSTRACT

Human annotations are the backbone of modern computer vision, but it is becoming clear that human data is an inefficient resource. Human annotations typically capture a *single fixed-view* of the otherwise rich visual information present in data. How can we move towards computer vision datasets that are *adaptively labeled*? We propose Instance-Level Retrieval, a method that adaptively builds datasets for object detection from large collections of unlabeled images. Given a handful of examples, our method finds and labels the most relevant training data by comparing self-supervised representations for objects. Starting from unlabeled images derived from the Pascal VOC training set, we rebuild Pascal VOC without human annotations. In experiments that control data scale, models trained on our data outperform training on the original Pascal VOC human annotations by an avearge of 0.041 mAP. Code for the method and examples are available at: instance-rag.github.io

1 INTRODUCTION

The current paradigm in modern computer vision is to gather a fixed set of human annotations based a particular downstream task like object detection. Human annotators are typically provided a set of instructions for labeling—often with a predetermined set of visual classes—and asked to locate regions of interest in provided images (Deng et al., 2009; Gupta et al., 2019; Lin et al., 2014; Everingham et al., 2010). These human annotations generally represent a *single fixed view* of the underlying rich visual information present in images. However, as computer vision applications are becoming increasingly *adaptive* in nature, needs can grow beyond a fixed set of labels. How do we build computer vision datasets that are *adaptively labeled* based on the ever-changing needs of downstream applications? We propose Instance-Level Retrieval, a method inspired by the success of Retrieval-Augmented Generation (Lewis et al., 2020) in language models, that adaptively builds a training set from a large collection of unlabeled images, given a handful of key examples.

Our method operates in three stages. First, we employ pretrained region proposal networks to generate candidate instances for retrieval. Second, we build representations for instances using features from a state-of-the-art self-supervised encoder. Finally, we search through the candidates to find instances most similar to a provided set of retrieval keys, and return the top k most relevant instances for training from the unlabeled source. We evaluate this approach by re-building the training set for Pascal VOC (Everingham et al., 2010) from a handful of key instances. Models trained on our adaptively labeled dataset demonstrate an average 0.041 mAP improvement over those trained on the original Pascal VOC human annotations; limited to the same number of instances. Code for the method and examples are available at: instance-rag.github.io

2 RELATED WORKS

Adaptive Computer Vision. Computer vision researchers have worked on adaptivity in prio works, focusing primarily on building adaptive models, rather than adaptive datasets. For example, Large Vision Models (Bai et al., 2023) are a class of autoregressive model that can perform new tasks via few-shot in-context learning. Similarly detection models like Owlv2 (Minderer et al., 2022; 2024) can have vision encoders attatched, which allow for detection of novel objects from a single example image. Perhaps the most relevant works to consider adaptive datasets are those in domain

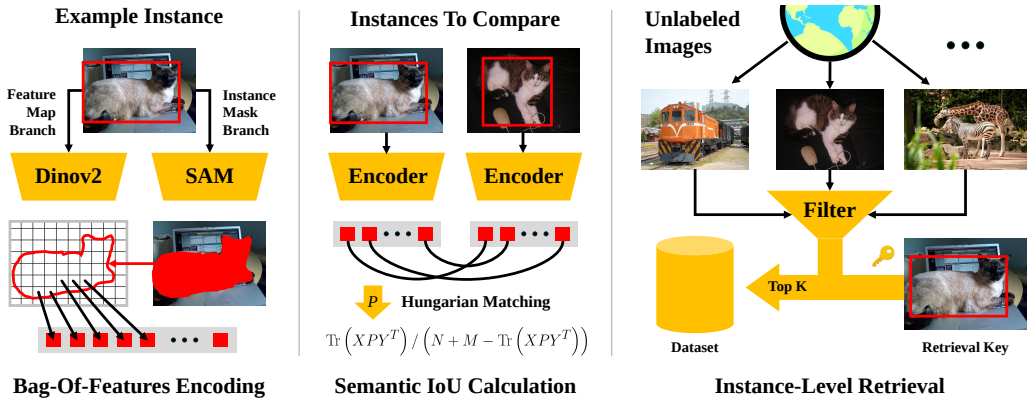

Figure 1: **Overview of Instance-Level Retrieval.** Starting from a large unlabeled set of images, we employ pretrained vision models to discover objects for training based on a handful of examples. Our method finds the most suitable instances by comparing self-supervised representations for objects using a proposed *Semantic IoU* metric sensitive to visual appearance and structural similarity.

