Exploring Minimally Sufficient Representation in Active Learning through Label-Irrelevant Patch Augmentation

Zhiyu Xue¹ Yinlong Dai² Qi Lei²
¹UC Santa Barbara, ²New York University
zhiyuxue@ucsb.edu, {yd2032,ql518}@nyu.edu

Deep learning models, which require abundant labeled data for training, are expensive and time-consuming to implement, particularly in medical imaging. Active learning (AL) aims to maximize model performance with few labeled samples by gradually expanding and labeling a new training set. In this work, we intend to learn a "good" feature representation that is both sufficient and minimal, facilitating effective AL for medical image classification. This work proposes an efficient AL framework based on off-the-shelf self-supervised learning models, complemented by a label-irrelevant patch augmentation scheme. This scheme is designed to reduce redundancy in the learned features and mitigate overfitting in the progress of AL. Our framework offers efficiency to AL in terms of parameters, samples, and computational costs. The benefits of this approach are extensively validated across various medical image classification tasks employing different AL strategies.[1]

1. Introduction

Deep learning models typically require training with abundant labeled data. However, annotating medical images requires prior domain expertise and is both costly and time-consuming. A potential mitigation for this challenge is through active learning (AL). AL aims to optimize model performance using the smallest number of labeled samples possible by incrementally expanding and labeling the training set. By prioritizing labeling informative samples rather than random selections, AL significantly enhances sample efficiency.[1]

Recent advances in AL largely attribute to the development of modern deep learning models (See reference therein [2-8]). Given that large-scale deep learning models are even more sample-demanding, it is urgent to develop effective and efficient active learning strategies. These strategies are essential to minimize the opportunity cost of labeling redundant samples, a significant concern in medical image classification where human annotations are notably scarce and costly.

In this work, we argue that the key to successful AL is to learn a "good" feature representation $\phi: x \rightarrow \phi(x)$. Ideally, this representation should ensure that label $y$ is linearly separable in the representation space. Such representation is preferable, since the process of AL for linearly separable data is well-understood[10]. As depicted in Fig.[1] establishing such a good representation requires the following conditions: (a) the representation we learned should be sufficient to predict $y$. This means $\phi(x)$ does not lose essential features in $x$ that are relevant to $y$. Mathematically speaking, our objective is to ensure $P(y|x)$ in the classification task aligns with $P(y|\phi(x))$, therefore constraining the

[1] Source Codes: https://github.com/chrisyxue/DA4AL

First Conference on Parsimony and Learning (CPAL 2024).
Figure 2: To evaluate the quality of representations \( \phi(x) \) for off-the-shelf SSL models in downstream tasks, we employ k-nearest neighbors (KNN) and logistic regression (LR) to model \( P(y|\phi(x)) \) based on the off-the-shelf SSL models including ViT-B with checkpoints released by DINO [14], MoCo-V3 [15], and MAE [16], and ResNet50 with checkpoint released by SimCLR [17]. The accuracy of these classifiers serves as our metric for assessing the linear separability of the representations produced by these off-the-shelf SSL models.

Figure 3: We compared the representation quality of adapting or not adapting the off-the-shelf ViT-B (MAE) to the unlabeled downstream tasks of medical images by reconstruction-based pretext training [16], where ViT-B (MAE)-Adapted/ViT-B (MAE)-Origin indicates the results of adapted/unadapted off-the-shelf ViTs gaining from the checkpoints released by MAE [16], respectively. Results show that ViT-B (MAE)-Origin can most times gain better representations than ViT-B (MAE)-Adapted.

classifier from a function on \( x \) to a function on \( \phi(x) \) introduces no additional bias. However, merely fulfilling this condition does not guarantee tangible benefits. For instance, a naive choice of sufficient \( \phi \) such as identity mapping fails in feature reduction. Therefore another essential requirement is (b) the representation should be \textbf{minimal}, in the sense that it preserves only the crucial information necessary to predict \( y \). By constraining the classifier to take in \( \phi(x) \) instead of \( x \), assuming \( \phi(x) \) excludes redundant features of \( x \), the model becomes more sample-efficient and will generalize better. The less redundant information \( \phi(x) \) contains, given its sufficiency, the more sample-efficient it is to model the relationship between \( \phi(x) \) to \( y \) [10,11]. This concept will be especially beneficial in the few-shot regime, aligned with the purpose of AL. (c) In line with AL’s principles, this representation initially can only be \textbf{approximated by off-the-shelf models} that have been pretrained on large-scale tasks with self-supervised learning (SSL) strategies, and gradually be corrected/fine-tuned with the progress of labeling new training samples from downstream tasks. Integrating conditions (a)(b) and (c), as proven in the theoretical work [12,13] for SSL, an approximately minimally sufficient representation established from the large-scale SSL tasks ensures our targets to be linearly separable in the representation space for downstream tasks.

