Towards Controlled Data Augmentations for Active Learning

Jianan Yang [1] **Haobo Wang** [1] **Sai Wu** [1] **Gang Chen** [1] **Junbo Zhao** [1]

Abstract

The mission of active learning is to identify the most *valuable* data samples, thus attaining decent performance with much fewer samples. The data augmentation techniques seem straightforward yet promising to enhance active learning by extending the exploration of the input space, which helps locate more valuable samples. In this work, we thoroughly study the coupling of data augmentation and active learning, thereby proposing **C**ontrollable **A**ugmentation **M**ani**P**ulator for **A**ctive **L**earning. In contrast to the few prior works that touched on this line, CAMPAL emphasizes a purposeful, tighten, and better-controlled integration of data augmentation into active learning in three folds: (i)-carefully designed augmentation policies applied separately on labeled and unlabeled data pools; (ii)-controlled and quantifiably optimizable augmentation strengths; (iii)-full and flexible coverage for most (if not all) active learning schemes. Theories are proposed and associated with the development of key components in CAMPAL. Through extensive empirical experiments, we bring the performance of active learning methods to a new level: an **absolute** performance boost of **16.99%** on CIFAR-10 and **12.25%** on SVHN with 1,000 annotated samples. Codes are available at https://github.com/jnzju/CAMPAL.

1. Introduction

The acquisition of labeled data serves as a foundation for the remarkable successes of deep supervised learning over the last decade, which also incurs great monetary and time costs. *Active learning* (AL) is a pivotal learning paradigm that puts the data acquisition process into the loop of learning (Settles, 2009; Zhang et al., 2020; Kim et al., 2021a; Wu et al., 2021). By decomposing the learning loop into several cycles that alternatively accumulate valuable data samples and update the model, AL attains much-lowered sample complexity but comparable performance compared to its supervised counterpart. Owing to this efficacy, active learning is widely used in real-world applications and ML productions (Bhattacharjee et al., 2017; Feng et al., 2019; Hussein et al., 2016). In spite of its meritorious practicality, active learning often suffers from unreliable data acquisition, especially from the early stages. Notably, the models obtained around the early stages are generally raw and undeveloped due to the insufficient data curated and sparse supervision signal being consumed.

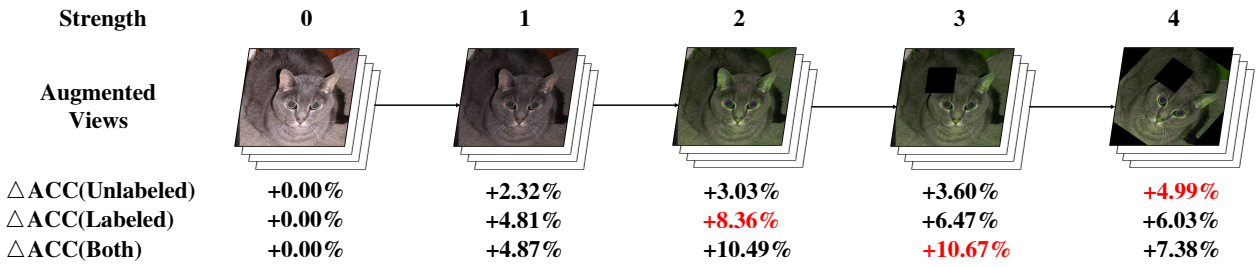
While this problem can probably be mitigated after adequate cycles are conducted, we argue that the problems at the early stages of AL cannot be overlooked. Indeed, few works have resorted to data augmentation techniques to generate additional data examples for active learning, e.g. GAN-based (Tran et al., 2019) and STN-based (Kim et al., 2021b) methods. In this work, we take a further step in investigating the role of data augmentation for AL.

To begin with, we provide a straightforward quantitative observation in Figure 1. The setup of these results is rather simple: we directly apply vanilla data augmentation (DA) operations, such as flipping and rotation, to data samples and stack them to increase the augmentation strengths. We may conclude from these scores as follows. First, the augmentations (loosely) integrated into AL have led to surprisingly enhanced results, albeit their simple designs. Secondly and perhaps more important, the same augmentation policy facilitated on different data pools manifests notably different impacts. As shown in Figure 1, the labeled and unlabeled data pools achieve the best performance at different levels of augmentation strengths. To incorporate DA into AL schemes, the augmentation ought to serve different objectives on the labeled/unlabeled pools. In particular, the labeled pool favors label-preserving augmentation in order to obtain a reliable classifier, when the unlabeled pool may require relatively more aggressive augmentations to maximally gauge the unexplored distribution. Noted, this observation has not been investigated by prior works (Wei et al., 2020; Gao et al., 2020; Kim et al., 2021b).

Motivated by it, we propose **C**ontrollable **A**ugmentation

[1]Department of Computer Science, Zhejiang University, Hangzhou, China. Correspondence to: Junbo Zhao <j.zhao@zju.edu.cn>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.



	0	1	2	3	4
△ACC(Unlabeled)	+0.00%	+2.32%	+3.03%	+3.60%	+4.99%
△ACC(Labeled)	+0.00%	+4.81%	+8.36%	+6.47%	+6.03%
△ACC(Both)	+0.00%	+4.87%	+10.49%	+10.67%	+7.38%

Figure 1. A visualization for data augmentation and their corresponding performance change as we stack augmentations over images when integrating them into active learning cycles, where the strength denotes the number of vanilla augmentations we stack. We test 3 cases where augmentations are applied to 1) unlabeled samples only; 2) labeled samples only; 3) Both. More details are in Appendix B.3.

Mani**P**ulator for **A**ctive **L**earning. Core to our method is a purposely designed form of better controlled and tightened integration of data augmentation into active learning. By proposing CAMPAL, we aim to fill this integration gap and unlock the full potential of data augmentation methods in active learning schemes. In particular, CAMPAL integrates several mechanisms into the whole AL framework:

- CAMPAL constructs separate augmentation flows distinctly on labeled and unlabeled data pools towards their own objectives;
- CAMPAL composes a *strength* optimization procedure for the applied augmentation policies;
- CAMPAL complies with the most common AL schemes, with carefully designed acquisition functions for both score- and representation-based methods;

Besides the theoretical justification of CAMPAL offered in Section 4, we extensively conduct wide experiments and analyses on our approach. The empirical results of CAMPAL are stunning: a **16.99%** absolute improvement at a 1,000-sample cycle and a **13.34%** lead with 2,000 samples on CIFAR-10, compared with previously best methods. As an extra juice, we demonstrate CAMPAL's versatility by embedding it into a parallel semi-supervised learning workflow, which further boosts the performance from the lens of both paradigms. Arguably, we postulate that these significantly-enhanced results may well extend the boundary of AL.

2. Methodology

In this section, we describe CAMPAL in detail. CAMPAL is chiefly composed of two stages. First, CAMPAL controls the augmentations being applied to labeled/unlabeled data pools with distinct optimization objectives (Section 2.2). Second, CAMPAL provides an enhanced data acquisition system based on the properly-controlled augmentations (Section 2.3). To this end, we may posit that CAMPAL forms a much more tightened integration of DA and AL, due to not only its controllable mechanism on both data pools but also its full adaptability for all common active learning schemes. With properly controlled augmentations and enhanced acquisitions, CAMPAL completes the integration of augmentations and active learning. The framework for CAMPAL is summarized in Figure 2.

2.1. Setup and Definitions

Active learning. The problem of active learning (AL) is defined with the following setup. Consider $\mathcal{D} \subset \mathbb{R}^d$ as the underlying dataset consisting of a labeled data pool $\mathcal{D}_L$ and an unlabeled data pool $\mathcal{D}_U$, with $|\mathcal{D}_U| \gg |\mathcal{D}_L|$. Based on a fully-trained classifier f_θ that assigns a label to each data point, a data acquisition function $h_{acq}(x, f_\theta) : \mathcal{D}_U \to \mathbb{R}$ calculates the score for each data instance. We also use $\mathcal{P}(y|x; f_\theta)$ to denote the probabilistic label distribution of x given by f_θ. Then AL selects the most valuable sample batch and updates the labeled set accordingly. In the remainder of this paper, we omit parameter f_θ in h_{acq} when the reliance on acquisitions over classifiers is clear.

Data augmentation. We denote the set consisting of vanilla augmentations (i.e. translation, rotation as shown in Table 8) by $\mathcal{T}$. Instead of designing new types of augmentations, we optimize the strength of augmentations derived from $\mathcal{T}$. In practice, several studies provide extended augmentation operators consisting of multiple vanilla operators (Hendrycks et al., 2019; Xu et al., 2022). Denoting an augmentation with T, the number of operators s in the composition is named the *strength* of T, then $\mathcal{T}^{(s)}$ denotes the augmentation set with strength s. Intuitively, s also quantifies how far an augmentation drifts images away from their original counterparts. Given a data point x, we also use $\mathcal{T}^{(s)}(x)$ to denote all its augmented views with strength s.

2.2. Controllable Augmentations for Active Learning

As shown in Figure 1, data augmentations serve different goals on different data pools for locating valuable samples with better acquisitions. On labeled data, it targets to improve the model prediction performance, thus providing a

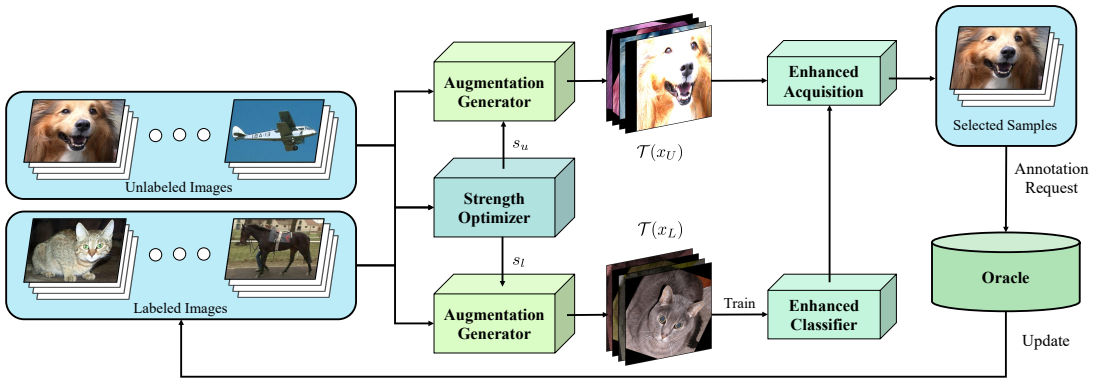


Figure 2. An active learning cycle for the CAMPAL framework. We optimize strengths s_u, s_l for the unlabeled/labeled pool separately, then generate the augmented images with the strengths given. We train an enhanced classifier over augmented labeled samples, then deduce an enhanced acquisition with the augmented unlabeled batch and the enhanced classifier. Valuable samples are selected accordingly.

reliable base model for acquisition. On unlabeled examples, it enlarges the exposed data distribution to deduce more precise informativeness evaluation over samples — in turn — better acquisitions. In this section, we propose a principled framework that searches for feasible DA configurations for different data pools in their own natural habits. It is worth noting that we adopt a dynamic control on the strength of augmentations across different cycles, making them adaptable to changes in AL as the cycle proceeds. It can be empirically verified that such dynamic control is better than a fixed augmentation strategy (Section 3.2). Through appropriate strength control, we expect to increase the quality of augmentations for AL.

