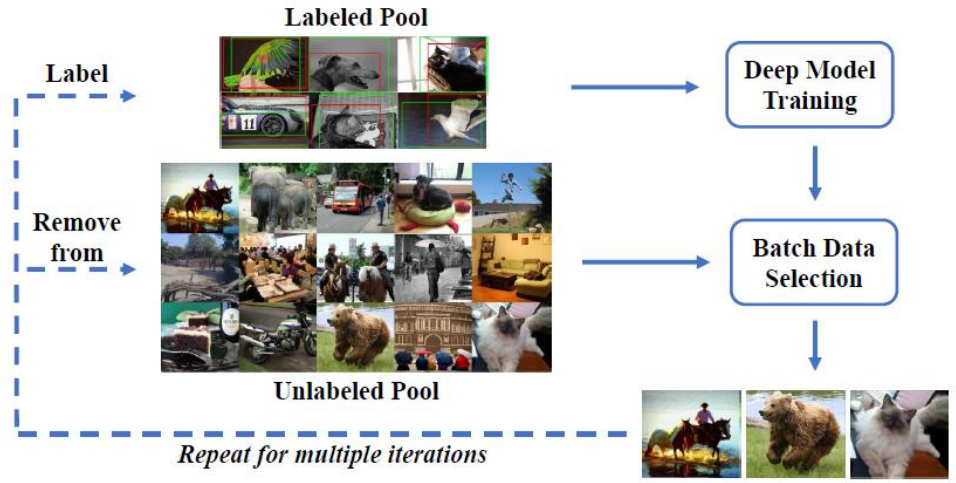
南京航空航天大學

Nanjing University of Aeronautics and Astronautics

Towards General and Efficient Active Learning

Yichen Xie, Masayoshi Tomizuka, Wei Zhan*
University of California, Berkeley

Introduction

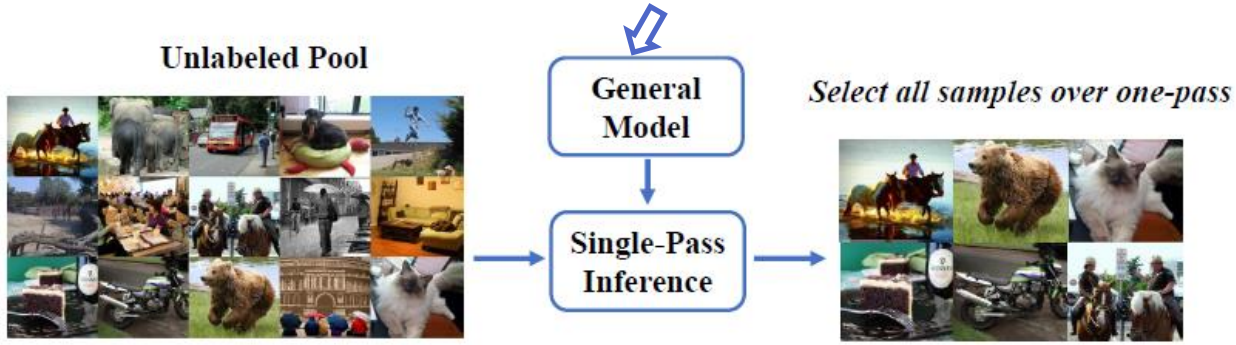


(a) **Traditional Pipeline of Active Learning:** Model training and batch data selection repeat multiple times on each dataset.

Three principles of our design：

- **Generality**: a general pre-trained model can work over multiple datasets
- **Efficiency**: query only once
- **Non-supervision**

self-supervised vision transformers



(b) **Our Pipeline:** Samples are selected with *one-pass* model inference on each dataset.