We investigate practical strategies to fulfill the above conditions and together propose an effective, sample- and computationally efficient AL framework.
First, as argued above, we employ off-the-shelf SSL models to gain approximately minimally sufficient representations for medical images. Specifically, we choose off-the-shelf ViTs (checkpoints as DINO [14], MoCo-V3 [15], and MAE [16]) as the backbone initially for AL.

Compared to other architectures for off-the-shelf SSL models like ResNets [18] (checkpoint as SimCLR [17]), off-the-shelf ViTs explicitly contain detailed information like object shapes and textures [14] which are essential for analyzing medical images; we also show in Fig. 2 that they have great potential to achieve linearly separable representations for medical images across different diseases and modalities. Besides, we surprisingly find that adapting off-the-shelf ViTs to the unlabeled downstream medical image tasks by pretext training can reduce but cannot improve the representation quality (Fig. 3). Therefore, utilizing off-the-shelf ViTs directly to AL on the downstream tasks of medical images is both an effective and efficient choice.

With the increasing labeling information in AL, we need to accordingly design effective algorithms to improve the feature representation (to gradually be closer to being minimally sufficient). This is challenging since labeled samples are very limited, and thus it is unrealistic to fine-tune the whole representation, with risks of distorting the originally good features and overfitting the little samples [19–21]. We resolve this problem by proposing label-irrelevant patch augmentations and by learning only a subsequent layer as an adapter on top of the fixed pre-trained representation. By investigating a diversity of label-irrelevant patches, we largely enrich the training set and ameliorate the overfitting issue in the early stages of AL. Unlike traditional data augmentations [22–24] that tend to modify the semantics in medical domains [25, 26] and cause additional errors due to misspecifications, our method utilizes the little labeling information to guarantee no semantics is changed in the augmented data, leading to more reliable and robust feature learning.

Based on our proposed framework above, our contributions can be concluded as follows:

- We design a parameter-, sample- and computationally efficient AL framework based on self-supervised pretrained ViTs, to initially gain nearly minimally sufficient representations. Unlike existing deep AL baselines [6, 27] that train deep models or even introduce additional discriminators [7, 28] in every data-selection round, our proposal only trains a light adapter, yielding simplified procedure with less computation and memory costs.
- As AL incorporates more labeled samples, we design a label-irrelevant patch augmentation scheme that preserves semantic information better than prior DAs. Together with our proposed framework, it gradually reduces redundant features and alleviates overfitting. Our DA scheme generally applies to different datasets and architectures and can potentially be extended to other learning tasks besides active learning.
- We extensively verified the improved performance on medical image classification tasks across various ViT architectures and AL strategies. Compared to existing widely-used AL paradigms, our proposed parameter-efficient AL framework can boost the overall performance of Few-shot AL by 5% – 7%. Based on this framework, our proposed label-irrelevant patch augmentation methods can generally surpass existing DA methods by 1% – 4%.

2. Related Work

Self-supervised Learning (SSL). SSL is used to derive feature representations from unlabeled samples [17, 29, 31]. Existing SSL tasks include predicting rotation angles [32], jigsaw puzzles [33], contrastive learning [34–36], and reconstruction-based training [16, 29, 30, 37, 38]. In our study, we broadly explored SSL and found the off-the-shelf ViTs to be superior (see Fig. 2), guiding our choice.

Data Augmentation (DA). DA is used to improve sample efficiency and mitigate overfitting [39 – 40]. Traditional methods used simple transformations or augmentations altering the labels [41 – 43]. Recent methods learn to combine existing strategies or add consistency regularization [23 – 44]. Most existing works concentrate on developing strategies to effectively combine these various transformations. AutoAug [23] uses a search algorithm to discover the optimal augmentation policies for the training dataset. RandAug [24] provides a simple and efficient strategy by randomly selecting
augmentation operations and magnitudes. However, none of them is designed to be label-irrelevant. Their potential alteration of the original semantic information has raised concerns [25, 45].

Active Learning (AL). The AL labeling strategies can be categorized into uncertainty-based and diversity-based methods. Uncertainty-based methods utilized the entropy [46], confidence [47], margin [48], and standard deviation [49] to measure informativeness. Diversity-based methods seek the most representative samples from the unlabeled dataset, such as Coreset [5] selected unlabeled samples that are furthest to their closest labeled samples, and Determinantal Point Process (DPP) [50] quantifies diversity based on a pairwise (dis)similarity matrix.