Strength for unlabeled data. The primary goal for augmenting unlabeled data is to offer precise informativeness evaluation with enriched data distribution, thus inducing more reliable acquisitions. A problem is that the weak augmentations contain trivial augmentations that contribute little to the distribution enrichment, while drastic augmentations introduce excessive distribution drifts that mislead the acquisition. We resolve this problem by proposing a proper strength that maximizes the overall informativeness of the unlabeled pool:

$$s_u = \arg\max_{s} \sum\nolimits_{x_U \in \mathcal{D}_U} \min \Big\{ \mathbb{H}(\tilde{x}_U), \\ \text{where } \tilde{x}_U \in \mathcal{T}^{(s)}(x_U) \text{ and } f_\theta(\tilde{x}_U) = f_\theta(x_U) \Big\}, \tag{1}$$

where $\mathbb{H}$ denotes the entropy. By adopting a max-min optimization procedure, we eliminate the potential harmful distribution drift caused by over-aggressively augmented samples with $\min\{\mathbb{H}(\tilde{x}_U)\}$, while also maximizing the overall informativeness of unlabeled data with $\arg\max$.

Strength for labeled data. By involving augmentations in model training, we aim at obtaining a dependable base model from limited labeled data and further enhance the acquisition. Different from the augmentations for unlabeled data that maximize overall informativeness, the augmentations for labeled data are prone to training stability and convergence. To give out proper control over labeled augmentations while avoiding extra training costs, we introduce a virtual loss term $\mathcal{L}_f$ and search the proper strength s_l for labeled samples by minimizing it:

$$s_l = \arg\min_{s} \frac{1}{|\mathcal{D}_L|} \sum\nolimits_{x_L \in \mathcal{D}_L} \mathcal{L}_f(x_L, s), \\ \text{where } \mathcal{L}_f(x,s) = \mathcal{L}(x) + \lambda \,\text{JS}\left(\{\mathcal{P}(y \mid \tilde{x}; f_\theta) \mid \tilde{x} \in \mathcal{T}^{(s)}(x)\}\right) \tag{2}$$

where $\mathcal{L}(x), \mathcal{L}_f(x)$ denotes the normal loss term and the augmented loss respectively, and λ denotes a fixed weight. Since we focus on training stability and convergence at this stage, we integrate the augmented information into the model by making them produce similar outputs, in which the dissimilarity is quantified with a Jensen-Shannon (JS) divergence term.

With the strengths s_u, s_l given above, we locate augmentations $\mathcal{T}_u$ that effectively enlarge the distribution, and the augmentations $\mathcal{T}_l$ that help deduce dependable classifiers. The combination of the two enables us to enhance acquisitions by making classifiers and informativeness evaluations in the AL framework work collaboratively. We will show how augmentations for unlabeled (UA) and labeled samples (LA) contribute to the acquisition in Section 3.2.

2.3. Controllable Augmentation-induced Acquisition for Active Learning

With the properly-controlled augmentations in Section 2.2, we proceed to integrate the augmentations into the data acquisition stage within active learning and propose controllable augmentation-induced acquisition. A key challenge for inducing the enhanced acquisition h_{acq} arises from the complicated forms for h_{base}, which denotes basic acquisitions and varies across different studies. In this section, we

highlight two types of acquisitions, i.e., score-based acquisition and representation-based acquisition and treat them differently. Since training a classifier f_θ with augmentations is straightforward, we focus on formulating enhanced acquisition of unlabeled data.

Integrating augmentations into score-based acquisitions. For score-based acquisitions, h_{base} provides an information score for each data point, according to which we select valuable samples with the highest score, like Max Entropy (Settles, 2009). To integrate data augmentations into score-based acquisitions , we calculate an information score $h_{base}(\tilde{x})$ for every augmented counterpart $\tilde{x} \in \mathcal{T}(x)$ and aggregate them into one score. We propose several variants of h_{acq}, including:

1. $h_{acq}(x) = \min_{\tilde{x}\in\mathcal{T}_u(x)} h_{base}(\tilde{x})$ provides the minimum acquisition score across all the augmented counterparts;
2. $h_{acq}(x) = \sum_{\tilde{x}\in\mathcal{T}_u(x)} h_{base}(\tilde{x})$ sums up all the information score provided by augmentations;
3. $h_{acq}(x) = \sum_{\tilde{x}\sim\mathcal{T}_u(x)} \text{sim}(x,\tilde{x}) h_{base}(x)$ weights the informativeness of $\tilde{x}$ by its similarity to its non-augmented counterpart.

Integrating augmentations into representation-based acquisitions. For representation-based acquisitions, h_{base} provides a feature vector embedded into a representation space and performs sampling according to this space, like Core-set (Sener et al., 2018). Notice that representation-based methods rely on a distance function to measure the correlation between instances, we generalize the distance functions between individual samples to point-set distance functions between augmented sample batches. By adopting set distance functions, we enhance the acquisition process by taking the correlation across augmentations over different samples into consideration. To this end, we focus on well-defined set distance functions and propose the corresponding variants as follows:

1. Standard distance $d(x,z) = \min_{\tilde{x}\in\mathcal{T}_u(x),\tilde{z}\in\mathcal{T}_u(z)} \|\tilde{x}-\tilde{z}\|_2^2$;
2. Chamfer distance $d(x,z) = \sum_{\tilde{x}\in\mathcal{T}_u(x)} \min_{\tilde{z}\in\mathcal{T}_u(z)} \|\tilde{x}-\tilde{z}\|_2^2 + \sum_{\tilde{z}\in\mathcal{T}_u(z)} \min_{\tilde{x}\in\mathcal{T}_u(x)} \|\tilde{x}-\tilde{z}\|_2^2$ considers pairwise similarities for the augmented views from two samples;
3. Pompeiu–Hausdorff distance that highlights the maximal potential difference between two samples is $d(x,z) = \max\{\max_{\tilde{x}\in\mathcal{T}_u(x)} d(\tilde{x},\mathcal{T}_u(z)), \max_{\tilde{z}\in\mathcal{T}_u(z)} d(\mathcal{T}_u(x),\tilde{z})\}$.

Controllable DA-driven active learning cycles. With those augmentation-induced acquisitions, we complete the active learning cycle within CAMPAL. First, we generate the labeled augmentations $\mathcal{T}_l$ with properly controlled strength s_l, then produce an enhanced classifier f_θ trained over them. This makes up for the insufficient labeled information and further brings a reliable model. Second, we generate the unlabeled augmentations with an optimized strength s_u and induce the enhanced acquisition h_{acq} with $\mathcal{T}_u$ and f_θ. Notably, CAMPAL offers a dynamic strength control on augmentations across cycles, which also leads to a controllable acquisition adapting itself to the changing data pools. This augmentation-induced acquisition step provides precise information evaluation and guarantees the positive impact of augmentations, which finally helps produce better querying results. As a result, these two steps jointly ensure the quality of data to label at the end of the active learning cycle, largely boosting the performance. Our experiments in Section 3 show their separate effects as well as the combined impacts in detail. We summarize the pseudo-code of our CAMPAL in Algorithm 1 as in the Appendix B.

3. Experiments

3.1. Baselines and Datasets

We instantiated our proposed CAMPAL with several existing strategies, including 1) Entropy, 2) Least Confidence (LC), 3) Margin, 4) Core-set (Sener et al., 2018), and 5) BADGE (Ash et al., 2020). We also implement several augmentation-aggregation modes that integrate augmentations into an enhanced acquisition, including 1) MIN, 2) SUM, 3) DENSITY for Entropy, LC, Margin, and 1) STANDARD, 2) CHAMFER, 3) HAUSDORFF for Core-set, BADGE, as shown in Section 2.3 and Table 2. In this section, we specify the instantiated augmentation-acquisition with basic strategy h_{base} as its subscript and the augmentation-aggregation mode as its superscript, e.g. $\text{CAMPAL}^{\text{MIN}}_{\text{Entropy}}$. We also denote the optimal version of CAMPAL, i.e. $\text{CAMPAL}^{\text{CHAMFER}}_{\text{BADGE}}$ as CAMPAL* in Table 2. We conduct experiments on four benchmark datasets: FashionMNIST, SVHN, CIFAR-10, and CIFAR-100.

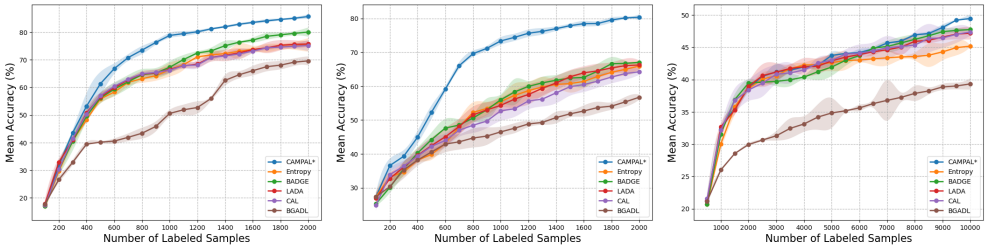
In this work, we compare our method to 1) Random, 2) Coreset, 3) BADGE (Ash et al., 2020), 4) Max Entropy, 5) Least Confidence 6) Margin. We also compare our method with other active learning strategies with data augmentations, including 1) BGADL (Tran et al., 2019), 2) CAL (Gao et al., 2020), and 3) LADA (Kim et al., 2021b) in Table 1. For a fair comparison, CAL does not use its original semi-supervised setting but uses a supervised procedure. Since LADA has multiple versions, we choose the one with the best performance for comparison in Table 1. We further prove the efficacy of CAMPAL by comparing its performance with its baseline versions in Table 3.

3.2. Main Empirical Results

CAMPAL achieves SOTA results. As shown in Table 1 and Figure 3, CAMPAL significantly outperforms their rivals on many datasets and data scales. Specifically, on the CIFAR-10 dataset, we improve upon the best baseline by **8.08**%, **16.99**%, **15.11**%, **13.34**%, where the labeled set has 500, 1000, 1500, 2000 instances respectively. Moreover,

Table 1. Comparison of the averaged test accuracy on benchmark datasets and different AL strategies. Since CAMPAL has multiple versions, we use $\text{CAMPAL}_{\text{BADGE}}^{\text{CHAMFER}}$ for comparison and denote it with CAMPAL*. The best performance in each category is indicated in boldface. N_L denotes the number of labeled samples at the end of active learning.