GEAL

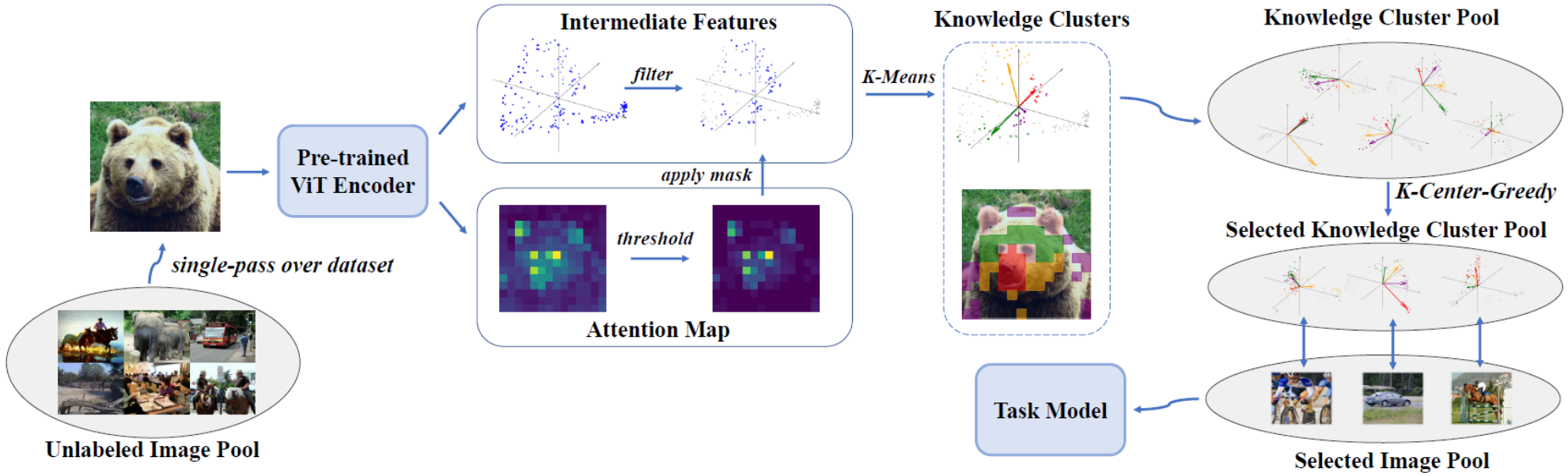


Figure 2. **Overview of our proposed GEAL:** Our method relies on a general pre-trained vision transformer to extract features from images. Knowledge clusters are derived from the intermediate features. Afterwards, we perform *K-Center-Greedy* algorithm to select knowledge clusters as well as the associated images. These selected images are labeled for downstream task model training.

Off-the-Shelf Features for Active Learning

Perform Core-Set algorithm over off-the-shelf global features extracted by ViT-Small model pre-trained on ImageNet dataset.

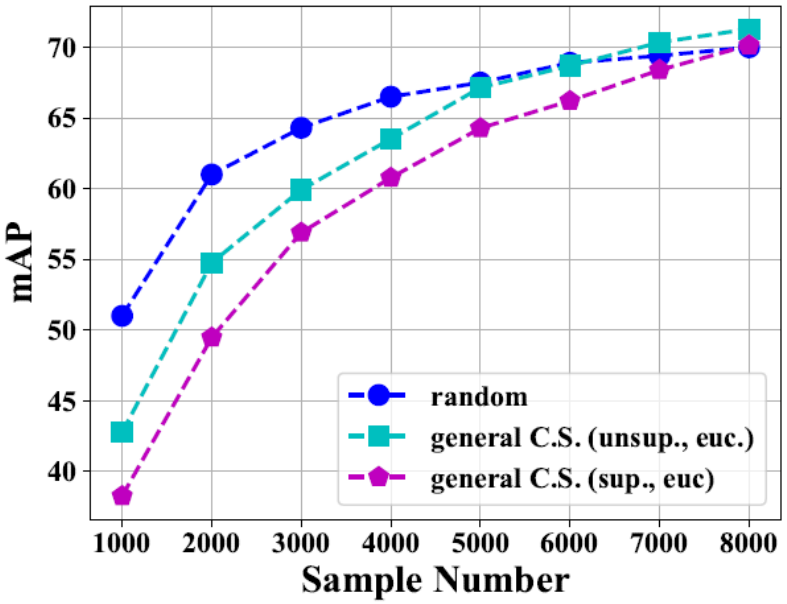


Figure 3. **Preliminary: Core-Set over *off-the-shelf* Global Features:** *C.S.* denotes Core-Set in this figure. We directly perform Core-Set algorithm over off-the-shelf global features extracted by ViT-Small model pre-trained on ImageNet dataset. We follow [49] to use euclidean distance (*euc.*). This setting causes a significant performance drop, especially when sample number is small.