Prior works have used SSL for AL representation learning, including SSLAL [51], MoBYv2AL [52], and PT4AL [53]. These methods require SSL in each AL round, increasing time costs. We show that an off-the-shelf ViT already offers a sufficient representation, reducing the need for repeated SSL (Fig. 3). In integrating AL with DA, DAST-AL [54] and others [55-57] use various augmentation techniques. Yet, these often neglect potential semantic loss and aren’t tailored for finetuning off-the-shelf ViTs that can outperform CNNs [58] [59].

3. Methodology

3.1. Setup

For an off-the-shelf ViT as $f_{\text{enc}}$ (off-the-shelf ViT pretrained as MAE also equips a decoder as $f_{\text{dec}}$), we denote the patched input as $x \in \mathbb{R}^{U \times (P^2 - C)}$, where $P$ and $U$ is the width and number of patches.

For AL, in $k$-th round, we denote the labeled/unlabeled set as $D_k^{\text{lab}} / D_k^{\text{unl}}$, respectively. The acquisition function $\alpha(x, M_k)$ will output a value measuring the informativeness for $x$ in $D_k^{\text{lab}}$ according to the trained predictive model $M_k$. Based on the outputted values, it can select a batch of $b$ samples in $D_k^{\text{unl}}$ as $B_k$, query their labels from human annotators and then move them from $D_k^{\text{unl}}$ to $D_k^{\text{lab}}$ with the queried labels. Such a process is illustrated as the blue lines in Fig. 4. As we introduced in section 2 existing AL strategies mainly focus on how to design acquisition function $\alpha(x, M_k)$ to select the most informative samples. For example, the least confidence such as $\max_{\hat{y}} p_{M_k}(\hat{y}|x)$ can be utilized to be the acquisition function, where $\hat{y}$ is the prediction with the highest probability.

3.2. Parameter-efficient AL on Off-the-shelf ViT

As illustrated in Fig. 4, in every AL round, we train an adapter $g$ on top of a pretrained and frozen ViT $f_{\text{enc}}$, where $g$ is designed to be a combination of a lightweight MLP $g_{\text{enc}}$ and a linear classification head $g_{\text{cls}}$. Within the $k$-th round, only $g$ is optimized via $D_k^{\text{lab}}$ while $f_{\text{enc}}$ is kept freezing, much more efficient than existing AL works that optimize the entire network [51-53] for each AL round. The AL acquisition function $\alpha(x, M_k)$ is then conducted based on the trained predictive model as $M_k(x) := g(f_{\text{enc}}(x))$. Here $g_{\text{enc}}$ is designed to gradually make the feature representations $\phi(x) := g_{\text{enc}}(f_{\text{enc}}(x))$ near minimally sufficient, and $g_{\text{cls}}$ aims to predict the labels based on $\phi(x)$.

3.3. Exploring the Diversity of Label-Irrelevant Patches

Based on our framework (Fig. 4), we further design a diverse set of data augmentations $A \in A$ that are label-irrelevant. When $A$ is label irrelevant, it means $y|x \equiv y|\phi(x) \equiv y|\phi(\hat{x})$ (equivalent in distribution), where $\hat{x} := A(x)$ and $A$ is randomly sampled from $A$. Therefore there is no bias when assigning the same label $y$ associated with $x$ to $\hat{x}$ and this ensures the representation $\phi$ to be sufficient. As we investigate a more diverse set of $A$, the dependence between $\phi(x)$ and $\phi(\hat{x})$ is reduced, and
the more sample-efficient the learning procedure will be. To see why this helps, one can think about the most ideal case where \( \phi(x) \perp \phi(\hat{x}) | y \), meaning their dependence is only through \( y \), and now the representation \( \phi \) is minimal and it only requires to learn \( y \) through \( \phi(x) \) with order \( |Y| \) samples, where \( |Y| \) is the number of classes [12].

The diverse set of label-irrelevant data augmentations \( A \) enforces \( \phi(x) := g_{\text{enc}}(f_{\text{enc}}(x)) \) to be invariant to the redundant information in \( x \), producing features closer to the minimal sufficient representation. Such representation nicely integrates with the AL procedure and the relationship between \( x \) and \( y \) will be learned rapidly and concretely. However, existing heuristic DA methods such as RandAug [24] do not satisfy label-irrelevance, causing the learned feature representation \( \phi(x) \) not near minimally sufficient, which cannot learn the relation between \( x \) and \( y \) effectively during AL. We design a diverse set of label-irrelevant DA strategies by exploring the diversity of label-irrelevant patches. Given the patch-wise feature extraction approach taken by ViTs, enhancing the diversity of label-irrelevant patches allows a better approximation of minimally sufficient \( \phi(x) \). Our proposed DA consists of two steps: (1) Localize the label-irrelevant patches, and (2) Augment the label-irrelevant patches. The details of (1) and (2) will be introduced as follows.

