When Active Learning Meets Implicit Semantic Data Augmentation

Zhuangzhuang Chen[1], Jin Zhang[1], Pan Wang[1], Jie Chen[1], and Jianqiang Li[1] *

Shenzhen University, Shenzhen 518060 , China

Abstract. Active learning (AL) is a label-efficient technique for training deep models when only a limited labeled set is available and the manual annotation is expensive. Implicit semantic data augmentation (ISDA) effectively extends the limited amount of labeled samples and increases the diversity of labeled sets without introducing a noticeable extra computational cost. The scarcity of labeled instances and the huge annotation cost of unlabelled samples encourage us to ponder on the combination of AL and ISDA. A nature direction is a pipelined integration, which selects the unlabeled samples via acquisition function in AL for labeling and generates virtual samples by changing the selected samples to semantic transformation directions within ISDA. However, this pipelined combination would not guarantee the diversity of virtual samples. This paper proposes diversity-aware semantic transformation active learning, or DAST-AL framework, that looks ahead the effect of ISDA in the selection of unlabeled samples. Specifically, DAST-AL exploits expected partial model change maximization (EPMCM) to consider selected samples' potential contribution of the diversity to the labeled set by leveraging the semantic transformation within ISDA when selecting the unlabeled samples. After that, DAST-AL can confidently and efficiently augment the labeled set by implicitly generating more diverse samples. The empirical results on both image classification and semantic segmentation tasks show that the proposed DAST-AL can slightly outperform the state-of-the-art AL approaches. Under the same condition, the proposed method takes less than 3 minutes for the first cycle of active labeling while the existing agreement discrepancy selection incurs more than 40 minutes.

Keywords: Active learning, implicit semantic data augmentation, expected partial model change maximization, diversity

1 Introduction

In recent years, deep learning has achieved a new height in performing various tasks like image classification, object detection, semantic segmentation *etc.* However, they suffer from huge annotation labor and incur long time due to the

* Corresponding author

requirement of large-scale labeled data to train the deep models [41]. In some tasks, it is difficult to accumulate the data and it requires skilled professional to annotate. Therefore, the dependency on large-scale labeled data has become a major bottleneck for deep learning methods [12]. To alleviate this dependency, many methods like, unsupervised learning [6,11,27,43], semi-supervised learning [17,20,40,34], weakly supervised learning [24,28,29,44], active learning [1,2,9,41] *etc.*, have received significant attentions. Although weakly supervised learning and semi-supervised learning have made rapid progresses, active learning remains the foundation of many vision tasks due to its simplicity and better performance [39].

AL is an iterative process. It overcomes the limited labeling budget by selecting a set of samples from an unlabeled pool at each iteration [41] and labels the selected ones. These unlabeled samples will be added to the labeled set after being labeled by an oracle. Different unlabeled samples will yield different results. Therefore, the key question for AL is how to acquire data that can achieve better performance.

To solve this problem, many state-of-the-art works [13,19,41] achieve competitive results by designing the customized modules that make full use of the labeled samples and unlabeled samples. For example, in [32], the authors propose variational adversarial active learning (VAAL) that uses variational autoencoder and a discriminator to learn the uncertainty of the unlabeled samples implicitly. To better leverage both the annotation and the labeled/unlabeled samples' information, state-relabeling adversarial active learning (SRAAL) [41] designs a compact model composed of the unified representation generator and a labeled/unlabeled state discriminator. In [13], the authors propose agreement discrepancy selection (ADS) by designing adversarial classifiers to the convolutional neural network for the selection of informative samples. The above approaches achieve satisfying results by training the customized module with the labeled and unlabeled data in an adversarial manner, involving excessive training time. However, for robots with minimal computing resources and limited runtime, it is hard to get adopted in a new scenario in quick time without extensive training. Hence, enhancing the efficiency of the AL scheme without designing the customized modules remains a critical challenge.

The approaches based on the *expected model change principle*(EMCP) [4,5] address the efficiency challenge by querying the examples that maximally change the current model without designing the customized modules. Specifically, EMCP follows the stochastic gradient descent rule to estimate the ability of a candidate example to change the model by the gradient of the loss at the current candidate example [4,5]. Notably, most of these methods deal with the regression problem with the small model, and as a result, the existing EMCP methods are not directly applicable to classification tasks for the deep networks.

Besides the existing EMCP methods, ISDA [36] is another efficient approach for training the deep networks that provide diverse instances, which can be generated by changing the original instance to semantic transformation directions sampled from the feature covariance matrix. Next, it is straightforward to con-

sider the feasibility of combining EMCP and ISDA. One simple direction is a naive pipelined combination, which selects unlabeled samples by an acquisition function that follows the EMCP, and generates virtual instances from the selected samples by ISDA afterward. However, such an acquisition function fails to consider the potential gain from ISDA with respect to diversity. Hence, without any feedback during the acquisition process, the augmented samples from ISDA would not guarantee diversity.

To solve the above problem, the assumption in this paper is two-fold: (1) unlabeled samples have an unequal contribution to increasing the diversity of the label set after being labeled and augmented by ISDA, (2) the augmented samples with a higher diversity should have a higher ability to change the classifier. Based on these assumptions, this paper proposes the diversity-aware semantic transformation active learning, or DAST-AL framework. DAST-AL develops EPMCM to look ahead the effect of ISDA in advance of the acquisition process, by selecting unlabeled samples considering the ability of their augmented virtual samples to change the current classifier. The proposed EPMCM algorithm of DAST-AL enables us to select the unlabeled samples that have higher gain for increasing the diversity of labeled sets when augmented via ISDA. Furthermore, DAST-AL realizes the previously mentioned gain by labeling and augmenting the selected samples by using ISDA.

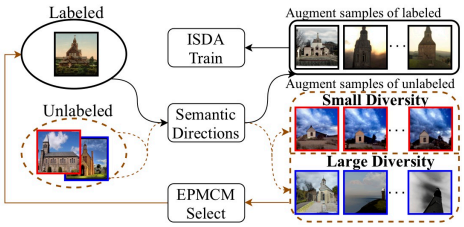


Fig. 1: An overview of DAST-AL. DAST-AL looks ahead the effect of ISDA in the process of acquisition while avoiding the costly sampling process. Note that the translated features are mapped to the image space and shown as augment samples in the above figure.