Off-the-Shelf Features don't work

Two potential reasons for this failure:

- Complex scenes are hard to represent globally
- K-Center-Greedy tends to select simple scenes

Extract reliable local information?

Knowledge Clusters inside Images

Two challenges in extracting reliable local information:

1. Given the low information density inside images, many regions are useless or even distracting for downstream tasks.
2. Local features extracted by DNNs inevitably contain noise.

Solution:

1. For the first concern, the class token self-attention map of the transformer can serve as a natural indicator of regional importance even without dense supervision.
2. For the second concern, to eliminate the noise, we perform K-Means clustering over the distilled intermediate features

$$
\begin{aligned}
{}^{I}c_j &= \frac{1}{|C_j|}\sum_{r\in C_j} f_{n-1}^{r} \\
&= \frac{1}{|C_j|}\sum_{r\in C_j} \hat{f}_{n-1}^{r} + \frac{1}{|C_j|}\sum_{r\in C_j} \epsilon_{n-1}^{r} \\
&\approx \frac{1}{|C_j|}\sum_{r\in C_j} \hat{f}_{n-1}^{r}
\end{aligned}
$$

$I_{C_j}, j = 1, 2, \ldots, K$ are defined as knowledge clusters inside the image I

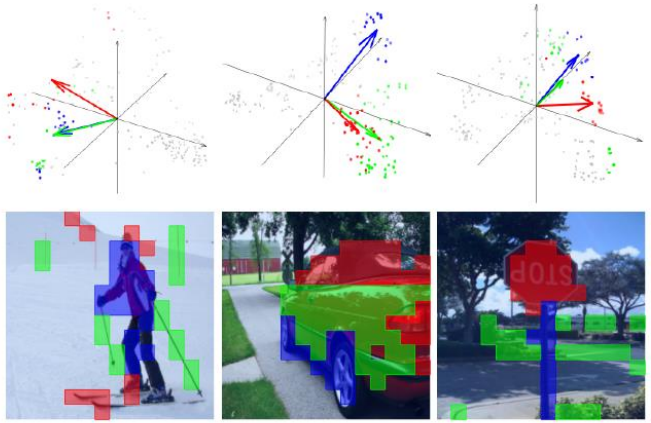


Figure 4. **Visualization of Knowledge Clusters:** *Top:* The distilled intermediate features of each image are grouped into different knowledge clusters. Gray features are eliminated in Eq. 2. Dimensions are reduced through PCA for visualization. *Bottom:* Regions inside images can be associated with corresponding knowledge clusters.

Sample Selection Strategy for GEAL

所选数据集合的分布在 Knowledge cluster level 更接近整个数据集，而不是 image level。

Apply a K-Center-Greedy algorithm w.r.t. knowledge clusters

Given an unlabeled image pool I_0

- Randomly select an initial image (included K clusters)
- Then, choose the image I that contains the knowledge cluster c_j farthest from already selected clusters

$$I, j = argmax_{\substack{I \in \mathcal{I} \\ j=1,\dots,K}} min_{c \in s_{\mathcal{K}}} d({}^{I}c_j, c)$$

Algorithm 1: Data Selection Algorithm

Input: all knowledge clusters $\mathbf{c}$, total annotation budget b

Output: selected image pool $s_{\mathcal{I}}$

Initialize $s_{\mathcal{I}} = \{I_0\}$

/* initialize the selected image pool with a random image I_0 */

Initialize $s_{\mathcal{K}} = \{{}^{I_0}c_j, j = 1, \dots, K\}$

/* initialize the selected knowledge cluster pool with knowledge clusters inside I_0 */

repeat

  $I, j = \arg\max_{\substack{I \in \mathcal{I} \\ j=1,\dots,K}} \min_{c \in s_{\mathcal{K}}} d({}^{I}c_j, c)$