adaptation (Csurka, 2017), which consider target domain with a potentially significant domain shift from a static source domain. The main difference between this work, and domain adaptation works lies in the formulation of the dataset. Our retrieval step dynamically rebuilds the data-distribution in its entirely based on images from the target domain, whereas previous works in domain adaptation often focus on loss function and model augmentations to mitigate domain shift (Csurka, 2017).

Retrieval Augmented Generation. Retrieval-Augmented Generation has recently emerges for language models as a way to mitigate hallucinations (Lewis et al., 2020). By allowing the model to augment its context with a relevant document retrieved from a larger source of relevant knowledge, language models can be extended to knowledge from beyond their initial training data. RAG has recently been extended to Diffusion models (Luo et al., 2024) to extend their capabilities to novel artistic concepts and styles created by users that such models were not initially trained to generate. Retrieval is especially promising in data-centric adaptivity, and to the best of our knowledge, ours is the first work to consider Instance-Level Retrieval for building adaptive object detection datasets.

Synthetic Data. Synthetic data in computer vision has recently emerged as an effective strategy for training data-efficient models with limited real data (Trabucco et al., 2023; He et al., 2023; Wu et al., 2023; Azizi et al., 2023). For object detection, synthetic data methods have explored generating synthetic images alongside their labels (Wu et al., 2023), and using pretrained object detectors to weakly annotate unlabeled images for training larger models (Minderer et al., 2024). However, these approaches ultimately do not solve the adaptivity problem. When the required visual task changes—often the case in real-world settings—synthetic data must often be re-generated.

3 Methodology

3.1 Bounding Box Proposal

For a given unlabeled image, the initial step involves generating a set of bounding box proposals. We achieve this using the Owlv2ForObjectDetection model (Minderer et al., 2024), which produces candidate bounding boxes that are subsequently filtered based on their objectness score. Each resulting bounding box, along with its corresponding image, is referred to as an instance.

To refine the set of proposed instances, we apply Non-Maximum Suppression (NMS) to eliminate bounding boxes with excessive overlap. We compute the pairwise Intersection over Union (IoU) for all bounding boxes within a given image. If the IoU between any two boxes exceeds a predefined threshold, we discard the instance with the lower objectness score. This process ensures that only distinct and high-confidence object proposals are retained.

3.2 BAG-OF-FEATURES ENCODING

Once the set of instances is obtained, we process each instance using the Segment Anything Model 2 (SAM2) (Kirillov et al., 2023) to generate a segmentation mask over the principal object. Simultaneously, we pass the image through the DINOv2 model (Oquab et al., 2024) to extract its feature map. We then retrieve the features from this feature map that correspond to the computed mask for each instance, ensuring that only the relevant regions are utilized for further processing. For each candidate instance, this process results in a set of patch features that were contained within the bounds of the segmentation mask. This set is defined below.

$$X = \{\vec{\mathbf{x}}_1, \vec{\mathbf{x}}_2, \cdots, \vec{\mathbf{x}}_N\} \text{ st } \vec{\mathbf{x}}_i \in \mathcal{R}^D \tag{1}$$

Here, each $\vec{\mathbf{x}}_i$ corresponds to a single patch feature selected from the final layer predictions of Dinov2 (Oquab et al., 2024). While it is the most accurate to think of X as a set, in that elements are not inherently ordered, we overload this notation and treat X as a matrix with N rows and D columns, where N is the number of patch locations within the segmentation mask for the object, and D is the dimensionality of the Dinov2 feature space.