Our contributions are summarized as follows:

• Firstly, we propose EPMCM for selecting the unlabeled samples. The augmented samples of these unlabeled samples bring maximum change to the current partial model, and achieve a higher gain from ISDA in the assessment of diversity contributes to the labeled set.

• Then, we propose DAST-AL that overcomes the limited labeling budget by using the proposed EPMCM to efficiently select the unlabeled samples. After labeling and adding these unlabeled samples to the labeled set, DAST-AL efficiently augments the labeled set by using ISDA without the costly burden of explicitly generating the augmented samples.

• Finally, we compare the performance of our proposed method with existing methods in [13,30,32,37,41]. Although we find performances of ADS and DAST-AL are closely comparable, ADS takes 44 minutes in the first iteration of AL which is much higher compared to the proposed DAST-AL.

2 Related work

With a decade's study, AL has proven its superiority over the other methods. Based on the existing methods, we group them into two categories: parameterized sampler and non-parameterized sampler. The difference between the two categories lies in whether the customized modules are introduced for selecting the most informative samples.

The parameterized sampler approaches can further be decomposed into the synthesizing approaches and the pool-based approaches. The synthesizing approaches introduce the generative model to produce new synthetic samples that are informative to the current model [23,26,42]. These approaches introduce or design various generative adversial networks (GAN) [23] or variational autoencoders (VAE) [33] to enhance the diversity of labeled set by generating diverse data. The pool-based approaches use the customized modules to query the most informative instances from the unlabeled pool. A loss prediction module is designed to select data that is likely to make the target model producing a wrong prediction [37]. VAAL [32] and *task aware* VAAL [19] build a latent space by a VAE that learns together with the ranking conditional GAN. SRAAL[41] build an unsupervised image reconstructor and a supervised target learner to help relabel the state of unlabeled data with different importance. ADS [13] introduces a customized classifier to play the min-max game to select the most informative samples. By introducing the customized modules, these methods achieve the state-of-arts results. However, these approaches suffer from excessive training time during their iterative procedure for selecting the unlabeled samples.

The non-parameterized sampler approaches can also be decomposed into two parts: the uncertainty-based and ECMP-based approaches. The uncertainty-based approaches select the most informative samples by evaluating the uncertainty in the model prediction. To do so, previous works [18,22,35] simply utilize class posterior probabilities to define uncertainty. The probability of a predicted class [22] or an entropy of class posterior probabilities [18,35] defines uncertainty of a data point. To better quantify uncertainty, multiple forward passes with Monte Carlo Dropout are used [15]. However, it involves large computation for large-scale learning, as each data point in the large-scale unlabeled pool needs to be performed with multiple forward passes to measure its uncertainty. The EMCP approaches are based on the decision-theoretic. They select the unlabeled data by estimating expected model changes [4,5] that is based on the current model. These approaches have been well applied on regression tasks. As deep network involves a large number of parameters to estimate their changes, the approaches are hard to be applied to these networks.

Our method fits into the category of EMCP approaches with an exception. We estimate the partial model change for the classification tasks, rather than estimating the expected model change of the full network for regression. Also, our method is different from the synthetic approaches in the sense that we do not design any customized module and have no need to explicitly generate augmented images.

3 Method

In this section, we introduce the proposed DAST-AL. We first introduce how to obtain augmented instances in the feature space by following ISDA in Section 3.1. Then, we present details of the proposed EPMCM that served as the acquisition function in Section 3.2. In each iteration of DAST-AL, the following two steps are successively carried out.

1. Train the backbone network (feature extractor) and classifier on the current labeled set by using ISDA for augmenting this set to introduce more diversity. At the same time, we use the extracted features of the labeled samples at hand to calculate the covariance matrix for each category, where the covariance matrix represents all the feature semantic transformation directions of each category.
2. Use the proposed EPMCM to select the unlabeled sample. Then, the translated feature (i.e., the augmented samples) of these selected samples along infinite semantic transformation directions result in the maximum change to the current partial model. To overcome the limited labeling budget, these selected samples, which are expected to have higher diversity, will be labeled and added to the labeled set for step 1 in the next iteration.

3.1 Implicit data augmentation via semantic transformation

Let $\mathbb{M}^e$ be the feature extractor, $\mathbb{M}^c$ be the classifier, L be the labeled set, U be the unlabeled set, C be the number of classes, and y_l be the label of a sample x_l from L.

For a sample x_l in a class y_l, we extract its feature with $\boldsymbol{a}_l = \mathbb{M}^e(x_l)$. By following ISDA, the semantic directions for class y_l can then be obtained by sampling random vectors from a zero-mean multi-variate distribution $\mathcal{N}(0, \Sigma_{y_l})$, where Σ_{y_l} is the class-conditional covariance matrix estimated from the features of the labeled samples in class y_l. As the semantic transformation directions of different categories are different, we use the online estimation algorithm [36] to get the covariance matrixes of all classes $\Sigma = \{\Sigma_1, \Sigma_2, \cdots, \Sigma_C\}$. Referring to ISDA, we can randomly sample along $\mathcal{N}(0, \Sigma_{y_l})$ to generate augmented features with different semantic transformations, i.e., $\tilde{\boldsymbol{a}}_l \sim \mathcal{N}(\boldsymbol{a}_l, \Sigma_{y_l})$.

Consequently, unlimited $\tilde{\boldsymbol{a}}_l$ can be generated to augment the labeled set for diversity by exploiting *expected* cross-entropy loss in ISDA [36]. As for the unlabeled samples, suppose an unlabeled sample x_u can be selected in the AL cycle and labeled by a human expert with the label y_u, then it will be added to L, and its augmented feature $\tilde{\boldsymbol{a}}_u$ can also be obtained from $\tilde{\boldsymbol{a}}_u \sim \mathcal{N}(\boldsymbol{a}_u, \Sigma_{y_u})$, where $\boldsymbol{a}_u = \mathbb{M}^e(x_u)$. The potential gain from ISDA in the assessment of diversity contributes to the labeled set is up to the diversity of the $\tilde{\boldsymbol{a}}_u$. Therefore, it is crucial to measure the diversity contributing to the labeled set under unlimited $\tilde{\boldsymbol{a}}_u$, so that we can confidently rank the unlabeled samples with the potential gain from ISDA and select the better ones.