Dataset	Method	$N_L = 500$	$N_L = 1,000$	$N_L = 1,500$	$N_L = 2,000$
SVHN	Random	52.42±2.10	64.38±1.91	68.55±1.30	71.43±1.34
	Entropy	55.86±1.66	66.42±2.49	73.08±2.84	75.40±2.43
	BADGE	56.19±1.97	67.30±2.19	76.35±0.57	80.03±1.68
	BGADL	40.18±0.43	50.58±1.30	64.56±1.34	69.73±1.34
	CAL	56.98±1.07	66.22±0.92	72.09±1.83	75.22±2.11
	LADA	56.61±1.50	66.56±1.21	72.48±1.66	75.84±1.12
	CAMPAL*	**61.34±4.26**	**78.81±0.93**	**82.86±0.42**	**85.66±0.79**
CIFAR-10	Random	38.54±2.28	49.77±3.08	58.61±2.75	61.49±2.06
	Entropy	39.80±1.60	55.43±1.71	60.76±2.64	65.95±1.36
	BADGE	44.18±2.09	55.97±1.57	62.40±2.15	67.03±0.62
	BGADL	37.54±1.88	47.57±1.38	51.81±1.00	56.73±0.75
	CAL	40.05±1.68	54.24±2.30	59.83±2.66	64.24±0.91
	LADA	41.87±2.33	56.37±2.24	62.76±1.99	66.26±1.29
	CAMPAL*	**52.26±2.01**	**73.36±1.11**	**77.87±0.61**	**80.37±0.86**



(a) SVHN (b) CIFAR-10 (c) CIFAR-100

Figure 3. Test accuracy on the number of labeled samples over different datasets.

Table 2. Performance of CAMPAL with different h_{base} and aggregation modes. The experiment is conducted over CIFAR-10 with 2,000 labeled samples.

Method		Aggregation Mode		
Type of h_{base}	h_{base}	MIN	SUM	DENSITY
Score	Entropy	76.90±0.76	75.87±0.32	78.89±0.74
	LC	76.82±0.62	72.74±1.31	76.76±0.54
	Margin	78.70±0.58	71.33±0.76	79.16±0.48
Type of h_{base}	h_{base}	STANDARD	CHAMFER	HAUSDORFF
Representation	Core-set	78.20±0.28	79.67±0.68	78.49±0.51
	BADGE	79.71±0.51	80.37±0.86	79.84±0.24

CAMPAL exhibits the most significant performance boost with a moderately small $N_{\mathcal{L}}$, which is approximately around 1,000 for CIFAR-10 and SVHN. Besides, we can see that different versions of CAMPAL consistently achieve superior results on CIFAR-10, as shown in Table 2. As shown in Table 3, in all combinations of baselines and datasets, CAMPAL variations exhibit the best performance. Notably, CAMPAL also brings a consistent performance boost.

In addition, it is worth noting that previous works (Tran et al., 2019; Kim et al., 2021a) are typically evaluated with a large number of labeled samples (e.g., $10\% \sim 40\%$ of labeled samples for CIFAR-10). We also challenge this by querying fewer samples over benchmark datasets, shown in Table 1. When $N_{\mathcal{L}} = 500$ or 2,000 on CIFAR-10, recent augmentation-based AL strategies fail to outperform other simple baselines like BADGE. Notably, BGADL performs the worst, because of the inadequate training with insufficient instances in the current active learning setting. Since CAL is originally designed for a semi-supervised setting, it fails to outperform simple baselines like BADGE under our supervised setting. LADA outperforms other baselines on CIFAR-10 but fails on SVHN since maximal-entropy augmentations can easily change the semantics of the digit data. In contrast, our proposed CAMPAL remains competitive at different data scales, indicating its superiority.

The learnt augmentation strength differs for unlabeled/labeled data. In Figure 4, we visualize the dynamics of the learnt strength s_u^*, s_l^* across active learning cycles. In particular, we conduct the experiment 5 times on CIFAR-10 with $\text{CAMPAL}_{\text{Entropy}}^{\text{MIN}}$ and $\text{CAMPAL}_{\text{BADGE}}^{\text{STANDATD}}$ and figure out the average optimal strength value. We can observe that the s_u^* is generally larger in comparison with s_l^* across the AL cycles. This verifies our postulations that labeled data requires moderate augmentation for label preserving. In con-

Table 3. Comparison of CAMPAL with its non-augmented counterpart with different AL strategies. Δ indicates the performance boost brought by CAMPAL. Specifically, the versions of CAMPAL for comparing with 5 basic acquisitions are: $\text{CAMPAL}^{\text{DENSITY}}_{\text{Entropy}}$, $\text{CAMPAL}^{\text{MIN}}_{\text{LC}}$, $\text{CAMPAL}^{\text{DENSITY}}_{\text{Margin}}$, $\text{CAMPAL}^{\text{CHAMFER}}_{\text{Coreset}}$, $\text{CAMPAL}^{\text{CHAMFER}}_{\text{BADGE}}$, which do not change across datasets.

Dataset	Method	Entropy	LC	Margin	Coreset	BADGE
Fashion	baseline	81.33±0.86	81.15±1.16	80.71±1.16	83.36±0.82	82.89±0.95
	+CAMPAL	85.89±0.29	84.63±1.31	84.82±0.62	84.36±0.48	86.24±1.04
	Δ	+4.56±0.91	+3.48±1.75	+4.11±1.32	+1.00±0.95	+3.35±1.41
SVHN	baseline	75.40±2.43	76.39±1.30	76.32±1.87	77.81±0.93	80.03±1.68
	+CAMPAL	85.36±0.45	84.34±0.62	84.90±0.57	84.35±0.81	85.66±0.79
	Δ	+9.96±2.47	+7.95±1.44	+8.58±1.95	+6.54±1.23	+5.63±1.86
CIFAR-10	baseline	65.95±1.36	66.97±1.87	66.76±1.77	66.90±0.93	67.03±0.62
	+CAMPAL	78.89±0.74	76.82±0.62	79.16±0.48	79.67±0.68	80.37±0.86
	Δ	+12.94±1.55	+9.85±1.97	+12.40±1.83	+12.77±1.15	+13.34±1.06
CIFAR-100	baseline	45.18±0.13	45.70±0.18	45.64±0.25	46.52±0.21	47.75±0.09
	+CAMPAL	48.76±0.30	49.24±0.70	49.63±0.65	46.80±0.27	49.46±0.65
	Δ	+3.58±0.33	+3.54±0.72	+3.99±0.70	+0.28±0.34	+1.71±0.66

Table 4. Test accuracy of CAMPAL when UA or LA are individually applied over CIFAR-10 with 2,000 labeled samples. The results are produced over 5 different AL strategies. The versions of CAMPAL for comparing with 5 basic acquisitions are: $\text{CAMPAL}^{\text{MIN}}_{\text{Entropy}}$, $\text{CAMPAL}^{\text{MIN}}_{\text{LC}}$, $\text{CAMPAL}^{\text{MIN}}_{\text{Margin}}$, $\text{CAMPAL}^{\text{STANDARD}}_{\text{Coreset}}$, $\text{CAMPAL}^{\text{STANDARD}}_{\text{BADGE}}$.

Components		Entropy	LC	Margin	Core-set	BADGE
UA	LA					
		65.95±1.36	66.97±1.87	66.76±1.77	66.90±0.93	67.03±0.62
✓		67.49±1.87	69.59±2.55	71.86±3.16	68.83±1.29	71.24±0.75
	✓	74.30±0.94	75.92±0.85	77.73±0.44	77.73±0.20	78.89±0.22
✓	✓	**76.90±0.76**	**76.82±0.62**	**78.70±0.58**	**78.20±0.28**	**79.71±0.51**

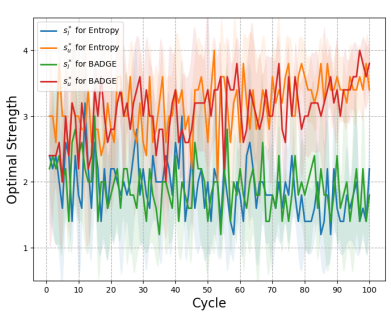
trast, unlabeled data prefers relatively stronger augmentations to enrich the data distribution such that a wider range of informative regions can be explored. We conclude that AL is better enhanced by DA with a combinatorial scheme of weak and strong augmentations applied to labeled/unlabeled data, corroborating to our theoretical findings in Section 4.

Impact of unlabeled/labeled augmentations. Here, we compare the performance boost of augmentation-induced acquisitions based on different AL strategies and the results are reported in Table 4. We can see that without augmented labeled information, the enhanced acquisition gives out a consistent performance boost over several strategies, and the maximal boost is presented by Margin ($\Delta 5.10\%$). The enhanced training process also plays an important role in promoting the performance of the existing strategies by $8.35\% \sim 11.86\%$. A combination of these two components also shows consistently best performance compared to other ablation versions. Additionally, we also compare CAMPAL with fixed augmentation strengths in Figure 4 and also see that the classifier f_θ prefers weakly labeled augmentations when stronger unlabeled augmentations induce better acquisitions, even without a dynamic strength control. We can also see that augmentations can be inefficient when strengths are not chosen appropriately. At last, we can conclude that both the augmented unlabeled information and the labeled ones help resolve the problem of unreliable judgment in AL.

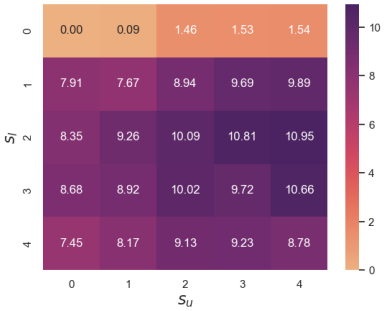
3.3. Further Extension: Integrating w/ SSL

Semi-supervised learning (SSL) is another hallmark in the history of machine learning. As is widely acknowledged, despite the similarity of AL and SSL from a high-level standpoint, their primary goals differ greatly in hindsight: the AL paradigms focus on enhancing the quality of unlabeled data queries and lowering the annotation cost, while SSL involves painstakingly proposed algorithms towards improving the representation learning by utilizing the unlabeled data during the training stage (Li et al., 2019; Zhang et al., 2021; Zheng et al., 2022; Wang et al., 2022b).

Regardless, as a versatile scheme, we entertain the probability of combining CAMPAL into the scope of several iconic SSL frameworks. In particular, we pick 1) FixMatch (Sohn et al., 2020); 2) Mean Teacher (Tarvainen & Valpola, 2017); 3) Pseudo Labeling; 4) UDA (Xie et al., 2020). In Table 5, we compare these SSL methods in their vanilla form with the CAMPAL-augmented counterpart. We see that CAMPAL yields a consistent performance boost, even on some very competitive baselines such as FixMatch.



(a) The learned strength on CIFAR-10, with CAMPAL instantiated with Entropy/BADGE.



(b) A heatmap visualization of performance boost brought by augmentations of different strengths.

Figure 4. Experiments that focus on the strengths of augmentations being applied to the labeled and unlabeled pool. The experiments are performed over CIFAR-10 with 2,000 labeled samples over Entropy.

Table 5. Comparison of SSL methods with its CAMPAL*-augmented counterpart when integrating them into AL strategies on CIFAR-10. We present our results with different sizes of $\mathcal{D}_L$. Since the labeled pool is not carefully selected and balanced as FixMatch did in an AL setting, the accuracy of FixMatch is not as high as the original paper claimed at the same data scale.