3.3 SEMANTIC IOU CALCUALTION

Equipped with a self-supervised representation for objects based on a set of patch-level features, we develop a similarity metric to compare two bag-of-features representations. Given two such representations X and Y defined in the previous section, we compute their *Semantic Intersection Over Union* (*Semantic IoU*) as follows. We first normalize all vectors within each set to unit length. Then, we apply the hungarian matching algorithm (Kuhn, 1955) to pair vectors from the set X to the set Y in order to maximize cosine similarity of paired vectors.

$$\hat{P} = \arg \max_{P \in S_{N,M}} \text{Tr}\left(X(PY)^T\right) \tag{2}$$

The matrix P^* found by the hungarian matching algorithm is a non-square permutation-like matrix from the set of non-square permutation-like matrices from N elements to M elements, noted $S_{N,M}$ in Equation 2. The matrix is optimized to maximize the sum of cosine similarities of paired vectors, induced by P^*, accomplished above using the matrix interpretations for X and Y. Here we assume that X has N elements as defined previously, and Y is defined similarly to X, but with M elements. With the optimal match between the sets, we compute *Semantic IoU* with the following.

$$\text{Semantic IoU} = \underbrace{\text{Tr}\left(X(\hat{P}Y)^T\right)}_{\textbf{intersection}} \Big/ \underbrace{\left(N + M - \text{Tr}\left(X(\hat{P}Y)^T\right)\right)}_{\textbf{union}} \tag{3}$$

Intuitively, this metric favors objects that are visually similar (have many patch features with high cosine similarity), and are of comparable size (the number of patches).

3.4 INSTANCE-LEVEL RETRIEVAL

The final step in generating adaptively labeled data involves a matching algorithm inspired by Retrieval-Augmented Generation (RAG) (Lewis et al., 2020). We begin by selecting a retrieval key, which is a human-labeled instance—referred to as the anchor instance—for which we aim to generate similar instances from unlabeled data.

After generating bounding box proposals for the unlabeled data, we extract bag-of-features encodings for both the generated instances and the anchor instance. We then compute the *Semantic IoU* between the anchor instance and all proposed instances.

Instead of directly selecting the top K instances, we initially retain the top $10 * K$ instances based on their similarity scores. We filter this superset by applying NMS as described in sub-section 3.1, but with a more stringent IoU threshold to ensure greater distinctiveness among selected instances. Finally, from this filtered subset, we select the top K instances and assign them the same label as the anchor instance.

4 RESULTS

4.1 LABEL ASSIGNMENT ACCURACY

We first evaluate the accuracy of the RAG-inspired label assignment methodology in sub-section 3.4. Instead of processing the entire dataset through the complete pipeline to generate bounding boxes, we utilize the pre-existing bounding boxes provided in the Pascal VOC dataset.

We iterate through the validation set, treating each instance as an anchor. For each anchor, we compute the *Semantic IoU* against all instances in the training set and identify the top K most similar instances. The anchor is then assigned the label corresponding to the mode of these K nearest instances.

To assess the label assignment performance we define two metrics. *Accuracy* is defined as the proportion of validation instances assigned a label that matches their ground truth. *Consistency* measures how many of the K selected instances share the same ground-truth label as the anchor instance.

K	**Accuracy**	**Consistency**
1	0.941	–
5	0.951	0.927
10	0.953	0.915
15	0.953	0.906
20	0.951	0.898

Table 1: Accuracy and Consistency for Different Values of K

4.2 OBJECT DETECTION PERFORMANCE

To demonstrate the training efficiency gained through the proposed adaptive labeling algorithm, we compare the performance of an object detection model trained on the original Pascal VOC human annotations with a model trained on an adaptively labeled dataset derived from the same source.