### 3.3.1. Label-Irrelevant Patches Localization

We localize the label-irrelevant patches by computing a patch-wise correlation vector \( \text{Cor} \in \mathbb{R}^{|U|} \) between \( x \) and \( y \), where its \( i \)-th coordinate \( \text{Cor}_i \) indicates the correlation between the \( i \)-th patch and \( y \). Existing model explanation techniques such as DeepLIFT [60], Saliency [61], CosineAttention-Map [62], and LastAttentionMap [63], can generate a saliency map that illuminates the semantic regions of a raw image \( x \) regarding ground truth label \( y \). Hence, we compute \( \text{Cor} \) by segmenting the saliency map into patches, aggregating values at the patch level, and subsequently normalizing the result. To establish a criterion for identifying irrelevant patches of \( x \), we operationally select them by masking the lowest \( r \% \) (e.g. 75\%) patches by \( M_{\text{Cor}} \) as \( \bar{x} = M_{\text{Cor}} \odot x \), where \( M_{\text{Cor}} \) is a patch-wise 0-1 mask generated to mask the lowest \( r \) patches in \( x \) in regard to \( \text{Cor} \). More discussions for these localization methods are deferred to section 4.2 and appendix A.3.

### 3.3.2. Label-Irrelevant Patches Augmentation

For label-irrelevant DA as \( \hat{x} := A(x) \), we primarily emphasize two effective methods (SelfPatchAug and SubstitutivePatchAug) in this section to augment the label-irrelevant patches localized by \( M_{\text{Cor}} \). Both of these two DA methods satisfy label-irrelevance since they only explore the diversity of label-irrelevant patches while keeping label-relevant patches unchanged. Moreover, we present Instance-adaptive Label Smoothing (IaLS) alongside these two label-irrelevant DA methods to achieve better performance by alleviating the negative impact caused by patch augmentation.
With checkpoints released by MAE due to the requirement of pretrained decoder, we propose an instance-adaptive label smoothing (IaLS) strategy for the augmented image to be more minimally sufficient, leading to a classifier $g_{cls}$ with better generalizability. However, the limitation of SelfPatchAug is that it can only be applied to the off-the-shelf ViTs with checkpoints released by MAE due to the requirement of pretrained decoder $f_{dec}$. Differently, SubstitutivePatchAug is compatible with off-the-shelf ViTs pretrained with various SSL strategies (e.g., DINO, MoCo-V3), which will be introduced as follows.

**SubstitutivePatchAug.** SubstitutivePatchAug is inspired by existing DA methods for text data [64, 65]. In the field of text classification, identifying suitable word substitutions and replacing the original words with substitutions is a kind of prevalent DA strategy [66–68]. The structure of Transformers is originally designed for NLP tasks [69], and ViTs treat an image as a sequence of non-overlapping patches, just like how Transformers handle tokens in a sentence. Therefore, the intuition of SubstitutivePatchAug is that we take the scope of ViTs and consider patches as words. Thus this method augments the label-irrelevant patches by substituting them with semantically related patches from $t$ similar images from a query set $Q$ selected from $D = D_{k}^{unl} \cup D_{k}^{lab}$.

As shown in Fig. 5b, we compute a pre-defined similarity matrix $\Psi \in \mathbb{R}^{N \times N}$ among $D$, where $\Psi(i, j)$ indicates the similarity between $f_{enc}(x_i)$ and $f_{enc}(x_j)$, and $N$ is the number of samples in $D$. With $x$ as a key, we search the top-$t$ samples as a query set $Q = \{(x_i, y_i)\}_{i=1}^{t}$ according to $\Psi$. In Fig. 5b, by leveraging query set $Q$ and key $x$, we construct a feature-level/raw-level patch similarity matrix $\Phi_{rep}/\Phi_{raw} \in \mathbb{R}^{U \times (t \times U)}$ based on $f_{enc}(x)/x$, referring to the substitutive patches defined from the features space and raw-data space, respectively. Subsequently, we use a trade-off factor $\lambda \in [0, 1]$ for linear combination as $\Phi = \lambda \Phi_{raw} + (1 - \lambda) \Phi_{rep}$. Such linear combination aims to select substitutive patches that can balance between the similarities in feature space and raw-data space. For every patch in $x$, we can select the most similar patch from $Q$ by $\Phi$, hence produce the augmented image $\hat{x} = A(x)$ by filling in the masked patches of $\bar{x}$ with the selected substitutive patches.