  /* find knowledge cluster ${}^{I}c_j$ farthest from currently selected ones */

  $s_{\mathcal{I}} = s_{\mathcal{I}} \cup \{I\}$

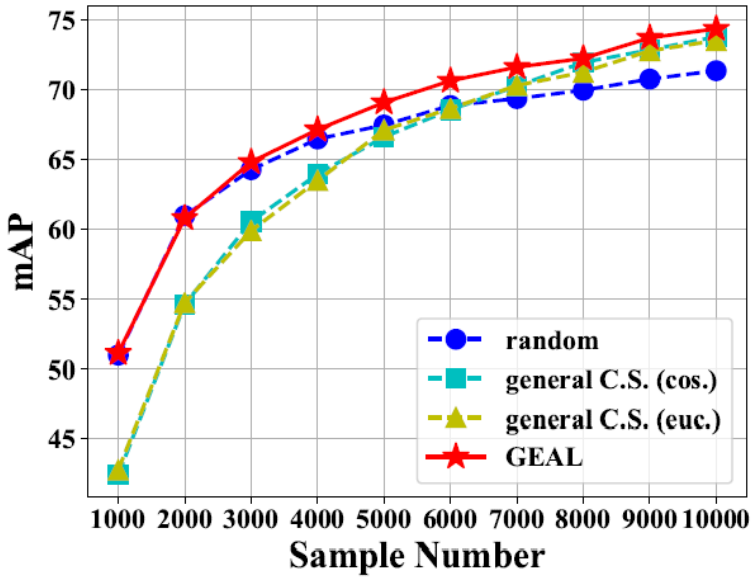
  /* add image I to selected image pool */

  $s_{\mathcal{K}} = s_{\mathcal{K}} \cup \{{}^{I}c_j, j = 1, \dots, K\}$

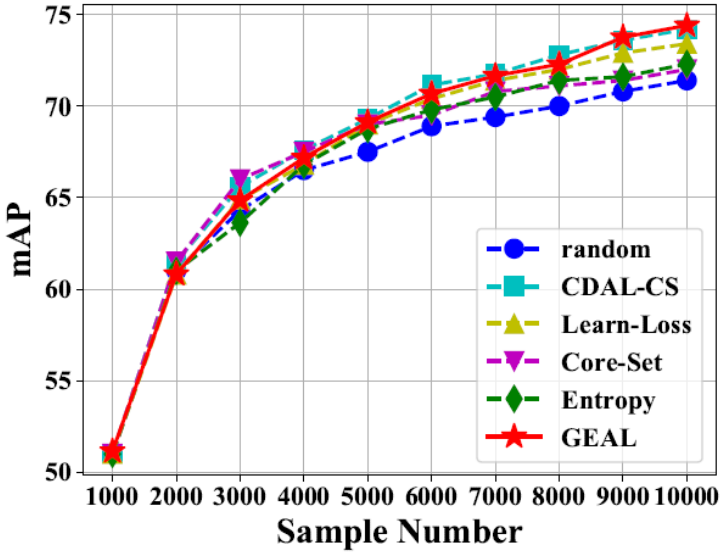
  /* add all knowledge clusters in image I to selected knowledge cluster pool */

until $|s_{\mathcal{I}}| = b$;

Experiment-Object Detection



(a) **Comparison with Random Selection and General Baselines:** *C.S.* means Core-Set, while *cos.* and *euc.* separately denote *cosine* and *euclidean* distance function. Our method has notable superiority in all cases.



(b) **Comparison with Traditional Pipeline Methods:** With significant advantage on generality and efficiency, our method is also equipped with a advanced performance competitive with or better than approaches following the traditional pipeline.