To construct the datasets, we follow the standard preprocessing pipeline. We generate synthetic instances for all the images in the training set. For each class, we randomly sample N anchor instances from the ground-truth annotations in the training set. We then carry out pre-selection, filtering and final selection of K instances for each anchor based on the algorithm described in sub-section 3.4.

To ensure a fair comparison, we subsample $N * K$ instances per class from the original training set, matching the dataset size of the adaptively labeled version.

We then train a YOLO11m model from scratch for 300 epochs on both datasets with image size set to 640 pixels. In table 2, we present the highest Mean Average Precision (mAP) achieved by the model on the Pascal VOC validation set. It is computed as the mean of the Average Precision (AP) across all object categories. The reported mAP@50 and mAP@50-95 is averaged over 5 seeds and correspond to the AP averaged over IoU thresholds of 0.50 and the range from 0.50 to 0.95, respectively.

Instances	N	K	**mAP@50**	**mAP@50-95**
1000	5	10	0.164 ± 0.017	0.098 ± 0.011
	Ground		0.116 ± 0.008	0.061 ± 0.006
4000	20	10	0.369 ± 0.011	0.246 ± 0.008
	Ground		0.328 ± 0.027	0.207 ± 0.021

Table 2: Object Detection Performance on Pascal VOC Validation Set

REFERENCES

Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J. Fleet. Synthetic data from diffusion models improves imagenet classification, 2023. URL https://arxiv.org/abs/2304.08466.

Yutong Bai, Xinyang Geng, Karttikeya Mangalam, Amir Bar, Alan Yuille, Trevor Darrell, Jitendra Malik, and Alexei A Efros. Sequential modeling enables scalable learning for large vision models, 2023. URL https://arxiv.org/abs/2312.00785.

Gabriela Csurka. Domain adaptation for visual applications: A comprehensive survey. *CoRR*, abs/1702.05374, 2017. URL http://arxiv.org/abs/1702.05374.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010.

Agrim Gupta, Piotr Dollár, and Ross B. Girshick. LVIS: A dataset for large vocabulary instance segmentation. *CoRR*, abs/1908.03195, 2019. URL http://arxiv.org/abs/1908.03195.

Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition?, 2023. URL https://arxiv.org/abs/2210.07574.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. URL https://arxiv.org/abs/2304.02643.

Harold W. Kuhn. The Hungarian Method for the Assignment Problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, March 1955. doi: 10.1002/nav.3800020109.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. *CoRR*, abs/2005.11401, 2020. URL https://arxiv.org/abs/2005.11401.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pp. 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48. URL https://doi.org/10.1007/978-3-319-10602-1_48.

Michael Luo, Justin Wong, Brandon Trabucco, Yanping Huang, Joseph E. Gonzalez, Zhifeng Chen, Ruslan Salakhutdinov, and Ion Stoica. Stylus: Automatic adapter selection for diffusion models, 2024.

Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection with vision transformers, 2022. URL https://arxiv.org/abs/2205.06230.

Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection, 2024. URL https://arxiv.org/abs/2306.09683.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

Brandon Trabucco, Kyle Doherty, Max Gurinas, and Ruslan Salakhutdinov. Effective data augmentation with diffusion models, 2023. URL https://arxiv.org/abs/2302.07944.

Weijia Wu, Yuzhong Zhao, Hao Chen, Yuchao Gu, Rui Zhao, Yefei He, Hong Zhou, Mike Zheng Shou, and Chunhua Shen. Datasetdm: Synthesizing data with perception annotations using diffusion models, 2023. URL https://arxiv.org/abs/2308.06160.

A CONCLUSION

The improved performance of the object detection model trained on fully synthetic labels demonstrates significant potential for scaling the training of Large Visual Models. This approach enables the creation of substantially larger labeled datasets while ensuring stable model training. With only a small set of human-labeled instances, we can generate a significantly larger collection of similar image-label pairs.

Future work will explore not only replicating performance within the same dataset but also leveraging large-scale unlabeled datasets to generate task-specific instances, further expanding the adaptability and generalization of Large Visual Models.