3.2 The proposed expected partial model change maximization

For selecting samples from the unlabeled set by considering their augmented features, the key is to find a reasonable way to evaluate the diversity contribution of each augmented feature sampling from the multivariate normal distribution. Intuitively, if the augmented feature is useless for the classifier updating, then this feature also has no use to augment the labeled set for more diversity. Inspired by this intuition, we propose EPMCM which is expected to evaluate the diversity contribution of the unlabeled sample under the unlimited augmented features by considering the partial classifier change. A detailed description of EPMCM is given below.

Different from existing EMCP-based methods that deal with regression problems at the instance level [4], we consider training the classifier $\mathbb{M}^c$ under an augmented labeled set $\mathcal{D} = \left\{\left(\tilde{\boldsymbol{a}}_l^i, y_l\right)\right\}_{i=1}^n$ in the feature space with cross-entropy loss $\mathcal{L}$, where $\tilde{\boldsymbol{a}}_l^i$ is sampled from $\mathcal{N}\left(\boldsymbol{a}_l, \Sigma_{y_l}\right)$ with label y_l. Then, following the empirical risk minimization principle [4], $\mathbb{M}^c$ is trained by minimizing the empirical error on $\mathcal{D}$. The corresponding empirical error is shown as follows:

$$\hat{\epsilon}_D = \sum_{i=1}^{n} \mathcal{L}\left[\mathbb{M}^c\left(\boldsymbol{a}_l^i\right), y_l\right]. \tag{1}$$

Suppose that $\mathbb{M}^c$ is defined by the last fully connected layer, and its parameters consist of the weight matrix $\boldsymbol{W} = \left[\boldsymbol{w}_1, \ldots, \boldsymbol{w}_C\right]^T \in \mathcal{R}^{C\times F}$ and biases $\boldsymbol{b} = \left[b_1, \ldots, b_C\right]^T \in \mathcal{R}^C$. For learning the best $\boldsymbol{W}$ and $\boldsymbol{b}$, stochastic gradient descent (SGD) [3] can be used to update the parameters iteratively according to the negative gradient of the loss $\mathcal{L}$ with respect to each augment features $\boldsymbol{a}_l^i$ that follows Eq. 2,

$$\{\boldsymbol{W}, \boldsymbol{b}\}_{\text{new}} \leftarrow \{\boldsymbol{W}, \boldsymbol{b}\} - \alpha \frac{\partial \mathcal{L}_{\boldsymbol{a}_l^i}(\{\boldsymbol{W}, \boldsymbol{b}\})}{\partial\{\boldsymbol{W}, \boldsymbol{b}\}}, \quad i = 1, \ldots, n, \tag{2}$$

where the α denotes learning rate.

With the understanding from the discussion above, now we describe the SGD rule in our AL scheme. Let us suppose the augmented feature $\tilde{\boldsymbol{a}}_u$ of unlabeled sample x_u with label y_u is added to the $\mathcal{D}$. The empirical error on the extended $\mathcal{D}^+ = \mathcal{D} \cup (\tilde{\boldsymbol{a}}_u, y_u)$ can be represented in the form of the following equation:

$$\hat{\epsilon}_{\mathcal{D}^+} = \sum_{i=1}^{n} \mathcal{L}\left[\mathbb{M}^c\left(\boldsymbol{a}_l^i\right), y_l\right] + \underbrace{\mathcal{L}\left[\mathbb{M}^c\left(\tilde{\boldsymbol{a}}_u\right), y_u\right]}_{:=\mathcal{L}_{\tilde{\boldsymbol{a}}_u}(\{\boldsymbol{W}, \boldsymbol{b}\})}. \tag{3}$$

As a consequence of the change in the augmented unlabeled set, the parameters $\boldsymbol{W}$ and $\boldsymbol{b}$ are also get changed. Considering the SGD update rule, as the model change is equivalent to parameters change, the parameters change $\mathbf{C}_{\{\boldsymbol{W}, \boldsymbol{b}\}}\left(\tilde{\boldsymbol{a}}_u\right)$ can be approximated as the gradient of the $\mathcal{L}$ at the $\tilde{\boldsymbol{a}}_u$, described by Eq. 4,

$$\mathbf{C}_{\{\boldsymbol{W}, \boldsymbol{b}\}}\left(\tilde{\boldsymbol{a}}_u\right) = \Delta\{\boldsymbol{W}, \boldsymbol{b}\} \approx \alpha \frac{\partial \mathcal{L}_{\tilde{\boldsymbol{a}}_u}(\{\boldsymbol{W}, \boldsymbol{b}\})}{\partial\{\boldsymbol{W}, \boldsymbol{b}\}}. \tag{4}$$

It should be noted that the dimension of $\boldsymbol{W}$ is $C \times F$ that will lead to a large computational burden with the increasing dimension of the feature. Hence, we only consider estimating the change of the $\boldsymbol{b}$, and then the partial classifier change can be represented as $\mathbf{C}_{\{\boldsymbol{b}\}}(\tilde{\boldsymbol{a}}_u)$. Since the goal of our AL acquisition function is to select the unlabeled sample, the augmented features of which lead to the maximum partial classifier change, we firstly consider an easy implementation that explicitly samples M times from the distribution $\mathcal{N}(\boldsymbol{a}_u, \Sigma_{y_u})$ to compose an limited augmented feature set $\mathcal{D}_{x_u} = \left\{\left(\tilde{\boldsymbol{a}}_u^1, y_u\right), \left(\tilde{\boldsymbol{a}}_u^2, y_u\right), \ldots, \left(\tilde{\boldsymbol{a}}_u^M, y_u\right)\right\}$ of size M. Here $\tilde{\boldsymbol{a}}_u^k$ denotes k^{th} sampled augmented features for the unlabeled sample x_u. Then, the potential of each x_u to augment the labeled set can be represented by summing the partial classifier change caused by each augmented feature in $\mathcal{D}_{x_u}$. Consequently, the acquisition function in our AL scheme can be formulated as follows:

$$x_u^* = \arg\max_{x_u \in U} \sum_{k=1}^{M} \left\| \mathbf{C}_{\boldsymbol{b}}\left(\tilde{\boldsymbol{a}}_u^k\right) \right\|, \tilde{\boldsymbol{a}}_u^k \in \mathcal{D}_{x_u}, \tag{5}$$

where x_u^* denotes the selected unlabeled samples.