Experiments on PASCAL VOC dataset

Methods	Train	Batch	Sup.	Time
CoreSet [49]	✓	✓	✓	$\sim$ 42 *hours* + *label query*
Learn Loss [61]	✓	✓	✓	
MC-dropout [20]	✓	✓	✓	
CDAL-CS [1]	✓	✓	✓	
VAAL [52]	✓	✓	✗	$\sim$ 42 *hours*
GEAL (ours)	✗	✗	✗	648 *seconds*

Experiment – Classification & Segmentation

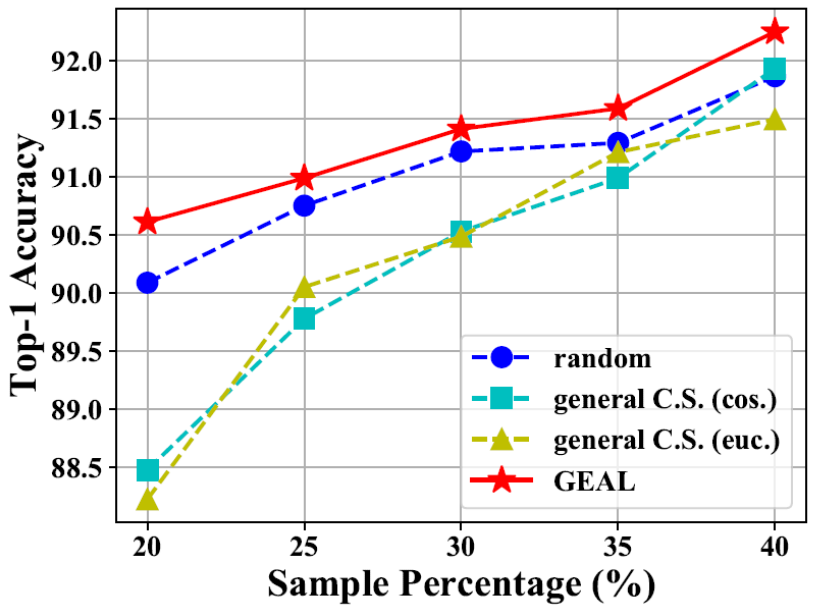


Figure 6. **Active Learning Performance on Image Classification Task:** *C.S.* means Core-Set. The Top-1 Accuracy on 100% training data is 93.13%.

Image Classification
Oxford-IIIT Pet dataset

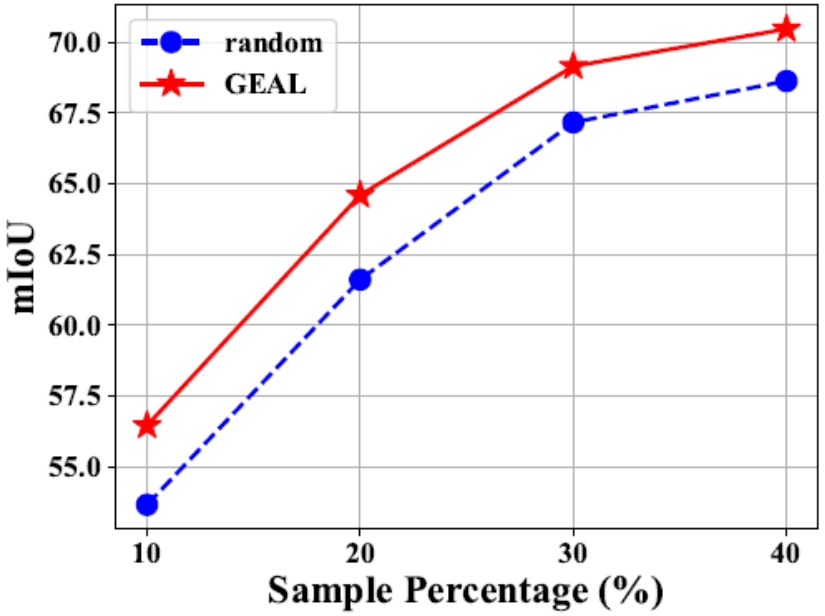


Figure 7. **Active Learning Performance on Semantic Segmentation Task:** The mIoU on 100% training data is 76.60.