**Instance-adaptive Label Smoothing (IaLS) for Augmented Data**

Inspired by existing works [70, 71] that utilize label smoothing to alleviate the degradation of semantic information caused by DA, we propose an instance-adaptive label smoothing (IaLS) strategy for the augmented image $\hat{x}$. The primary motivation behind IaLS is that, since label-irrelevant patches are selected by Cor with a hard threshold ratio of $r_{\%}$, reconstructing (SelfPatchAug) or replacing (SubstitutivePatchAug) these patches risks of losing some label-relevant information. In other words, $\phi(x) \perp \phi(\hat{x}) | y$ is merely an idealized condition, and cannot be guaranteed with absolute certainty in practical applications, particularly when $(1 - M_{Cor})^T Cor$ is large.

To address this issue, we introduce an instance-adaptive factor $\beta_x$ to reduce the confidence of the model’s prediction for the augmented image $\hat{x}$. Specifically, the smoothed label vector of the augmented sample $\hat{x}$ is computed as $\hat{y} = \beta_x \frac{1}{M_{Cor}} + (1 - \beta_x) \cdot y$. For each image $x$ with its corresponding $M_{Cor}$ and Cor, we smooth its ground truth vector by an instance-adaptive factor $\beta_x = (1 - M_{Cor})^T Cor$.

### 4. Experiments

#### 4.1. Experimental Setups

**Datasets.** We evaluate our methods on MedMNIST [73], an ensemble evaluation benchmark encompassing various classification tasks for medical imaging. MedMNIST is a widely adopted collection of standardized biomedical image datasets designed for image classification tasks. We conduct our methods on DermaMNIST, BloodMNIST, PneumoniaMNIST, OrganAMNIST, OrganCMNIST, and OrganSMNIST, respectively.

**AL Settings.** We conduct our experiments with various AL strategies to demonstrate that our method can achieve promising performance consistently across different AL strategies. Our work investigates
two AL paradigms: Few-shot AL ($N_{lab}^0 = 10$, $K = 50$, $b = 5$) and Many-shot AL ($N_{lab}^0 = 1000$, $K = 5$, $b = 500$). Notably, Few-shot AL poses a sterner challenge due to its greater potential of overfitting, induced by the paucity of labeled data especially at the early stage of AL. Besides, Few-shot AL is more practical in performing medical image classification due to the high labeling cost. Due to the page limitation, all the results posted in tables and figures in section 4.2 are produced under the least confidence strategy. Mean here denotes the averaged values of results among 6 different datasets.

![Figure 6: The AUBC results of comparing existing AL model structures (ResNet) and our efficient model structures (ViT-B + Adapter) via AL strategies as Least Confidence and Margin. Mean here denotes the averaged values of results among 6 different datasets.](image)

**Table 1:** Results of the comparison between our proposed label-irrelevant patch augmentation methods and other DA methods across 6 datasets for medical image classification by utilizing Least Confidence and Margin as the AL strategy. We conduct those DA methods on our efficient AL framework via different off-the-shelf ViTs. Note that Mean represents the averaged AUBC across 6 datasets, while Avg Rank is computed by ranking the AUBC performance on each dataset and then taking the average.