To calculate Eq. 5, the following two issues must be tackled. **Firstly**, the true label y_u is unknown before querying. Therefore, calculation over all possible labels y_u can be a costly affair for an increasing number of classes. **Secondly**, when considering sampling M times, the sampling variance is unstable and limited. An ideal way is to generate as much data as possible (i.e., set M as large as possible). However the increasing M will incur an excessive time complexity.

To address the first issue, we use a maximum *a-posteriori* approximation by only considering the most likely label $\hat{y}_u$ of $\boldsymbol{a}_u$, predicted by the current $\mathbb{M}^c$. As for the second issue, aiming at simplifying computation while generating more data, we try to implicitly generate the unlimited augmented samples. When M grows to infinity, it is equivalent to considering the expectation of the Eq. 5 under all possible augmented features. Then, the Eq. 5 under the cross-entropy loss can be rewritten as:

$$\begin{aligned}
x_u^* &= \arg\max_{x_u \in U} \left(\mathbb{E}_{\tilde{\boldsymbol{a}}_u \sim \mathcal{N}(\boldsymbol{a}_u, \Sigma_{\hat{y}_u})} \left\| \mathbf{C}_{\boldsymbol{b}}(\tilde{\boldsymbol{a}}_u) \right\| \right) \\
&= \arg\max_{x_u \in U} \left(\mathbb{E}_{\tilde{\boldsymbol{a}}_u \sim \mathcal{N}(\boldsymbol{a}_u, \Sigma_{\hat{y}_u})} \left\| \frac{\partial \mathcal{L}_{\tilde{\boldsymbol{a}}_u}(\boldsymbol{b})}{\partial \boldsymbol{z}} \cdot \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{b}} \right\| \right) \\
&= \arg\max_{x_u \in U} \left(\mathbb{E}_{\tilde{\boldsymbol{a}}_u \sim \mathcal{N}(\boldsymbol{a}_u, \Sigma_{\hat{y}_u})} \left[1 - \frac{e^{\boldsymbol{w}_{\hat{y}_u}^T \tilde{\boldsymbol{a}}_u + \boldsymbol{b}_{\hat{y}_u}}}{\sum_{j=1}^{C} e^{\boldsymbol{w}_j^T \tilde{\boldsymbol{a}}_u + \boldsymbol{b}_j}} \right] \right) \\
&= \arg\max_{x_u \in U} \left(\mathcal{E}_{x_u}^{\infty} \right),
\end{aligned} \tag{6}$$

where $\boldsymbol{z}$ is the output of the $\tilde{\boldsymbol{a}}_u$ by using $\mathbb{M}^c$. Specifically speaking, $\boldsymbol{z}$ can be obtained by the formula $\boldsymbol{z} = \boldsymbol{w}^T \tilde{\boldsymbol{a}}_u + \boldsymbol{b}$, and then we have $\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{b}} = \mathbf{1}$, where $\mathbf{1}$ represents a vector of dimension C with each value 1. By assuming that the

learning rate α is identical for each augmented feature, the third equation is then obtained by unfolding the second equation with the expanded cross-entropy loss.

However, it is infeasible to compute $\mathcal{E}_{x_u}^{\infty}$ precisely, we derive an upper bound $\overline{\mathcal{E}_{x_u}^{\infty}}$ as an alternative to that. Although the maximum upper bound is not strictly guaranteed to be numerically maximum, in Sec. 4.1, we demonstrate that the selection of unlabeled samples by the upper bounds is effective. The reason for this effect is that the gap between the upper-bound and $\mathcal{E}_{x_u}^{\infty}$ will decrease with the increase of sampling times (shown in Tab. 2). Since we consider the infinite sampling times, the large upper-bound could indicate that $\mathcal{E}_{x_u}^{\infty}$ would be large too. Moreover, to prove that the selected unlabeled samples are able to augment the labeled set for more diversity, we visualize the augmented features of the selected unlabeled samples (ref. to Sec. 4.3). The corresponding $\overline{\mathcal{E}_{x_u}^{\infty}}$ is derived in the following manner:

$$
\begin{aligned}
\mathcal{E}_{x_u}^{\infty} &= \mathbb{E}_{\tilde{\boldsymbol{a}}_u \sim \mathcal{N}(\boldsymbol{a}_u, \Sigma_{\hat{y}_u})} \left[1 - \frac{e^{\boldsymbol{w}_{\hat{y}_u}^T \tilde{\boldsymbol{a}}_u + \boldsymbol{b}_{\hat{y}_u}}}{\sum_{j=1}^{C} e^{\boldsymbol{w}_j^T \tilde{\boldsymbol{a}}_u + \boldsymbol{b}_j}} \right] \\
&\leq \mathbb{E}_{\tilde{\boldsymbol{a}}_u \sim \mathcal{N}(\boldsymbol{a}_u, \Sigma_{\hat{y}_u})} \left[\frac{\sum_{j=1}^{C} e^{\boldsymbol{w}_j^T \tilde{\boldsymbol{a}}_u + \boldsymbol{b}_j}}{e^{\boldsymbol{w}_{\hat{y}_u}^T \tilde{\boldsymbol{a}}_u + \boldsymbol{b}_{\hat{y}_u}}} \right] - 1 \\
&= \mathbb{E}_{\tilde{\boldsymbol{a}}_u \sim \mathcal{N}(\boldsymbol{a}_u, \Sigma_{\hat{y}_u})} \sum_{j=1}^{C} \left(e^{\left(\boldsymbol{w}_j^T - \boldsymbol{w}_{\hat{y}_u}^T\right) \tilde{\boldsymbol{a}}_u + \boldsymbol{b}_j - \boldsymbol{b}_{\hat{y}_u}} \right) - 1 \\
&= \sum_{j=1}^{C} e^{\left(\boldsymbol{w}_j^T - \boldsymbol{w}_{\hat{y}_u}^T\right) \boldsymbol{a}_u + (\boldsymbol{b}_j - \boldsymbol{b}_{\hat{y}_u}) + \frac{1}{2} \left(\boldsymbol{w}_j^T - \boldsymbol{w}_{\hat{y}_u}^T\right) \Sigma_{\hat{y}_u} (\boldsymbol{w}_j - \boldsymbol{w}_{\hat{y}_u})} - 1 \\
&= \overline{\mathcal{E}_{x_u}^{\infty}}.
\end{aligned}
\tag{7}
$$