Semantic Segmentation
Cityscapes dataset

Analysis

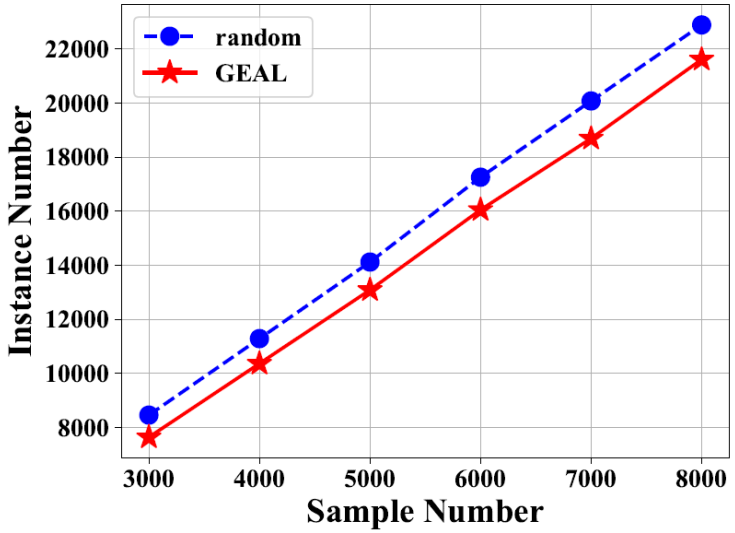


Figure 9. **Number of Instances inside Selected Images** The number is counted on PASCAL VOC dataset. GEAL selects fewer instances than random selection when the image number is same.

Instance Number

τ	K	d$(\cdot,\cdot)$	Sample Number $3k$	$5k$	$7k$
0.4	25	cos.	64.18	68.61	71.02
0.5			**64.85**	**69.12**	71.66
0.6			64.19	69.07	**71.87**
0.5	5	cos.	64.62	69.00	71.74
	15		64.36	68.95	71.68
0.5	25	euc.	63.70	68.70	71.76

Ablation Study on Object Detection

南京航空航天大學

Nanjing University of Aeronautics and Astronautics

Reducing Label Effort: Self-Supervised meets Active Learning

Javad Zolfaghari Bengar[1,2] Joost van de Weijer[1,2] Bartlomiej Twardowski[1]
Bogdan Raducanu[1,2]
Computer Vision Center (CVC)[1], Univ. Autònoma of Barcelona (UAB)[2]
{jzolfaghari,joost,btwardowski,bogdan}@cvc.uab.es