<table>
<thead>
<tr>
<th></th>
<th>DermaMNIST</th>
<th>BloodMNIST</th>
<th>PneumoniaMNIST</th>
<th>OrganAMNIST</th>
<th>OrganCMNIST</th>
<th>OrganSMNIST</th>
<th>Mean</th>
<th>Avg Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Least Confidence</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ViT-B (MAE)</td>
<td>0.6113</td>
<td>0.5900</td>
<td>0.5930</td>
<td>0.6740</td>
<td>0.6420</td>
<td>0.5140</td>
<td>0.6952</td>
<td>4.17</td>
</tr>
<tr>
<td>+ Rand/Aug</td>
<td>0.6410</td>
<td>0.5800</td>
<td>0.8410</td>
<td>0.6840</td>
<td>0.6760</td>
<td>0.5280</td>
<td>0.6382</td>
<td>3.00</td>
</tr>
<tr>
<td>+ Auto/Aug</td>
<td>0.6140</td>
<td>0.6090</td>
<td>0.8160</td>
<td>0.7050</td>
<td>0.6410</td>
<td>0.5120</td>
<td>0.6540</td>
<td>3.83</td>
</tr>
<tr>
<td>+ Normal/Aug</td>
<td>0.6330</td>
<td>0.6220</td>
<td>0.8400</td>
<td>0.5440</td>
<td>0.5410</td>
<td>0.5110</td>
<td>0.6152</td>
<td>4.83</td>
</tr>
<tr>
<td>+ SubstitutivePatchAug</td>
<td>0.6700</td>
<td>0.6340</td>
<td>0.8280</td>
<td>0.6850</td>
<td>0.6710</td>
<td>0.5360</td>
<td>0.6707</td>
<td>2.33</td>
</tr>
<tr>
<td>+ SelfPatchAug</td>
<td>0.6920</td>
<td>0.6270</td>
<td>0.8330</td>
<td>0.6840</td>
<td>0.6390</td>
<td>0.5170</td>
<td>0.6653</td>
<td>2.67</td>
</tr>
<tr>
<td>ViT-B (DINO)</td>
<td>0.6510</td>
<td>0.6900</td>
<td>0.8420</td>
<td>0.6740</td>
<td>0.6220</td>
<td>0.5590</td>
<td>0.6735</td>
<td>3.83</td>
</tr>
<tr>
<td>+ Rand/Aug</td>
<td>0.6240</td>
<td>0.7090</td>
<td>0.8610</td>
<td>0.6790</td>
<td>0.6380</td>
<td>0.5430</td>
<td>0.6775</td>
<td>2.67</td>
</tr>
<tr>
<td>+ Auto/Aug</td>
<td>0.6960</td>
<td>0.6960</td>
<td>0.8490</td>
<td>0.6870</td>
<td>0.6140</td>
<td>0.5510</td>
<td>0.6777</td>
<td>2.83</td>
</tr>
<tr>
<td>+ Normal/Aug</td>
<td>0.6720</td>
<td>0.6750</td>
<td>0.8220</td>
<td>0.5560</td>
<td>0.5740</td>
<td>0.5460</td>
<td>0.6408</td>
<td>3.50</td>
</tr>
<tr>
<td>+ SubstitutivePatchAug</td>
<td>0.6640</td>
<td>0.6890</td>
<td>0.8370</td>
<td>0.6930</td>
<td>0.6480</td>
<td>0.5830</td>
<td>0.6840</td>
<td>2.17</td>
</tr>
</tbody>
</table>

**Margin**

<table>
<thead>
<tr>
<th></th>
<th>DermaMNIST</th>
<th>BloodMNIST</th>
<th>PneumoniaMNIST</th>
<th>OrganAMNIST</th>
<th>OrganCMNIST</th>
<th>OrganSMNIST</th>
<th>Mean</th>
<th>Avg Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>ViT-B (MAE)</td>
<td>0.7632</td>
<td>0.8620</td>
<td>0.8330</td>
<td>0.6690</td>
<td>0.6630</td>
<td>0.5410</td>
<td>0.6656</td>
<td>4.90</td>
</tr>
<tr>
<td>+ Rand/Aug</td>
<td>0.6270</td>
<td>0.6240</td>
<td>0.8410</td>
<td>0.7200</td>
<td>0.6980</td>
<td>0.5050</td>
<td>0.6845</td>
<td>2.50</td>
</tr>
<tr>
<td>+ Auto/Aug</td>
<td>0.6460</td>
<td>0.6170</td>
<td>0.8160</td>
<td>0.7190</td>
<td>0.6690</td>
<td>0.5400</td>
<td>0.6679</td>
<td>4.83</td>
</tr>
<tr>
<td>+ Normal/Aug</td>
<td>0.6427</td>
<td>0.6230</td>
<td>0.8400</td>
<td>0.6427</td>
<td>0.5940</td>
<td>0.5270</td>
<td>0.6449</td>
<td>5.00</td>
</tr>
<tr>
<td>+ SubstitutivePatchAug</td>
<td>0.6640</td>
<td>0.6410</td>
<td>0.8280</td>
<td>0.7240</td>
<td>0.7270</td>
<td>0.5850</td>
<td>0.6948</td>
<td>2.17</td>
</tr>
<tr>
<td>+ SelfPatchAug</td>
<td>0.6880</td>
<td>0.6470</td>
<td>0.8330</td>
<td>0.7220</td>
<td>0.7000</td>
<td>0.5650</td>
<td>0.6925</td>
<td>2.00</td>
</tr>
</tbody>
</table>

**Table 1:** Results of the comparison between our proposed label-irrelevant patch augmentation methods and other DA methods across 6 datasets for medical image classification by utilizing Least Confidence and Margin as the AL strategy. We conduct those DA methods on our efficient AL framework via different off-the-shelf ViTs. Note that Mean represents the averaged AUBC across 6 datasets, while Avg Rank is computed by ranking the AUBC performance on each dataset and then taking the average.