Method	$N_L = 40$	$N_L = 250$	$N_L = 1,000$
FixMatch	67.89±0.22	82.37±0.05	92.13±0.13
+CAMPAL*	68.34±0.10	82.98±0.07	92.28±0.09
Δ	+0.55±0.24	+0.61±0.09	+0.15±0.16
Mean Teacher	28.79±2.03	60.79±1.80	83.35±0.85
+CAMPAL*	33.24±1.05	63.22±1.28	84.40±0.99
Δ	+4.45±2.29	+2.43±2.05	+1.05±1.30
Pseudo Label	25.73±0.83	41.19±1.83	64.48±0.73
+CAMPAL*	30.78±1.01	54.39±1.33	74.46±0.88
Δ	+5.05±1.31	+13.20±2.26	+9.98±1.14
UDA	56.34±1.34	77.57±1.08	82.66±0.17
+CAMPAL*	58.35±0.78	79.33±0.92	83.35±0.09
Δ	+1.99±1.55	+1.76±1.42	+0.79±0.19

Table 6. Comparison of performance on CAMPAL*, FixMatch, and the framework integrated with both of them when applying them on imbalanced CIFAR-10. We present our results with different imbalance factors IF with 1,000 samples.

Method	$IF = 0.2$	$IF = 0.1$	$IF = 0.01$
Baseline	58.82±1.24	54.72±0.34	46.74±0.18
+CAMPAL*	65.74±0.21	63.82±0.22	51.02±0.26
+FixMatch	68.41±0.77	64.18±0.12	50.34±0.37
+Both	**71.14±0.45**	**67.34±0.15**	**53.83±0.14**

In spite of the decent performance attained from advanced SSL methods like FixMatch, they often assume an underlying balanced distribution in the labeled sample pool. However, in real-world applications, this balancing condition can hardly be satisfied (Wei et al., 2022b; Wang et al., 2022b; Wei et al., 2022a). As such, in an imbalanced setup, the FixMatch algorithm plummets as displayed in Table 6. In hindsight, we may posit that this is very much due to an inherent but indispensable lack of data querying/selection mechanism to form the labeled data pool. To fill this void, we integrate CAMPAL (see Appendix B.4 for implementation details) with the FixMatch algorithm. The results are summarized in Table 6. We observe that CAMPAL performs synergistically with FixMatch, respectively attaining **12.48%** and **7.09%** performance gain over the baseline, under an imbalancing factor of 0.1 and 0.01.

Remark. The paradigm of SSL has marked a few milestones in recent years. Nevertheless, as is analyzed previously, we postulate that its methodology remains room to progress. In particular, its innate lack of appropriate data selection mechanism evidently causes performance plunge in certain scenarios. In that regard, CAMPAL is justified and capable to interplay with the iconic SSL framework in a holistic manner. We believe the complementing marriage of the AL and SSL frameworks can further foster their deployment towards real-world applications.

4. Theoretical Analysis

In this section, we theoretically analyze why weak and strong augmentations being strategically applied to labeled and unlabeled data exhibit the best performance. Following the previous sections, we use f_θ to denote the model fully trained. When an unlabeled sample lies within the augmented region for a particular labeled sample, we can propagate the labeled information to the corresponding unlabeled samples. Formally, with a feature map f_θ^{emb} derived from f_θ we define a covering relation:

Definition 4.1. Given an augmentation set $\mathcal{T}$, we say that an image x is covered by x_i with respect to the augmentation set $\mathcal{T}$, if $f_\theta^{\text{emb}}(x)$ lies within the convex hull of the augmented views of x_i: $f_\theta^{\text{emb}}(x) \in \text{conv}\left(f_\theta^{\text{emb}}\left(\mathcal{T}(x_i)\right)\right)$. We denote the covering relation by $x \triangleleft x_i$.

Without loss of generality, assume there are L labeled sam-

ples $x_1, \ldots, x_L$, together with the unlabeled samples covered by its augmentations, constituting L components. For each component $C_i (i = 1, \ldots, L)$, let P_i be a probability that a data point sampled from the underlying data distribution covered by C_i. To make the analysis tractable, we assume that the properly controlled augmentations for labeled samples never overlaps:

Assumption 4.2. With moderately weak augmentations for labeled samples, C_i's do not overlap with each other, i.e. $\forall i \neq j$, $\mathcal{P}(C_i \cap C_j \neq \emptyset) = 0$.

With Assumption 4.2, the error for f_θ can be estimated by how these components cover the data space. To further illustrate this, we provide a comparison between different augmentations in Figure 5. The following proposition characterizes the relationship between error and components.

Proposition 4.3. *Let $\mathcal{E}$ denote the probability that the f_θ cannot infer the correct label of a test example. Then $\mathcal{E}$ is upper bounded:*

$$\mathcal{E} \leq \sum\nolimits_{i=1}^{L} P_i(1-P_i)^m + \left(1 - \sum\nolimits_{i=1}^{L} P_i\right), \quad (3)$$

where m denotes the number of samples that lie within the labeled components.

In Eq. (3), the first term denotes the risk brought by ill-defined augmentations, while the second term denotes subsample empirical risk. With Eq. (3), we continue to reduce the error as much as possible by acquiring informative samples. By adding a newly queried sample x_{L+1}, the error reduction is estimated as follows:

$$\begin{aligned}\Delta\mathcal{E}(\Delta m, P_{L+1}) \approx & \sum\nolimits_{i=1}^{L} P_i(1-P_i)^m \left(1-(1-P_i)^{\Delta m}\right) \\ & - P_{L+1}\left(1-(1-P_{L+1})^{m+\Delta m}\right),\end{aligned} \quad (4)$$

where Δm is the number of samples newly covered.

We take a step further by illustrating two terms in Eq. 4. The first term denotes the performance boost brought by newly-annotated samples. Specifically, the samples that drift farthest from the existing components better cover the under-explored data space, indicating a larger Δm – in turn – the performance boost. This is also consistent with the max-min optimization objective for unlabeled samples described in Eq. (1), with the intuition provided in Figure 5(c),(d). The second term characterizes the potential error induced from augmentations on unlabeled samples, i.e., too strong augmentation excessively increases the value of P_{L+1}, leading to its increase. Therefore, it is important to locate moderately strong augmentations for unlabeled data.

Theorem 4.4. *With properly selected augmentation sets and sufficient large L, the maximal value for error reduction $\Delta\mathcal{E}(\Delta m, P_{L+1})$ with newly-annotated samples can be estimated as follows:*

$$\Delta\mathcal{E}(\Delta m, P_{L+1}) \lessapprox \mathcal{E}\left(1 - Ke^{-m/L}\right), \quad (5)$$

where U denotes the number of unlabeled samples, with $K = \frac{m+L(\log(L+U)-\log L-1)}{L+U}$.

From the theorem, we can see that properly selected samples and augmentations give out a significant error reduction. Specifically, m/L denotes the average number of samples covered by each component, which indicates better coverage induced from properly controlled components when being larger. With all the discussions above, augmentation-acquisition integration effectively relies on the quality of augmentations, where better augmentations result in more dependable classifiers and larger error reduction.

5. Related Works

Data augmentation is a technique that improves the generalization ability of models by increasing the number of images in a dataset (Xu et al., 2022). The most commonly used augmentation techniques include geometric transformations (Shorten & Khoshgoftaar, 2019), generative adversarial networks (Bowles et al., 2018), and image mixing (Zhang et al., 2018; Yun et al., 2019). Instead of designing new types of augmentations, recent studies optimize the strength of an augmentation group (Cubuk et al., 2020).

Active learning is a machine learning paradigm that actively selects the data it wants to learn from the unlabeled data sources (Ren et al., 2021). The most crucial part of active learning is exactly the data acquisition. Current studies can be roughly categorized as follows: (a) Uncertainty-based methods that prefer the hardest samples (Mai et al., 2022; Wang et al., 2022a); (b) Representation-based methods searching for the samples that are the most representative of the underlying data distribution (Sener et al., 2018; Ash et al., 2020). To date, the unreliable informativeness evaluation with few samples remains a critical issue.

6. Conclusions

In this work, we propose a novel active learning framework CAMPAL. By devising adaptable control policies on data augmentation integrated for active learning, CAMPAL attains state-of-the-art performance with a significant boost. Our theoretical analysis further justifies the innate reliance of AL on the quality of introduced augmentations. In addition, we evidently exhibit that CAMPAL is capable to enhance popular semi-supervised learning frameworks, expressing as a holistic system. In the future, we hope to generalize CAMPAL to more tasks.

Acknowledgements

This work is majorly supported by the National Key Research and Development Program of China (No. 2022YFB3304101), and in part by the NSFC Grants (No. 62206247). JY, HW, and JZ also thank the sponsorship by the Fundamental Research Funds for the Central Universities and CAAI-Huawei Open Fund.

References

Ash, J. T., Zhang, C., Krishnamurthy, A., Langford, J., and Agarwal, A. Deep batch active learning by diverse, uncertain gradient lower bounds. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=ryghZJBKPS.

Bhattacharjee, S. D., Talukder, A., and Balantrapu, B. V. Active learning based news veracity detection with feature weighting and deep-shallow fusion. In *2017 IEEE International Conference on Big Data (Big Data)*, pp. 556–565. IEEE, 2017.

Bowles, C., Chen, L., Guerrero, R., Bentley, P., Gunn, R. N., Hammers, A., Dickie, D. A., del C. Valdés Hernández, M., Wardlaw, J. M., and Rueckert, D. GAN augmentation: Augmenting training data using generative adversarial networks. *CoRR*, abs/1810.10863, 2018. URL http://arxiv.org/abs/1810.10863.

Cao, K., Wei, C., Gaidon, A., Arechiga, N., and Ma, T. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.

Caramalau, R., Bhattarai, B., and Kim, T.-K. Sequential graph convolutional network for active learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9583–9592, 2021.

Cubuk, E. D., Zoph, B., Shlens, J., and Le, Q. V. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 702–703, 2020.

Feng, D., Wei, X., Rosenbaum, L., Maki, A., and Dietmayer, K. Deep active learning for efficient training of a lidar 3d object detector. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pp. 667–674. IEEE, 2019.

Gao, M., Zhang, Z., Yu, G., Arık, S. Ö., Davis, L. S., and Pfister, T. Consistency-based semi-supervised active learning: Towards minimizing labeling cost. In *European Conference on Computer Vision*, pp. 510–526. Springer, 2020.

Hendrycks, D., Mu, N., Cubuk, E. D., Zoph, B., Gilmer, J., and Lakshminarayanan, B. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019.

Hussein, A., Gaber, M. M., and Elyan, E. Deep active learning for autonomous navigation. In *International Conference on Engineering Applications of Neural Networks*, pp. 3–17. Springer, 2016.

Kim, K., Park, D., Kim, K. I., and Chun, S. Y. Task-aware variational adversarial active learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8166–8175, 2021a.

Kim, Y.-Y., Song, K., Jang, J., and Moon, I.-C. Lada: Look-ahead data acquisition via augmentation for deep active learning. *Advances in Neural Information Processing Systems*, 34:22919–22930, 2021b.

Li, Y.-F., Wang, H., Wei, T., and Tu, W.-W. Towards automated semi-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 4237–4244, 2019.

Mai, X., Avestimehr, S., Ortega, A., and Soltanolkotabi, M. On the effectiveness of active learning by uncertainty sampling in classification of high-dimensional gaussian mixture data. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4238–4242. IEEE, 2022.