ICCV 2021

Overview

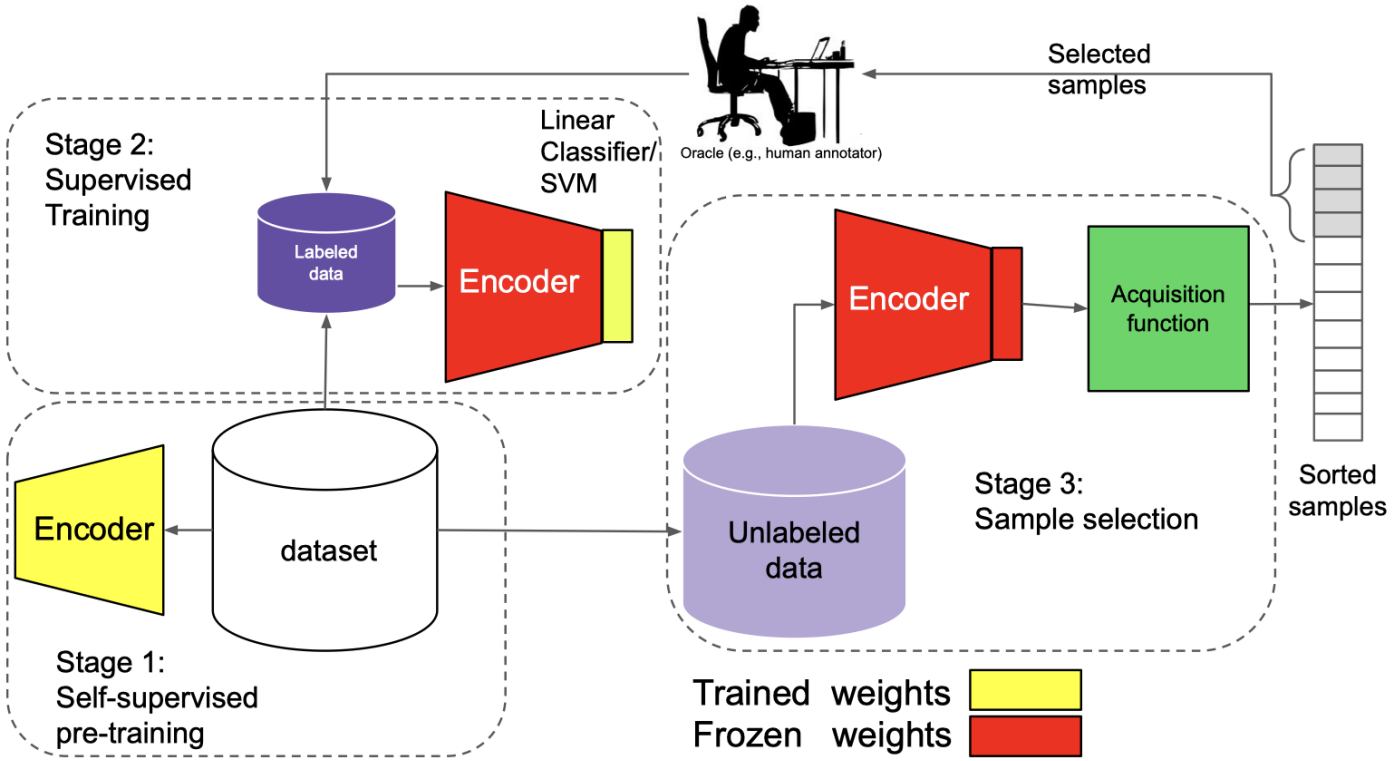


Figure 1. **Overview of active learning framework enhanced by self supervised pre-training.** The framework consists of 3 stages: (i) Self supervised model is trained on the entire dataset. (ii) Given the frozen backbone and few labeled data, a linear classifier or an SVM is fine-tuned on top of the features in supervised way. (iii) Running the model as inference on the unlabeled data and sort the samples from least to highest informative/representative via acquisition function. Finally the top samples are queried to oracle for labeling and added to labeled set. Stages (i) & (ii) are repeated until the total labeling budget finishes.