Evaluation Metrics. To evaluate the performance of AL, we report area under the budget curve (AUBC) \(^2\) in our experimental results, where the AUBC value is calculated by the trapezoidal
method over a given budget curve, and higher AUBC values indicate better overall performances for varying budgets during the AL process.

4.2. Experimental Results and Analysis

Effectiveness and Efficiency of Our AL Framework. As Fig. 6 illustrated, to show the superiority of our proposed efficient AL framework (Fig. 4), we employ ResNet18/34/50 as the baseline models for comparison across various AL strategies. We choose ResNet18/34/50 as baselines due to their varying degrees of parameter complexity and widespread use in the AL field. Results illustrated in Fig. 6 show that our proposed efficient AL framework on off-the-shelf ViTs can surpass the selected baselines via various AL methods across six datasets selected from MedMNIST. Comparing the results of off-the-shelf ViTs (MAE/MoCo-V3/DINO) between Fig. 6 and Fig. 2, it obviously supports our claim in section 1 that sufficient feature representations are the key to a successful AL, since the relative magnitude of AUBC shown in Fig. 6 is almost consistent with the relative magnitude of LR/KNN Accuracy shown in Fig. 2, demonstrating that sufficient representations lead the AL acquisition function to select promising informative samples. Besides, in Table 2, we provide a comparative analysis of the parameter size for several widely-used AL architectures, where our framework demonstrates a significant reduction in the number of parameters.

Effectiveness of SelfPatchAug & SubstitutivePatchAug. Based on our proposed efficient framework (Fig. 4), we further compare the performance of our label-irrelevant patch augmentation methods against existing DA methods, such as AutoAug and RandAug. Additionally, we design a practical DA strategy as a baseline named NormalAug by simply combining random horizontal and vertical flips.

Our label-irrelevant patch augmentations (SubstitutivePatchAug and SelfPatchAug), as shown in Table 1, markedly outperform existing DA methods in Few-shot AL across various medical datasets. This success, reflected in the Mean and Avg Rank metrics, underscores the significant and consistent enhancement our methods bring to our efficient AL framework. The Mean metric together with the Avg Rank offers a comprehensive evaluation of the DA methods. While our SubstitutivePatchAug and SelfPatchAug methods excel in most datasets, they falter in cases like PneumoniaMNIST where lesions pervade the entire image, making almost all patches label-relevant and risking the loss of semantic information through any patch augmentation. However, in datasets like DermaMNIST, where lesions are localized in some specific regions, our methods demonstrate effectiveness by safely augmenting label-irrelevant patches.

Flexibility of Label-irrelevant Patch Augmentation. Besides the label-irrelevant DA methods we mentioned in section 3.3.2, our label-irrelevant patch augmentation approach exhibits high flexibility for various DA methods. We can plug some other DA methods including color jitter (ColorPatchAug),

Figure 7: The AUBC curve for DermaMNIST and OrganSMNIST based on the AL strategies of Least Confidence and Margin. The lines marked with ‘-o-’ denote the widely used baselines in AL, and the lines marked with ‘-x-’ represent existing DA methods while the lines with ‘-*-’ indicate various label-irrelevant patch augmentation methods based on our efficient AL framework.

<table>
<thead>
<tr>
<th>Parameters Size (10^9)</th>
<th>ResNet18</th>
<th>ResNet34</th>
<th>ResNet50</th>
<th>Adapter (Ours)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.8240</td>
<td>3.6787</td>
<td>4.1337</td>
<td>0.0006</td>
</tr>
</tbody>
</table>
RandAug (RandPatchAug), and zero-masking (ZeroMaskPatchAug) to explore the diversity of the localized label-irrelevant patches. These methods are called as extended label-irrelevant patch augmentation methods as follows.

Fig. 7 illustrates the curve for AUBC results for varying budgets during the AL process, where extended patch augmentation methods perform competitively. In most cases like Fig. 7(a)(b)(d), most extended patch augmentation methods perform better AUBC results on varying budgets than existing DA methods (RandAug/AutoAug/NormalAug), but underperform SubstitutivePatchAug and SelfPatchAug. However, as shown in Fig. 7(c), some of the extended patch augmentation methods like RandPatchAug and ZeroMaskPatchAug can even surpass our carefully designed label-irrelevant patch augmentation methods (SelfPatchAug & SubstitutivePatchAug), showing the great potential of the way to plug different DA methods in the flexible label-irrelevant patch augmentation approach.