where the second inequality is hold by obeying the Jensen's inequality $2 \leq x + \frac{1}{x} \iff 1 - x \leq \frac{1}{x} - 1$, where $x = \frac{e^{\boldsymbol{w}_{\hat{y}_u}^T \tilde{\boldsymbol{a}}_u + \boldsymbol{b}_{\hat{y}_u}}}{\sum_{j=1}^{C} e^{\boldsymbol{w}_j^T \tilde{\boldsymbol{a}}_u + \boldsymbol{b}_j}}, 0 \leq x \leq 1$. Because of $\tilde{\boldsymbol{a}}_u \sim \mathcal{N}(\boldsymbol{a}_u, \Sigma_{\hat{y}_u})$, we can obtain that $\left(\boldsymbol{w}_j^T - \boldsymbol{w}_{\hat{y}_u}^T\right) \tilde{\boldsymbol{a}}_u + \boldsymbol{b}_j - \boldsymbol{b}_{\hat{y}_u}$ is also a Gaussian random variable, i.e., $\left(\boldsymbol{w}_j^T - \boldsymbol{w}_{\hat{y}_u}^T\right) \tilde{\boldsymbol{a}}_u + (\boldsymbol{b}_j - \boldsymbol{b}_{\hat{y}_u}) \sim \mathcal{N}\left(\left(\boldsymbol{w}_j^T - \boldsymbol{w}_{\hat{y}_u}^T\right) \boldsymbol{a}_u + (\boldsymbol{b}_j - \boldsymbol{b}_{\hat{y}_u}), \left(\boldsymbol{w}_j^T - \boldsymbol{w}_{\hat{y}_u}^T\right) \Sigma_{\hat{y}_u} (\boldsymbol{w}_j - \boldsymbol{w}_{\hat{y}_u})\right)$. Then, the fourth equation can be obtained by leveraging the moment-generating function $\mathbb{E}\left[e^{tX}\right] = e^{t\mu + \frac{1}{2}\sigma^2 t^2}$, where $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$ [8]. Finally, by calculating the $\overline{\mathcal{E}_{x_u}^{\infty}}$, we can select the unlabeled samples more efficiently. Since we do not need to explicit sampling and design the customized modules, the proposed EPMCM can be smoothly integrated into the AL scheme without excessive training time.

4 Experiments

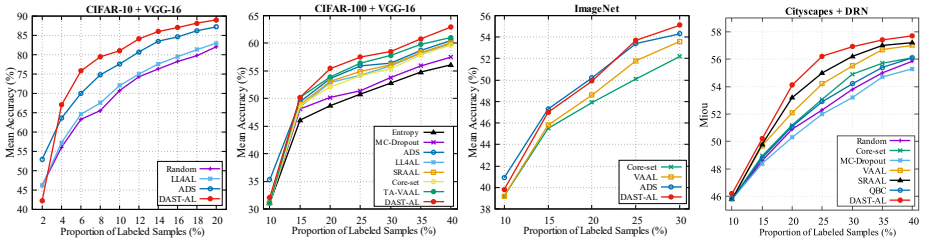
Under the scope of this part, we evaluate DAST-AL against state-of-the-art AL approaches on image classification in Sec. 4.1 and segmentation task in Sec. 4.2.

To further verify the efficiency of our method, we perform ablation study in Sec. 4.3 and time analysis in Sec. 4.4.

4.1 Active learning for image classification

Dataset. To verify our methods, we follow the same experimental settings proposed in [13,32,37,41] that fine-tune the network from the previous cycle if available. We choose commonly used CIFAR-10 and CIFAR-100 datasets for the image classification task. CIFAR-10 consists of 60000 images of $32 \times 32 \times 3$ pixels where 5000 images are used for the training and 1000 images are used for the testing. The CIFAR-10 and CIFAR-100 have 10 categories and 100 categories, respectively, while each category contains 600 images. We also follow the same setting in [32] that uses ImageNet [10] for the validation.
Compared methods. For image classification tasks, we evaluate our method against state-of-the-art AL approaches, including Core-set [30], LL4AL [37], VAAL [32], SRAAL [41], ADS [13]. We also use the random selection method as the baseline. It is important to note that these methods are evaluated by the same target model that consists of feature extractor and classifier.
Training settings. We use ResNet-18 [16] and VGG-16 [31] as the feature extractor in the target model to evaluate the accuracy. By following the experiment setting in [13], we initialize the labeled set L by randomly sampling 1000 data points from the whole unlabeled set U for the CIFAR-10 when using ResNet-18 or VGG-16, and randomly sampling 2500 data points for the CIFAR-100 when using ResNet-18. In the each iteration of AL, the number of labeled samples added to L for CIFAR-10 and CIFAR-100 are 1000 and 2500, respectively. We then re-train the target model. We adopt the same image normalization as reported in the experiment part [13], and the data augmentation strategies including 32×32 random image crop and horizontal flip. In each AL iteration, according to the previous work experiment setting [13], the training epoch is set to 200, mini-batch size is set to 128, the initial learning rate is set to 0.1 before 160 epochs and it decreases to 0.01 after 160 epochs on CIFAR-10. The momentum and weight decay are set to 0.9 and 0.0005, respectively. To obtain the mean and standard deviation of performance, each experiment is repeated three times.
Sub-set sampling. To make a fair comparison, we adopt the sub-set sampling from [13]. Since the entire training set is considered as the initial unlabeled set U, the sample size is very large, *e.g.,* 50,000 for CIFAR-10 and CIFAR-100. According to the study [25,30], it is less efficient to directly select top-k samples from the U, because of the information overlap among the samples [13]. To address this problem, we follow the same settings reported in [2,13] that first selects a random subset S_R and then selects top-k samples from S_R by different methods. Here, the sample size R is set to 10000 based on the study [13].

Performance on CIFAR-10 Fig. 2(a) shows the performances on the CIFAR-10 with the VGG-16 as the feature extractor. We can observe that, first, our



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

Fig. 2: Comparison of DAST-AL with Core-set [30], LL4AL [37], VAAL [32], SRAAL [41], ADS [13], and random selection method as a baseline: (a) on the CIFAR-10 using the VGG-16 as the target model, (b) on CIFAR-100 using the VGG-16 as the target model, (c) on ImageNet using the same target model in [32], (d) on cityscapes using the DRN as the target model.