Experiment

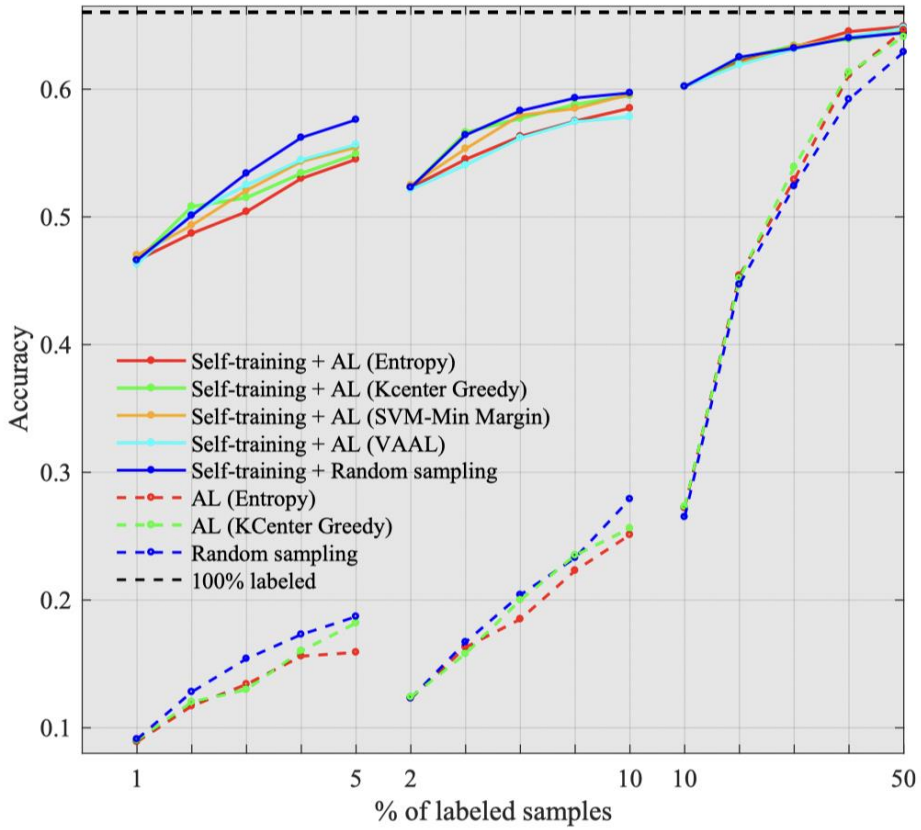


Figure 4. **AL performance on cifar100** performance comparison between the addition of self-training to AL methods (solid lines) and AL methods (dashed lines). The initial and per cycle budget are equal in all the curves.

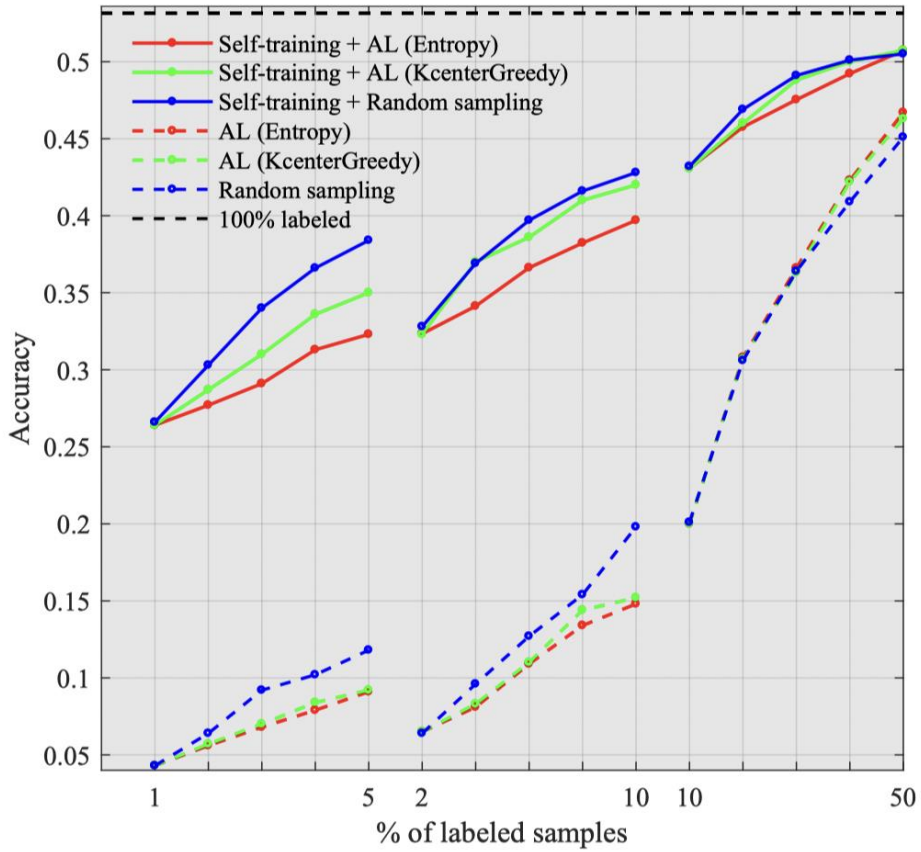


Figure 5. **AL performance on Tiny ImageNet** performance comparison between the addition of self-training to AL methods (solid lines) and AL methods (dashed lines). The initial and per cycle budget are equal in all the curves.

THANKS