Ren, P., Xiao, Y., Chang, X., Huang, P.-Y., Li, Z., Gupta, B. B., Chen, X., and Wang, X. A survey of deep active learning. *ACM computing surveys (CSUR)*, 54(9):1–40, 2021.

Sener et al., S. S. Active learning for convolutional neural networks: A core-set approach. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=H1aIuk-RW.

Settles, B. Active learning literature survey. 2009.

Shorten, C. and Khoshgoftaar, T. M. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

Sinha, S., Ebrahimi, S., and Darrell, T. Variational adversarial active learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5972–5981, 2019.

Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C. A., Cubuk, E. D., Kurakin, A., and Li, C.-L. Fixmatch: Simplifying semi-supervised learning with

consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

Tarvainen, A. and Valpola, H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

Tran, T., Do, T.-T., Reid, I., and Carneiro, G. Bayesian generative active deep learning. In *International Conference on Machine Learning*, pp. 6295–6304. PMLR, 2019.

Wang, T., Li, X., Yang, P., Hu, G., Zeng, X., Huang, S., Xu, C.-Z., and Xu, M. Boosting active learning via improving test performance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8566–8574, 2022a.

Wang, X., Lian, L., and Yu, S. X. Unsupervised selective labeling for more effective semi-supervised learning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXX*, pp. 427–445. Springer, 2022b.

Wei, T., Shi, F., Wang, H., Li, W.-W. T., et al. Mixpul: consistency-based augmentation for positive and unlabeled learning. *arXiv preprint arXiv:2004.09388*, 2020.

Wei, T., Liu, Q.-Y., Shi, J.-X., Tu, W.-W., and Guo, L.-Z. Transfer and share: semi-supervised learning from long-tailed data. *Machine Learning*, pp. 1–18, 2022a.

Wei, T., Shi, J.-X., Li, Y.-F., and Zhang, M.-L. Prototypical classifier for robust class-imbalanced learning. In *Advances in Knowledge Discovery and Data Mining: 26th Pacific-Asia Conference, PAKDD 2022, Chengdu, China, May 16–19, 2022, Proceedings, Part II*, pp. 44–57. Springer, 2022b.

Wu, T.-H., Liu, Y.-C., Huang, Y.-K., Lee, H.-Y., Su, H.-T., Huang, P.-C., and Hsu, W. H. Redal: Region-based and diversity-aware active learning for point cloud semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15510–15519, 2021.

Xie, Q., Dai, Z., Hovy, E., Luong, T., and Le, Q. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33: 6256–6268, 2020.

Xu, M., Yoon, S., Fuentes, A., and Park, D. S. A comprehensive survey of image augmentation techniques for deep learning. *arXiv preprint arXiv:2205.01491*, 2022.

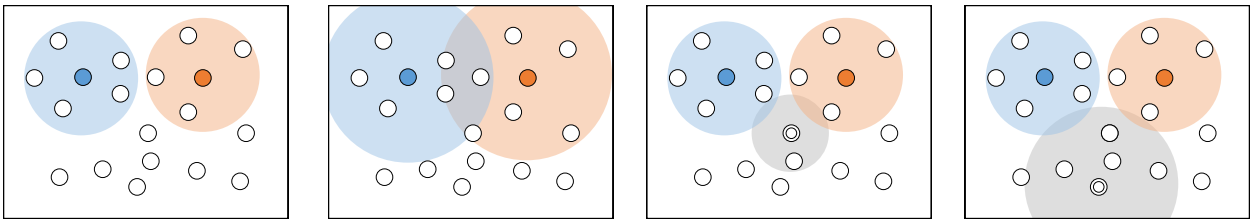
Yun, S., Han, D., Chun, S., Oh, S. J., Yoo, Y., and Choe, J. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 6022–6031. IEEE, 2019. doi: 10.1109/ICCV.2019.00612. URL https://doi.org/10.1109/ICCV.2019.00612.

Zhang, B., Li, L., Yang, S., Wang, S., Zha, Z.-J., and Huang, Q. State-relabeling adversarial active learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8756–8765, 2020.

Zhang, B., Wang, Y., Hou, W., Wu, H., Wang, J., Okumura, M., and Shinozaki, T. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34: 18408–18419, 2021.

Zhang, H., Cissé, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=r1Ddp1-Rb.

Zheng, M., You, S., Huang, L., Wang, F., Qian, C., and Xu, C. Simmatch: Semi-supervised learning with similarity matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14471–14481, 2022.



(a) (b) (c) (d)

Figure 5. The coverage on the data space presented by augmentations, where colored circles are labeled samples, white circles are unlabeled samples and the colored shade denotes the region covered by corresponding augmentations. A double circle denotes the unlabeled sample to be annotated. The figures above show: a) Proper augmentation for labeled samples; b) Drastic augmentation for labeled samples; c) Sub-optimal unlabeled sample with the corresponding augmentation; d) A proper unlabeled sample with the corresponding augmentation.

A. Theoretical Analysis

This section provides a complete derivation for the analysis given in Section 4. An intuition for this is given in Figure 5. Before the actual acquisition process, we must ensure convergence for the underlying classifier. Specifically, with proper augmentations over labeled data and the approximate loss term in (3), we can deduce the upper bound for $Pr(\mathcal{A})$ and guarantee the convergence for training, shown as follows:

Theorem A.1. *Under the setting for CAMPAL, Let $\mathcal{E}$ denote the probability that the classifier cannot infer the label of newly given samples drawn from the underlying data space, with L labeled samples given in $\mathcal{D}_L$ and augmentation set $\mathcal{T}$. Then $\mathcal{E}$ is upper bounded by $\hat{\mathcal{E}}$ as follows:*

$$\mathcal{E} \leq \hat{\mathcal{E}}(\mathcal{D}_L, f_\theta, \mathcal{T}) = \sum_{l=1}^{L} P_i(1-P_i)^m + \left(1 - \sum_{i=1}^{L} P_i\right), \tag{6}$$

With properly selected augmentation set $\mathcal{T}$ and sufficient large L, $\hat{\mathcal{E}}$ can be estimated by $O(\varepsilon)$ with $O(L/\varepsilon)$ samples covered by labeled components, i.e.

$$m = O(L/\varepsilon) \Rightarrow \hat{\mathcal{E}} \lessapprox O(\varepsilon) \Rightarrow \mathcal{E} \leq O(\varepsilon). \tag{7}$$

Proof. With proper control over augmentations, we assume that each component does not overlaps with at most one other component in Proposition 3, which can be controlled with appropriate augmentations, and generalizable to multiple components. Let x be the sampled example, the probability of x not covered only in one of C_i's is

$$\begin{aligned}\hat{\mathcal{E}} &= \mathcal{P}\left(\exists i \neq j,\, x' \in C_i \cap C_j\right) + \mathcal{P}\left(x' \text{ is uncovered}\right) \\ &= \sum_{l=1}^{L} P_i(1-P_i)^m + \left(1 - \sum_{i=1}^{L} P_i\right)\end{aligned}$$

With sufficiently large L, we can also have a component set that covers the entire dataset, leading to $\sum_i P_i = 1$. Now it remains to find the maximum value of $\sum_{l=1}^{L} P_i(1-P_i)^m$ to bound the error term, with the following optimization objective:

$$\min_C -\sum_i P_i(1-P_i)^m, \text{ s.t. } \sum_i P_i = 1.$$

With the KKT condition, we attain its maximum value when all P_i is set to $\frac{1}{L}$, i.e. $\hat{\mathcal{E}} \lessapprox (1-\frac{1}{L})^m$. With $O(L/\varepsilon)$ and sufficiently large L, we have

$$\hat{\mathcal{E}} \lessapprox \exp\left(-\frac{m}{L}\right) = \exp\left(-O(\frac{1}{\varepsilon})\right) \leq O(\varepsilon).$$

□

With the conditions in theorem A.1, it remains to consider the approximate boost provided by the reduction on upper bound $\hat{\mathcal{E}}$:

$$\begin{aligned}\Delta\hat{\mathcal{E}}(\Delta m, P_{L+1}) &= \hat{\mathcal{E}}(\mathcal{D}_L \cup \{x_{L+1}\}, f_\theta, \mathcal{T}\}) - \hat{\mathcal{E}}(\mathcal{D}_L, f_\theta, \mathcal{T})) \\ &= \sum_{i=1}^{L} P_i(1-P_i)^m \left(1-(1-P_i)^{\Delta m}\right) - P_{L+1}\left(1-(1-P_{L+1})^{m+\Delta m}\right).\end{aligned}$$

where Δm is the number of samples newly covered after labeling x_{L+1}.

Theorem A.2. *With the conditions given in Theorem A.1, the maximal value for error bound reduction $\Delta\hat{\mathcal{E}}(\Delta m, P_{L+1})$ with newly-annotated samples can be estimated as follows:*

$$\Delta\mathcal{E}(\Delta m, P_{L+1}) \lessapprox \mathcal{E}\left(1 - Ke^{-m/L}\right), \tag{8}$$

where U denotes the number of unlabeled samples, with

$$K = \frac{m\log(L+U) - L\left(\log L - 1\right)}{L+U}$$

.

Proof. Under this setting, P_{L+1} appears to be proportional to Δm, when no unnecessary overlap appears across components (guaranteed by Theorem A.1). Therefore, we can estimate $P_{L+1} \approx \Delta m/(L+U)$, where U denotes the number of unlabeled samples. With those conditions, we estimate the relative error reduction as follows:

$$\begin{aligned}\frac{\Delta\hat{\mathcal{E}}}{\hat{\mathcal{E}}} &= \frac{\sum_{i=1}^{L} P_i(1-P_i)^m\left(1-(1-P_i)^{\Delta m}\right) - P_{L+1}\left(1-(1-P_{L+1})^{m+\Delta m}\right)}{\sum_{i=1}^{L} P_i(1-P_i)^m + \left(1-\sum_{i=1}^{L} P_i\right)} \\ &\lessapprox \left(1-\left(1-\frac{1}{L}\right)^{\Delta m}\right) - \exp\left(-\frac{m}{L}\right)\frac{\Delta m}{L+U}\left(1-\left(1-\frac{\Delta m}{L+U}\right)^{m+\Delta m}\right)\end{aligned}$$

Since m is large with sufficient labeled samples, we can further estimate this term as:

$$\frac{\Delta\hat{\mathcal{E}}}{\hat{\mathcal{E}}} \lessapprox 1 - \left(1-\frac{1}{L}\right)^{\Delta m} - \exp(-\frac{m}{L})\frac{\Delta m}{L+U}.$$

Then the maximum value for this is attained when Δm reaches

$$\begin{aligned}\Delta m^* &= \frac{1}{\log\left(1-\frac{1}{L}\right)}\left(\frac{m}{L} - \log\left(\frac{1}{L+U}\right) - \log\left(-\log\left(1-\frac{1}{L}\right)\right)\right) \\ &\approx L\left(\frac{m}{L} - \log\left(\frac{1}{L+U}\right) + \log(\frac{1}{L})\right) = m + L\left(\log\left(L+U\right) - \log L\right)\end{aligned}$$

Then

$$\begin{aligned}\frac{\Delta\hat{\mathcal{E}}}{\hat{\mathcal{E}}} &\approx 1 + \frac{1}{L+U}\left(\frac{1}{\log(1-\frac{1}{L})} - \Delta m^*\right)\exp\left(-\frac{m}{L}\right) \\ &\approx 1 - \frac{m + L\left(\log(L+U) - \log L - 1\right)}{L+U}\exp\left(-\frac{m}{L}\right).\end{aligned}$$

□

Algorithm 1 An active learning cycle for CAMPAL.