DAST-AL achieves an accuracy close to 90% by using 20% of the labeled samples. The highest accuracy of the ResNet-18 with full dataset reaches 93.5% as reported in [41], and this is only 3.05% better than DAST-AL with 20% samples. Second, although the state-of-the-art ADS outperforms proposed DAST-AL when using 2% samples, The proposed DAST-AL can outperform ADS when using over 4% samples. The following two are the reasons: (1) At early iterations, ADS can improve feature representation by taking a long time (see the time analysis in Sec. 4.4) to train the customized classifier with a large number of unlabeled samples. (2) Our proposed EPMCM is able to consistently select the unlabeled sample, augmented samples of which have a large diversity contribution for the label set. Notably, the effect of the ADS would decay with the decrease of the unlabeled samples. Moreover, using ResNet-18 as the feature extractor, the proposed DAST-AL has the higher mean accuracy as shown in Tab. 1, demonstrating the robustness of proposed method comparing the others.

Performance on CIFAR-100 Although CIFAR-10 and CIFAR-100 have the same number of training images, CIFAR-100 has 100 categories while CIFAR-10 has 10 categories only. Hence, CIFAR-100 is much more challenging to tackle and needs larger proportions of training samples for achieving gratifying performance. As we can see from Fig. 2(b), at early iteration, DAST-AL outperforms all other methods, except ADS. Meanwhile, when using over 20% samples, our method is marginally better than ADS. The primary reason for the performance improvement of DAST-AL at the above iterations lies in the ability to select the unlabeled samples that can augment the challenging labeled set for more diversity. And, our DAST-AL outperforms the state-of-the-art SRAAL when using the same initial labeled samples. This indicates that DAST-AL can increase the diversity of the labeled samples by using ISDA. In addition, since DAST-AL does not use any unlabeled samples for training, our method does not suffer from the decrease in the number of the unlabeled samples.

4.2 Active learning for semantic segmentation

Dataset. Semantic segmentation tasks can be viewed as pixel-level classification tasks, which is more challenging than the image-level classification [41]. Here, we follow the experiment in [41], and choose the dataset Cityscapes [7] to evaluate DAST-AL against state-of-the-art AL approaches. The Cityscapes dataset consists of 3475 frames with instance segmentation annotations. To make a fair comparison, we also modify this dataset into 19 classes following the experiment in [41].

Compared methods. We evaluate our DAST-AL against a number of existing AL methods that reports performance on the semantic segmentation Cityscapes dataset. These methods contain Core-set [30], MC-Dropout [14], VAAL [32], QBC [21], and SRAAL [41]. As mentioned earlier, we introduce the random selection method as the baseline.

Training settings. Following the works in [32,41], the target model in our semantic segmentation experiment consists of the dilated residual networks (DRN) [38] as the feature extractor and a convolution layer as a classifier. Similar to the previous image classification setting, we initialize the labeled set L by randomly sampling 348 data points from the whole unlabeled dataset U. In the i^{th} iteration of AL, the number of labeled samples added to L is 150, and then re-train the target model. In addition, we only adopt the random horizontal flips as the data augmentation strategy in [32]. In each AL iteration, according to the previous work experiment setting [32,41], the training epoch is set to 50, mini-batch size is 8, the initial learning rate is set to 5×10^{-4}. To obtain the mean performance, each experiment is repeated 5 times with the same initial labeled pool.

Performance comparison. For the semantic segmentation task, following the setting in the SRAAL [41], we use the mean intersection over union, denoted as Miou to evaluate the performances of various methods. Since semantic segmentation tasks can be viewed as pixel-level classification tasks, we select the unlabeled samples by averaging $\overline{\mathcal{E}_{x_u}^{\infty}}$ from Eq. 7 of each pixel. In our experiments, we use the same initial labeled set and the same selection budget for different methods.

Fig. 2 (d) shows our result on various AL methods. We can observe that, first, SRAAL and VAAL obtain better performance than other methods, such as QBC, MC-Dropout, and core-set. This is because both VAAL and SRAAL take a long time to train the VAE module with a large number of unlabeled samples, and then they can select the most informative unlabeled samples. Second, our DAST-AL outperforms the SRAAL and with a large margin. It verifies that although the maximum upper bound in the Eq. 7 is not strictly guaranteed to be numerically maximum, DAST-AL still can effectively select the unlabeled samples, the augmented samples of which have large diversity shown in Eq. 7. Moreover, our method achieves high performance without being trained with the unlabeled samples and designing any extra modules.

4.3 Ablation study and discussion

To evaluate the effect of ISDA and EPMCM in DAST-AL, we conduct a series of ablation studies on CIFAR-10 with the ResNet18 as the feature extractor. As we can see from Tab. 1, by using ISDA and EPMCM, DAST-AL significantly boosts the performance at later iterations. By using ISDA only, the accuracy of Ran^+ increases close to 2% when compared with Ran under 10% labeled samples. The reason behind the performance improvement of Ran^+ is that ISDA can be used to increase the diversity of the labeled set, particularly when the labeled set is small. With the increase of the labeled sample selected by the random method, the accuracy of Ran^+ is less than the Ran when using 40% labeled data. This observation firstly verifies that the effect of ISDA is decayed with the random method, and then it proves that the proposed EPMCM can enhance the power of ISDA by selecting the unlabeled samples, augmented samples of which have a larger diversity contribution for the labeled set. For further verifying our remark, we conduct the visualization experiment in Sec. 4.3. Moreover, DAST-AL achieves better results compared to the Maxp^+. Notably, it illustrates that ISDA works better with EPMCM as ISDA and EPMCM share the same intuition that how to effectively increase the diversity for the labeled set.

Method	Accuracy (%) on Labeled Proportion (%)						
	5	10	15	20	25	30	35
Ran	67.13	80.06	85.25	87.14	89. 25	90.36	91.21
Maxp	67.13	76.37	79.88	81.40	82.84	83.46	84.60
Ran^+	69.27	80.76	85.60	87.82	89.30	90.67	91.15
Maxp^+	69.27	76.35	81.13	81.72	82.85	83.24	84.03
DAST-AL	**69.27**	82.98	**87.84**	**90.42**	**92.10**	**93.06**	**93.32**

Table 1: Comparison of ISDA and EPMCM in DAST-AL on CIFAR-10 under different proportion of labeled samples. Ran denotes DAST-AL random selects the unlabeled samples without ISDA, Maxp denotes DAST-AL select the unlabeled samples by its max prediction probability without ISDA. $(\cdot)^+$ represents the supervision under ISDA. It should be noted that Maxp^+ represent the simple pipeline that combines ISDA with existing AL methods.

The gap coming from the upper-bound term: We randomly select 1000 queried samples from CIFAR-10 and get their upper-bound. Meanwhile, we also explicitly generate the different number of augmented samples to compute an acquisition score of the queried sample. Then, we get the mean and std of the gap. The following table shows that the mean of the gap decrease with the increase of the sampling times. Hence, we are allowed to use upper-bound as we are considering the case of infinite augmented samples.

Sampling Times	100	1000	10000	100000
Gap (ResNet18)	0.85 ± 0.49	0.57 ± 0.34	0.35 ± 0.10	0.11 ± 0.07

Table 2: We compute he gap coming from the upper-bound term with ResNet18.

Visualization results To demonstrate that EPMCM can select the unlabeled samples, augmented features of which are able to increase diversity for the labeled set, we obtain the 'Augmented Images' by utilizing the reversing convolutional networks [36] to map the augmented features of the unlabeled samples back to the image space. Specifically speaking, since the ImageNet [10] has a high resolution, we compose a high resolution labeled set by randomly selecting 20000 labeled images from the ImageNet. In our case, 10000 images served as an unlabeled set. Then, we train the labeled set with ISDA to obtain the semantic directions, and select the images from the unlabeled set by random method and EPMCM. For the selected samples, their 'Augmented Images' of the corresponding augmented features are shown in right.

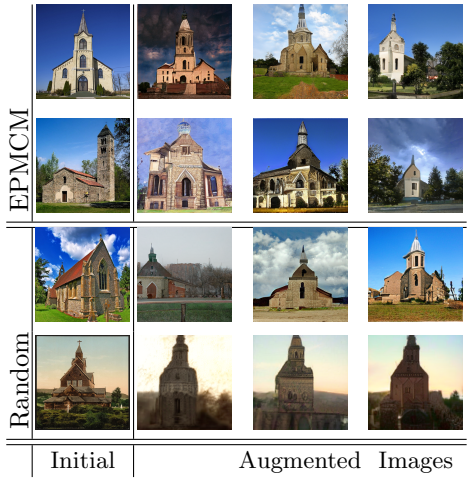


Table 3: Visualization of the semantically augmented feature of the selected samples from EPMCM and random. These 'Augmented Images' are generated with features sampled from feature distribution. It should be noted that these 'Augmented Images' are only for visualization.

In Tab. 3, the first two pictures from the first column and the last two from the same column represent the unlabeled samples selected by EPMCM and random, respectively. The 'Augmented Images' columns denote the images generated by the augmented features of the unlabeled samples. It can be observed that the unlabeled samples selected by EPMCM are more diverse.

4.4 Timing analysis

Tab. 4 shows the comparison results including DAST-AL and other methods on CIFAR-10. For a fair comparison, all of these methods are tested in the same torch version using the same NVIDIA TITAN Xp. Tab. 4 shows the extra params for the customized modules in these methods, the time needed to train for the first iteration in AL, sample a fixed budget of samples from the unlabeled set, and the total time. LL4AL only takes 2.06 minutes for one iteration in AL but it does not perform as well as DAST-AL considering its achieved mean accuracy. VAAL and SRAAL introduce the customized modules that involve

Method	EP (M)	TT (s)	ST (s)	ToT (m)
Core-set [30]	-	114.98	73.29	3.14
LL4AL [37]	0.2	120.72	2.97	2.06
VAAL [32]	88.18	62855	36.12	1048
SRAAL [41]	90.22	17897	41.67	298.9
ADS [13]	0.98	2688.6	4.83	44.91
DAST-AL	-	128.30	7.59	2.26

Table 4: Comparison of the extra params (EP) introduced by DAST-AL and other methods, the sampling time (ST) is taken to select the data from the unlabeled set on the CIFAR-10 dataset, the training time (TT) is taken to train the target model in AL. The total time (ToT) is composed by the training time and sampling time for one iteration in AL.

88.12 M and 90.22 M extra parameters, which need to be trained with the labeled and unlabeled samples in an adversarial manner. This explains why both the methods appear to be very slow in the training steps. ADS is the most competitive baseline to DAST-AL in terms of its achieved accuracy when using the small proportion of the labeled samples. However, DAST-AL takes 128.3 seconds while ADS requires 2688.6 seconds for the training in one iteration in AL. This can be explained by the fact that ADS introduces adversarial classifiers with 0.98 M extra parameters to minimize classifiers' prediction discrepancy and maximize prediction agreement with a large number of the unlabeled samples.

5 Conclusion and future work

In this paper, we propose a novel diversity-aware semantic transformation active learning method to overcome a limited labeling budget for achieving better performance. By looking ahead the effect of ISDA in the process of acquisition, we can select the unlabeled samples to augment the labeled set for more diversity with ISDA. Since we can not always guarantee that the expected partial model change to be numerically maximum, we will continue our research to calculate this change more accurately. In addition, as our method is able to construct a high-quality label set, we do believe our method can be a complement for existing works i.e., semi-supervised learning which employs unlabeled samples for training. By combining the proposed method with such a method, our method will have much potential for larger-scale datasets.

Acknowledgements. This work was supported by the National Nature Science Foundation of China under Grants U2013201, 62073225, 62072315, 61836005 and 62006157, the Natural Science Foundation of Guangdong Province-Outstanding Youth Program under Grant 2019B151502018, the Guangdong "Pearl River Talent Recruitment Program" under Grant 2019ZT08X603, the Guangdong "Pearl River Talent Plan" under Grant 2019JC01X235, and the Shenzhen Science and Technology Innovation Commission R2020A045.