Input: Labeled data pool $\hat{\mathcal{D}}_L$, Unlabeled data pool $\mathcal{D}_U$, Model f_θ.
$\theta \leftarrow \arg\min_\theta \frac{1}{|\mathcal{D}_L|} \sum_{x \in \mathcal{D}_L} \mathcal{L}\left(f_\theta(x), y\right)$;
$m_l = \arg\min_m \frac{1}{|\mathcal{D}_L|} \sum_{x_L \in \mathcal{D}_L} \mathcal{L}_f(x_L, m)$, where L_f is shown in (9);
Generate an augmentation set $\mathcal{T}^{(m_l)}$ with strength m_l;
$\theta \leftarrow \arg\min_\theta \frac{1}{|\mathcal{T}^{(m_{lab})}(\mathcal{D}_L)|} \sum_{x \in \mathcal{T}^{(m_l)}(\mathcal{D}_L)} \mathcal{L}\left(f_\theta(x), y\right)$;
$m_u = \arg\max_m \sum_{x_U \in \mathcal{D}_U} \min\{\mathbb{H}(\tilde{x}_U) \,|\, \tilde{x}_U \in \mathcal{T}^{(m)}(x_U),\ f_\theta(\tilde{x}_U) = f_\theta(x_U)\}$;
Generate an augmentation set $\mathcal{T}^{(m_u)}$ with strength m_u;
Deduce the enhanced acquisition h_{acq} with $\mathcal{T}^{(m_u)}$ and f_θ as shown in Section 2.3;
$\mathcal{Q}$ according to h_{acq};
$\mathcal{D}_U \leftarrow \mathcal{D}_U - \mathcal{Q}$;
$\mathcal{D}_L \leftarrow \mathcal{D}_L \cup \mathcal{Q}$.

Table 7. The list of all the hyper-parameters used in the experiments across different SSL algorithms.

Parameters	FixMatch	Mean Teacher	Pseudo Label	UDA
unlabeled data ratio	7	1	1	7
unlabeled loss ratio	1.0	50.0	1.0	1.0
temperature	0.5	None	None	0.4

B. Additional Experimental Setups and Results

B.1. Pseudo Code of CAMPAL

We given the pseudo code of CAMPAL as shown in algorithm 1.

B.2. Implementation Details

We conduct experiments on four benchmark datasets: FashionMNIST, SVHN, CIFAR-10, and CIFAR-100. We will construct a random initial dataset with 100 instances for FashionMNIST, SVHN, and CIFAR-10, and 1,000 instances for CIFAR-100. Then we acquire 100 instances for FashionMNIST, SVHN, and CIFAR-10, and 500 instances for CIFAR-100 at each cycle. We repeat the cycle 20 times. Then we generate 10 single-image augmentations and 5 mix-up augmentations for each sample. We normalize the images with the channel mean and standard deviation over all the datasets. For CIFAR-10 and CIFAR-100, we apply a standard augmentation after conducting augmentations in the pipeline. We adopt ResNet-18 as the architecture and train the model for 300 epochs with an SGD optimizer of learning rate 0.01, momentum 0.9, and weight decay 5e-4. For the virtual loss term in (9), we also set $\lambda = 1$.

B.3. Implementation Details for the Simple Application of DA for AL in Figure 1

We integrate DA into AL with fixed augmentations $\mathcal{T}$ as follows. This experiment is also conducted on dataset CIFAR-10 with a ResNet-18 architecture. The basic acquisition here is Max Entropy. First, we augment the labeled pool with $\mathcal{T}$, and train the classifier f_θ accordingly. Then we augment the unlabeled pool with $\mathcal{T}$ and perform acquisitions directly on the augmented unlabeled pool. Other settings are the same as the main empirical experiments.

B.4. Implementation Details for Integrating Semi-supervised Learning into CAMPAL

Active learning and semi-supervised learning are two different paradigms that serve different purposes. Specifically, active learning aims at selecting and querying the most efficient samples from unlabeled data pools and labeling them, emphasizing the quality of selected samples. The performance of the classifier or other types of models trained accordingly serves as a type of quality evaluation for those samples. In SSL, the labeled pool and the unlabeled pool are fixed, and researchers hope to design a training procedure that maximally utilizes unlabeled samples to boost performance. Therefore, the quality of SSL depends on the design of training algorithms. Since both of them have a labeled pool and an unlabeled pool, it is natural to consider a combination of them, constituting another subject: semi-supervised active learning. Contrary to typical

active learning that doesn't involve unlabeled samples in training, Semi-supervised Active Learning involves the unlabeled data pool $\mathcal{D}_U$ in classifier training.

By adopting the semi-supervised training paradigm into active learning, i.e. involving unlabeled samples in classifier training, we can easily make our work incorporate with recent SSL works. In detail, the labeled pool is exactly the augmented labeled pool produced by CAMPAL, when we do not involve augmented unlabeled samples in training. For works like MeanTeacher and Pseudo Labeling that are not guided augmentations, we simply apply their training paradigms into an active learning framework. For UDA, since it involves augmentation for unlabeled samples into training, we use the augmented unlabeled samples optimized by CAMPAL when integrating them. For FixMatch, we refer augmentations with strength 2 as weak augmentations and those with strength 4 as strong augmentations. For other hyper-parameters, we follow the work of FlexMatch (Zhang et al., 2021), which is a comprehensive SSL framework that contains several SSL baselines mentioned above. Specifically, all the experiments run 1048576 training iterations with a batch size 64, the model of ResNet-18, an optimizer of SGD of learning rate 0.03, momentum 0.9 and weight decay 0.0005. Some different parameters across these algorithms are shown in Table 7.

For SSL and AL under imbalanced settings, we delete samples from the original balanced set CIFAR-10 and make imbalanced datasets with EXP algorithm presented by (Cao et al., 2019) and it is commonly used to make imbalanced datasets from balanced ones. Other settings are just the same as aforementioned. The imbalance factor (IF) is the most commonly used measure to describe the imbalance extent of a dataset. IF is defined as $IR = N_{maj}/N_{min}$, where N_{maj} is the sample size of the majority class and N_{min} is the sample size of the minority class.

B.5. Augmentations Included

The details of the 19 augmentations in the (CAMPAL) with their parameters are shown in Table 8. In brief, the augmentations we use can be categorized into single-image augmentations and image-mixing. Formally, we provide an augmentation functional set that covers (i)-singular input augmentation means such as rotation for low-level image processing. The corresponding functional set is denoted by $\mathcal{T}_{single} = \{\omega(x; \lambda)\}$ where ω points to an instantiated augmentation function. The sample x is taken as an input to ω together with varying augmentation hyper-parameters λ, such as the angle in the image rotation function. Similarly, we also construct a combinatorial augmentation functional set, $\mathcal{T}_{mix} = \{\gamma(x, x'; \lambda)\}$, where the augmentation function γ takes two input samples x and x' together with hyper-parameter. With slight abuse of notations, we uniformly use λ to refer to augmentation-related hyper-parameters. In the implementation of CAMPAL, we simply adopt MixUp for combinatorial augmentation. As we can see, upon fixed input, both singular and combinatorial augmentation functional sets can be arbitrarily expanded, by varying λ in a continuous scalar space.

B.6. Additional Results Compared to RandAugment.

Since CAMPAL locate feasible augmentations guided by their strength, we also compare CAMPAL with RandAugment (Cubuk et al., 2020). To show the effectiveness of a separate control on unlabeled/labeled data in CAMPAL, we trained RandAugment on the labeled data within each AL cycle, then applied the optimized augmentation to both the labeled pool and unlabeled pool. As shown in Table 9, CAMPAL shows better performance than the RandAugment, indicating the superiority of the separate control. It should be noted that RandAugment is originally designed for training over full labeled data, but is obliged to be conducted over the labeled pool with limited samples under the AL setting. Therefore, directly adopting RandAugment to AL is infeasible, since it can be heavily biased towards limited labeled data, contributing little to the distribution enrichment on unlabeled data.

We also supplemented a set of experiments on CIFAR-10 and observe their full-cycle performances. Table 10 shows comparative results on CIFAR-10 with fewer samples, showing the performance boost is significant with 200 annotated samples (6.30%). Since CAMPAL optimizes augmentation strengths for labeled/unlabeled samples at each cycle towards their own objectives, it does not rely on data quantity and is much more flexible than RandAugment which relies on fixed parameters. The performance difference seems to amplify especially with fewer labeled samples at early stages for AL. For approaches like RandAugment, there are normally two ways of specific implementation: For one thing, one can always obtain a new RandAugment policy on the provided task associated with the dataset. However, this is not quite realistic in the active learning setup where the labeled data is limited and every step of label acquisition extracts a toll. What is more adverse is that the original RandAugment approach evidently requires a large number of validation samples (e.g. 10,000 samples on CIFAR-10) to be involved. For the other, one can obtain a policy from external source then transfer it to the current setting. While this is functionally feasible, we do observe no strong performance gain from the above table. While

Table 8. The list of all the augmentations used in the experiments. The letter x or x_* denotes given images. $\mathcal{U}(a, b)$ denotes a continuous uniform distribution at interval $[a, b]$, when $\mathcal{B}(a, b)$ denotes a beta distribution with parameters a and b.

Augmentation	Parameters	Description
AutoContrast(x)		Maximizing the (normalize) image contrast
Brightness(x,v)	$v \sim \mathcal{U}(1, 1.18)$: an enhancing factor	Enhancing the brightness of a given image
Color(x,v)	$v \sim \mathcal{U}(1, 1.18)$: an adjustment factor	Adjust the color balance of a given image
Contrast(x,v)	$v \sim \mathcal{U}(1, 1.18)$:	Enhancing the contrast of a given image
CutOut(x,v)	$v \sim \mathcal{U}(0.09, 0.11)$: CutOut ratio	Cut out a part of image and fill with black
CutOutAbs(x,v)	$v \sim \mathcal{U}(0.09, 0.11)$: CutOut ratio	Cut out a part of image and fill with gray
Equalize(x)		Equalize the image histogram
Identity(x)		Return the image itself
Invert(x)		Invert all pixel values
Posterize(x,v)	$v \sim \mathcal{U}(6.0, 6.4)$: Posterizing degree	Posterizing the image
Rotate(x,v)	$v \sim \mathcal{U}(20, 30)$: Rotation degree	Rotating the image
Sharpness(x,v)	$v \sim \mathcal{U}(1, 1.18)$: Sharpen degree	Sharpen the image
ShearX(x,v)	$v \sim \mathcal{U}(0.15, 0.18)$: Affinity degree	Affine transformation in x-axis
ShearY(x,v)	$v \sim \mathcal{U}(0.15, 0.18)$: Affinity degree	Affine transformation in y-axis
Solarize(x,v)	$v \sim \mathcal{U}(96, 128)$: Solarization degree	Solarizing the image
SolarizeAdd(x,v)	$v \sim \mathcal{U}(50, 60)$: Solarization degree	Solarizing the image and add back
TranslateX(x,v)	$v \sim \mathcal{U}(0.1, 0.15)$: translation ratio	Translating the image in x-axis
TranslateY(x,v)	$v \sim \mathcal{U}(0.1, 0.15)$: translation ratio	Translating the image in y-axis
MixUp(x,x_*, λ)	λ: the mixing ratio	Mix up the two given images

Table 9. Test accuracy of CAMPAL and augmentation-induced acquisition with learned RandAugment.