References

1. Abraham, I., Murphey, T.D.: Active learning of dynamics for data-driven control using koopman operators. IEEE Transactions on Robotics **35**(5), 1071–1083 (2019)
2. Beluch, W.H., Genewein, T., Nürnberger, A., Köhler, J.M.: The power of ensembles for active learning in image classification. In: CVPR. pp. 9368–9377 (2018)
3. Bottou, L.: Stochastic gradient descent tricks. In: Neural networks: Tricks of the trade, pp. 421–436 (2012)
4. Cai, W., Zhang, M., Zhang, Y.: Batch mode active learning for regression with expected model change. IEEE transactions on neural networks and learning systems **28**(7), 1668–1681 (2016)
5. Cai, W., Zhang, Y., Zhou, J.: Maximizing expected model change for active learning in regression. In: EEE international conference on data mining. pp. 51–60 (2013)
6. Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: ECCV. pp. 132–149 (2018)
7. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR. pp. 3213–3223 (2016)
8. Curtiss, J.H.: A note on the theory of moment generating functions. The Annals of Mathematical Statistics **13**(4), 430–433 (1942)
9. Dasgupta, S., Hsu, D.: Hierarchical sampling for active learning. In: International Conference on Machine Learning. pp. 208–215 (2008)
10. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255 (2009)
11. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: CVPR. pp. 1422–1430 (2015)
12. Ebrahimi, S., Rohrbach, A., Darrell, T.: Gradient-free policy architecture search and adaptation. In: Conference on Robot Learning. pp. 505–514 (2017)
13. Fu, M., Yuan, T., Wan, F., Xu, S., Ye, Q.: Agreement-discrepancy-selection: Active learning with progressive distribution alignment. In: AAAI. pp. 7466–7473 (2021)
14. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: International Conference on Machine Learning. pp. 1050–1059 (2016)
15. Gal, Y., Islam, R., Ghahramani, Z.: Deep bayesian active learning with image data. In: International Conference on Machine Learning. pp. 1183–1192 (2017)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
17. Jiang, B., Zhang, Z., Lin, D., Tang, J., Luo, B.: Semi-supervised learning with graph learning-convolutional networks. In: CVPR. pp. 11313–11320 (2019)
18. Joshi, A.J., Porikli, F., Papanikolopoulos, N.: Multi-class active learning for image classification. In: CVPR. pp. 2372–2379 (2009)
19. Kim, K., Park, D., Kim, K.I., Chun, S.Y.: Task-aware variational adversarial active learning. In: CVPR. pp. 8166–8175 (2021)
20. Kingma, D.P., Mohamed, S., Rezende, D.J., Welling, M.: Semi-supervised learning with deep generative models. In: NeurIPS. pp. 3581–3589 (2014)
21. Kuo, W., Häne, C., Yuh, E., Mukherjee, P., Malik, J.: Cost-sensitive active learning for intracranial hemorrhage detection. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 715–723 (2018)

22. Lewis, D.D., Gale, W.A.: A sequential algorithm for training text classifiers. pp. 3–12 (1994)
23. Li, J., Chen, Z., Chen, J., Lin, Q.: Diversity-sensitive generative adversarial network for terrain mapping under limited human intervention. IEEE transactions on cybernetics (2020)
24. Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., Van Der Maaten, L.: Exploring the limits of weakly supervised pretraining. In: ECCV. pp. 181–196 (2018)
25. Making, M.O.D.: Synthesis lectures on artificial intelligence and machine learning (2012)
26. Mayer, C., Timofte, R.: Adversarial sampling for active learning. In: IEEE Winter Conference on Applications of Computer Vision. pp. 3071–3079 (2020)
27. Noroozi, M., Pirsiavash, H., Favaro, P.: Representation learning by learning to count. In: ICCV. pp. 5898–5906 (2017)
28. Peyre, J., Sivic, J., Laptev, I., Schmid, C.: Weakly-supervised learning of visual relations. In: CVPR. pp. 5179–5188 (2017)
29. Saito, S., Yang, J., Ma, Q., Black, M.J.: Scanimate: Weakly supervised learning of skinned clothed avatar networks. In: CVPR. pp. 2886–2897 (2021)
30. Sener, O., Savarese, S.: Active learning for convolutional neural networks: A core-set approach. In: ICLR (2018)
31. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. ICLR (2015)
32. Sinha, S., Ebrahimi, S., Darrell, T.: Variational adversarial active learning. In: ICCV. pp. 5972–5981 (2019)
33. Sohn, K., Lee, H., Yan, X.: Learning structured output representation using deep conditional generative models. NeurIPS **28**, 3483–3491 (2015)
34. Wang, D., Zhang, Y., Zhang, K., Wang, L.: Focalmix: Semi-supervised learning for 3d medical image detection. In: CVPR. pp. 3951–3960 (2020)
35. Wang, K., Zhang, D., Li, Y., Zhang, R., Lin, L.: Cost-effective active learning for deep image classification. IEEE TCSVT **27**(12), 2591–2600 (2016)
36. Wang, Y., Huang, G., Song, S., Pan, X., Xia, Y., Wu, C.: Regularizing deep networks with semantic data augmentation. IEEE TPAMI (2021)
37. Yoo, D., Kweon, I.S.: Learning loss for active learning. In: CVPR. pp. 93–102 (2019)
38. Yu, F., Koltun, V., Funkhouser, T.: Dilated residual networks. In: CVPR. pp. 472–480 (2017)
39. Yuan, T., Wan, F., Fu, M., Liu, J., Xu, S., Ji, X., Ye, Q.: Multiple instance active learning for object detection. In: CVPR. pp. 5330–5339 (2021)
40. Zhai, X., Oliver, A., Kolesnikov, A., Beyer, L.: S4l: Self-supervised semi-supervised learning. In: ICCV. pp. 1476–1485 (2019)
41. Zhang, B., Li, L., Yang, S., Wang, S., Zha, Z.J., Huang, Q.: State-relabeling adversarial active learning. In: CVPR. pp. 8756–8765 (2020)
42. Zhu, J.J., Bento, J.: Generative adversarial active learning. arXiv preprint (2017)
43. Zhuang, C., Zhai, A.L., Yamins, D.: Local aggregation for unsupervised learning of visual embeddings. In: ICCV. pp. 6002–6012 (2019)
44. Zhukov, D., Alayrac, J.B., Cinbis, R.G., Fouhey, D., Laptev, I., Sivic, J.: Cross-task weakly supervised learning from instructional videos. In: CVPR. pp. 3537–3545 (2019)