Method	Fashion	SVHN	CIFAR-10	CIFAR-100
Ent w. RA	86.15±0.89	82.84±1.12	76.83±0.82	46.70±0.34
$\text{CAMPAL}^{\text{DENSITY}}_{\text{Entropy}}$	86.17±0.58	83.49±0.96	78.89±0.74	48.76±0.30

this way of transferring a pre-trained RandAugment policy is promising, we believe it is also pivotal in terms of how to transfer/fine-tune it in the course of an AL setting. This may have gone beyond the scope of this paper.

B.7. Ablation Studies over Types of Augmentations

The impact of each single-image augmentation operator on CAMPAL. To further dive into the impact of the contribution of augmentations, we also provide the results when each augmentation is separately applied to CAMPAL with different strengths, shown in Table 11 on CIFAR-10 with $\text{CAMPAL}^{\text{DENSITY}}_{\text{Entropy}}$. We can see the impact of different types of single-image augmentations varies. An interesting observation is that different augmentation operator does not contribute equally at the different AL cycles. For example, Sharpness performs better than Rotate when $\mathcal{N}_L = 500$, but underperforms Rotate when $\mathcal{N}_L = 2000$. It reveals a sophisticated mechanism of the benefit of these augmentation operators on AL. However, the profound theory behind why data augmentation works have not been fully revealed to date, making it difficult to principally pick up the best optimal augmentation type. Hence, we naively adopt a simple strategy that uniformly selects and stacks these operators to enjoy their mixed benefits to AL.

Effect of single-image augmentations and mix-up. To prove the efficacy of including both single-image augmentations of image-mixing into one query batch, we further explore the effect of these two kinds of augmentations separately. To verify this, we conduct experiments over two variants of CAMPAL that only use one type of augmentations, i.e. single-image augmentations and MixUp. The tests are performed by the ResNet-18 model with 4% (2000) data from CIFAR-10. For fairness, when only one kind of augmentation is used, we generate 15 augmented samples of this type. In Table 12, we can

Table 10. Test accuracy of CAMPAL and augmentation-induced acquisition on CIFAR-10 with different pretrained parameters and data scales.

Method	$N_L = 200$	$N_L = 500$	$N_L = 1,000$	$N_L = 2,000$
Ent w. RA(CIFAR-10)	33.96	55.55	68.17	76.83
Ent w. RA(ImageNet)	34.39	53.14	66.90	77.85
CAMPAL	40.26	58.06	71.40	78.89

Table 11. Comparison of the averaged test accuracy when each type of augmentation is separately integrated into CAMPAL. We ran each experiment on CIFAR-10 with 2,000 samples annotated at the last cycle, and repeat them 5 times. $\mathcal{N}_L$ denotes the number of labeled samples.

Augmentation	$\mathcal{N}_L = 500$	$\mathcal{N}_L = 1,000$	$\mathcal{N}_L = 1,500$	$\mathcal{N}_L = 2,000$
None	39.80±1.60	55.43±1.71	60.76±2.64	65.95±1.36
AutoContrast	48.59±1.12	63.35±0.30	72.77±0.12	76.36±0.28
Brightness	45.63±0.26	60.16±2.25	69.50±0.26	74.25±0.16
Color	50.84±3.50	62.04±1.75	72.24±0.25	76.64±0.46
Contrast	49.77±3.91	56.52±1.32	68.95±1.61	74.47±0.63
CutOut	47.91±2.66	62.29±3.69	71.37±1.19	76.62±0.34
CutOutAbs	53.59±0.30	63.96±0.57	67.94±0.19	71.94±0.18
Equalize	49.38±1.93	63.18±1.86	69.86±0.78	74.25±0.27
Invert	51.73±0.29	63.65±0.12	71.89±0.33	75.98±0.28
Posterize	49.02±1.16	64.24±1.30	72.25±0.90	75.62±0.64
Rotate	44.60±1.39	56.47±0.21	63.08±1.22	67.60±0.45
Sharpness	47.31±2.35	62.40±1.15	70.98±1.26	74.38±2.46
ShearX	45.19±0.39	58.22±0.99	67.96±0.54	72.16±1.20
ShearY	48.09±4.96	62.13±1.98	70.98±0.05	76.75±0.96
Solarize	48.72±1.55	63.58±1.12	70.14±0.73	73.94±0.02
SolarizeAdd	52.48±2.06	64.95±0.34	69.17±0.20	71.91±0.49
TranslateX	41.41±1.39	54.97±0.94	64.88±0.39	70.03±0.68
TranslateY	55.12±1.12	68.23±0.53	73.44±0.02	76.98±0.48

see a consistent performance boost when using both kinds of augmentations over Entropy (Δ 1.03), LC (Δ 0.78), Margin (Δ 0.94), Coreset (Δ 1.58), and BADGE (Δ 1.93). In conclusion, an integration of both single-image augmentations and image-mixing better unleash the potential information of each sample than they separately do.

A more detailed exploration of the virtual loss term. Since single-image augmentations and image-mixing have different impacts over CAMPAL, we also examine whether it is necessary to assign different weights for them within our virtual loss term, which we formulate as follows:

$$\begin{aligned} s_l &= \arg\min_s \frac{1}{|\mathcal{D}_L|} \sum\nolimits_{x_L \in \mathcal{D}_L} \mathcal{L}_f(x_L, s), \\ \text{where } \mathcal{L}_f(x,s) &= \mathcal{L}(x) + \lambda_1 \,\text{JS}\left(\{\mathcal{P}(y \mid \tilde{x}; f_\theta) \mid \tilde{x} \in \mathcal{T}^{(s)}_{single}(x)\}\right) + \lambda_2 \,\text{JS}\left(\{\mathcal{P}(y \mid \tilde{x}; f_\theta) \mid \tilde{x} \in \mathcal{T}^{(s)}_{mix}(x)\}\right) \end{aligned} \tag{9}$$

To optimize m_l, i.e. the strength for augmentations performed over labeled samples, we use λ_1, λ_2 to trade off the impact of single-image augmentations and image mixing. We dive deeper into this scheme by applying different combinations of λ_1, λ_2, shown in Table 13. Specifically, the experiment is conducted on the following versions: 1) CAMPAL$^{\text{MIN}}_{\text{Entropy}}$; 2) CAMPAL$^{\text{MIN}}_{\text{LC}}$; 3) CAMPAL$^{\text{MIN}}_{\text{Margin}}$; 4) CAMPAL$^{\text{STANDARD}}_{\text{Coreset}}$; 5) CAMPAL$^{\text{STANDARD}}_{\text{BADGE}}$. Since we constantly achieve superior results with $\lambda_1 = \lambda_2 = 1$, we may conclude that single-image augmentations and image-mixing are rough of the same impact over CAMPAL. Therefore, we do not distinguish between these two types of augmentations within our main context.

Table 12. Test accuracy of CAMPAL when integrated with different combinations of single-image augmentations and the MixUp.

Augmentations		Entropy	LC	Margin	Core-set	BADGE
Single	MixUp					
✓		75.87±0.32	76.04±0.41	77.61±0.91	76.62±0.50	77.78±0.45
	✓	75.41±0.25	74.89±0.64	77.76±0.48	75.21±0.32	75.69±0.62
✓	✓	**76.90±0.76**	**76.82±0.62**	**78.70±0.58**	**78.20±0.28**	**79.71±0.51**

Table 13. Test accuracy of CAMPAL when integrated with different combinations of λ_1, λ_2.

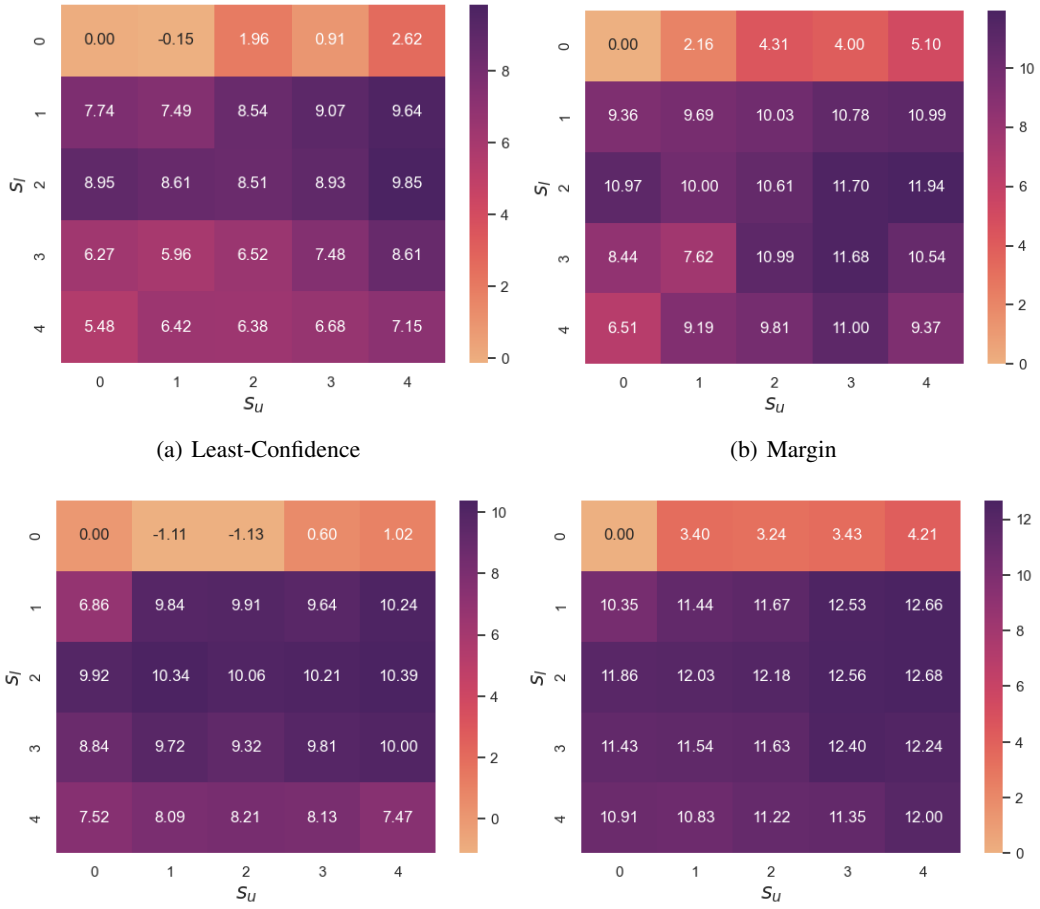
Coefficients		Entropy	LC	Margin	Core-set	BADGE
λ_1	λ_2					
1.0	0	75.87±0.32	73.36±2.25	70.15±2.19	76.62±0.50	77.78±0.45
1.0	0.5	75.62±0.76	75.89±0.73	77.20±0.52	79.23±0.51	77.07±0.38
0	1.0	75.41±0.25	72.93±0.65	67.69±0.72	73.99±0.03	75.69±0.62
0.5	1.0	75.95±0.62	76.55±0.62	76.25±0.60	75.21±0.32	74.88±0.14
1.0	1.0	76.90±0.76	76.82±0.62	78.70±0.58	78.20±0.28	79.71±0.51

B.8. Further Extension: Augmentations vs. Unsupervised Training

Recall that several studies tried to involve unlabeled samples in training auxiliary networks to assist querying(Sinha et al., 2019; Zhang et al., 2020; Kim et al., 2021a; Caramalau et al., 2021), which inevitably brings high computational costs. We claim that data augmentations are sufficient to enforce the acquisition process without much extra cost over unsupervised training. To verify this, we compare the running time and performance of augmentation-based strategies and those utilizing extra unsupervised architectures, shown in Table 14. We can see that augmentation-based methods with the best performance consistently outperform other strategies when becoming computationally efficient. Since active learning usually faces the problem of heavy computational cost in acquisitions, data augmentation may serve as an effective tool for both boosting the speed and performance at once. More importantly, this thought restricts the training process merely over labeled data, thus reducing the need for numerous unlabeled data in AL and making AL paradigms more applicable. We also adopt augmentations for labeled samples for methods with unsupervised representations.

Table 14. Comparison of the averaged test accuracy and the run-time of a single AL cycle over CIFAR-10. The run-time is calculated as the ratio to Random Sampling. Bold indicates the best performance of different data scales within each category.

Method		$\mathcal{N}_L = 500$	$\mathcal{N}_L = 1,000$	$\mathcal{N}_L = 1,500$	$\mathcal{N}_L = 2,000$	Time
Random		38.54±2.28	49.77±3.08	58.61±2.75	61.49±2.06	1
Entropy		39.80±1.60	55.43±1.71	60.76±2.64	65.95±1.36	1.03
LC		38.50±1.10	53.83±2.71	59.74±2.12	66.97±1.87	1.01
Margin		40.03±2.49	54.22±2.47	62.61±1.91	66.76±1.77	1.07
Core-set		43.42±2.09	53.54±2.74	62.00±1.44	66.90±0.93	1.33
BADGE		44.18±2.09	55.97±1.57	62.40±2.15	67.03±0.62	1.28
Unsupervised Representation	TA-VAAL	61.72±0.47	66.67±0.92	70.53±0.50	74.41±0.70	5.82
	SRAAL	60.53±0.89	67.08±0.28	71.02±0.66	75.05±0.15	6.04
	CoreGCN	56.03±1.73	59.81±1.31	65.19±1.49	69.61±2.34	2.73
CAMPAL-based Augmentation	Entropy	62.78±1.33	69.34±1.35	71.84±1.35	76.90±0.76	5.13
	LC	61.89±0.80	69.06±1.00	73.49±0.92	76.82±0.62	5.08
	Margin	65.46±0.63	72.77±0.55	75.96±0.85	78.70±0.58	5.10
	Core-set	62.59±0.89	71.55±0.29	75.69±0.62	78.20±0.28	5.42
	BADGE	**66.40±1.01**	**73.48±0.42**	**77.38±0.53**	**79.71±0.51**	5.54



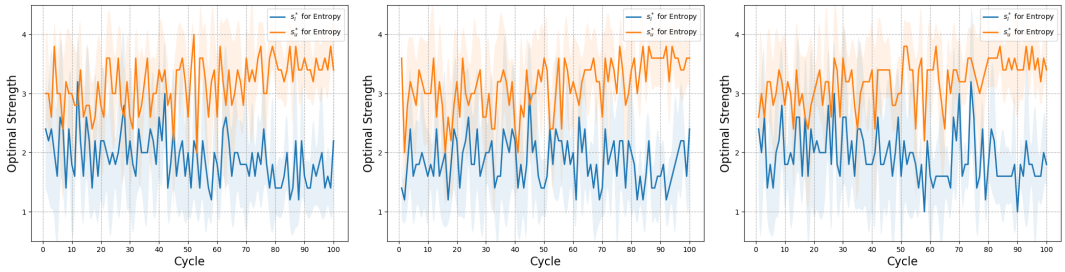
(a) Least-Confidence (b) Margin

(c) Core-set (d) BADGE

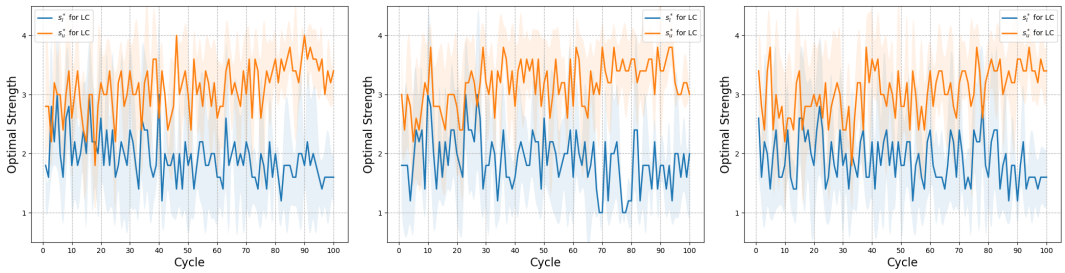
Figure 6. A heatmap visualization of performance boost brought by augmentations of different strengths, when attached to the labeled and unlabeled pool. The experiments are performed over CIFAR-10 with 2,000 labeled samples and are conducted over LC, Margin, Core-set and BADGE.

B.9. Addtional Results for Ablation Studies over Strengths

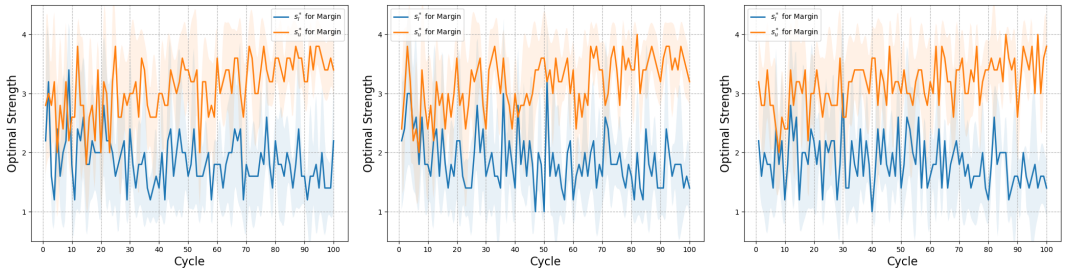
Compare with fixed augmentation strengths. Since we emphasize the importance of a strength control over s_l, s_u in Section 2.2, we will provide more details here. In brief, augmentations with various strengths contribute to the performance but can be inefficient when strengths are not chosen appropriately. To further look at the impact of augmentation sets with different strengths, we fix the value for s_l, s_u and see how they decide the final performance. Specifically, we test different combinations of s_l and s_u in the range $[0, 4]$, with other settings following the main empirical studies. The relative performance boost compared to their non-augmented counterparts is shown in Figure 6. Without proper strength control, the performance boost can decrease. For instance, $\text{CAMPAL}^{\text{MIN}}_{\text{Margin}}$ with $s_l = 3, s_u = 1$ leads to a 4.32% performance drop compared to the optimal one, when the worst case in $\text{CAMPAL}^{\text{MIN}}_{\text{Entropy}}$ causes a 3.28% drop. In addition, we can also see a trend similar to Section 3.2 that the classifier f_θ prefers weakly labeled augmentations when stronger unlabeled augmentations induce better acquisitions, even without a dynamic strength control.



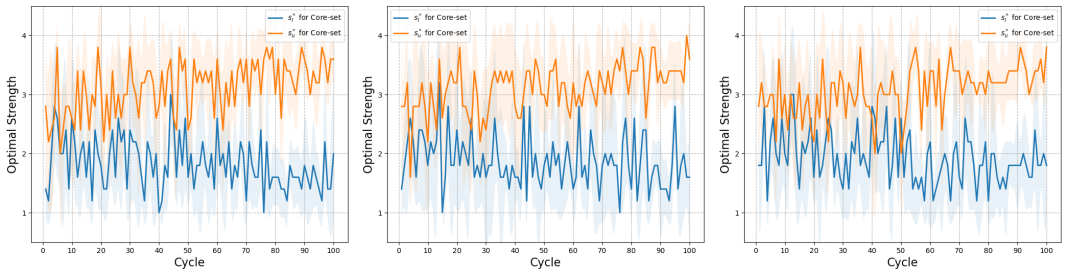
(a) $\text{CAMPAL}^{\text{MIN}}_{\text{Entropy}}$ (b) $\text{CAMPAL}^{\text{SUM}}_{\text{Entropy}}$ (c) $\text{CAMPAL}^{\text{DENISTY}}_{\text{Entropy}}$



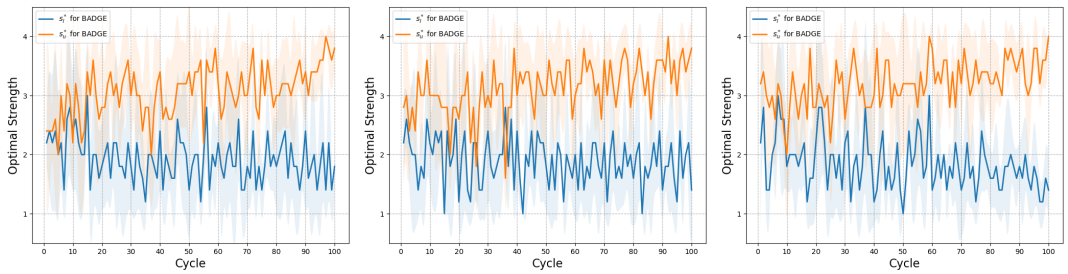
(d) $\text{CAMPAL}^{\text{MIN}}_{\text{LC}}$ (e) $\text{CAMPAL}^{\text{SUM}}_{\text{LC}}$ (f) $\text{CAMPAL}^{\text{DENISTY}}_{\text{LC}}$



(g) $\text{CAMPAL}^{\text{MIN}}_{\text{Margin}}$ (h) $\text{CAMPAL}^{\text{SUM}}_{\text{Margin}}$ (i) $\text{CAMPAL}^{\text{DENISTY}}_{\text{Margin}}$



(j) $\text{CAMPAL}^{\text{STANDARD}}_{\text{Coreset}}$ (k) $\text{CAMPAL}^{\text{CHAMFER}}_{\text{Coreset}}$ (l) $\text{CAMPAL}^{\text{HAUSDORFF}}_{\text{Coreset}}$



(m) $\text{CAMPAL}^{\text{STANDARD}}_{\text{BADGE}}$ (n) $\text{CAMPAL}^{\text{CHAMFER}}_{\text{BADGE}}$ (o) $\text{CAMPAL}^{\text{HAUSDORFF}}_{\text{BADGE}}$

Figure 7. The average optimal strength m_l, m_u across different AL cycles on CIFAR-10 with different instantiated versions for CAMPAL.