Do We Need Pretext SSL Training for the Downstream Task? As we claimed in section 1 and section 2, one key difference between our work and existing AL+SSL works is that our method does not require SSL pretext training on the downstream dataset D. The reason is that conducting pretext SSL training on the medical image dataset is not efficient (costing time for training) and can even distort the representations (Fig. 3). This statement was further demonstrated by Table 4, where ViT-B (MAE)-Adapted performed worse than ViT-B (MAE)-Origin with respect to AUBC for various AL strategies on DermaMNIST.

Different r and Localization Methods. The ablation study for different r and localization methods are shown in Table 3. We report the results marked as blue (not the best) in Table 1 since we need to maintain consistent hyperparameters across datasets for fair comparison. It’s important to note that dataset-specific hyperparameter tuning could further enhance the performance of the label-irrelevant patch augmentation. As shown in Table 3 in most cases, CosineAttentionMap and LastAttentionMap can outperform DeepLIFT and Saliency with respect to the AUBC results. However, for Margin sampling with r = 75%, DeepLIFT performs much better than others.

In terms of efficiency, CosineAttentionMap and LastAttentionMap outperform DeepLIFT and Saliency for localizing label-irrelevant patches, as they are produced simultaneously during the forward process without additional computational costs, while DeepLIFT and Saliency require multiple backward propagations, significantly increasing time cost.

5. Conclusion

In this paper, we argue that the key to successful AL is to learn a minimally sufficient representation. We presented an efficient AL framework leveraging off-the-shelf ViTs to gain a relatively good representation at the initial stage of AL. We further propose a DA method for localizing and augmenting label-irrelevant patches, to gradually train a lightweight encoder to transfer the representation closer to minimally sufficient. The effectiveness and efficiency of our framework are widely evaluated across various datasets and AL strategies.
References


[16] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 15979–15988, 2022.


A. Details

In the appendix, we present the details of datasets, hyperparameters, metrics, and our algorithm in appendix A.1 and appendix A.2. Details of the label-irrelevant patch location methods can be found in appendix A.3, and other patch augmentation methods besides SelfPatchAug and SubstitutivePatchAug can be found in appendix A.4. For additional experimental results, appendix B.1 contains results for more comparison and ablation studies, and appendix B.2 includes the results for visualization studies.

A.1. Details of Experimental Settings

Details of Datasets. The details of the datasets we used in our experiments are presented in Table 5. These datasets comprise medical imaging data from diverse modalities, varying numbers of classes, distinct pathological conditions, and different anatomical regions, providing a comprehensive evaluation of our proposed framework.

Details of Hyperparameter Settings. For Tables 1 and 4, and Figs. 6 and 7, the results of SelfPatchAug are produced by fixed $r = 75\%$, and the results of SubstitutivePatchAug are produced by $r = 75\%$, $t = 5$, and $\lambda = 0.5$.

Details of Computing Resources. All the experiments can be run on a single NVIDIA A100 (80GB).

Details of Metrics. In this section, we will introduce the metrics evaluating the representation quality presented in Fig. 2 and Fig. 3.

- **LR Accuracy.** This metric is utilized to verify the linear separability of the learned feature representation space by modeling $P(y|\phi(x))$ with LR. For the implementation details of the LR classifier, we use `LogisticRegression(random_state=66,solver='sag')` from sklearn.

- **KNN Accuracy.** Similar to LR Accuracy, it measures the representation quality by modeling $P(y|\phi(x))$ with KNN. We use the `KNeighborsClassifier(n_neighbors=5,algorithm='auto')` from sklearn as the KNN classifier.

- **Representation Metric.** We denote a matrix as $M \in \mathbb{R}^{c \times c}$, where $c$ is the number of classes. For $M_{i,j}$, it represents a score measuring the distance between the representations from $i$-th class and $j$-th class center, which is computed as eq. (1)

$$M_{i,j} = \frac{1}{N_i} \sum_{z \in \{z|y=i\}} S(z; u_j), \quad u_j = \frac{1}{N_j} \sum_{z \in \{z|y=j\}} z$$

where $S(a; b)$ denotes the Euclidean distance between vector $a$ and $b$. Generally, we compute the intra-class and inter-class distance based on $M$, and then the representation metric $\tau$ shown in Fig. 3 is defined as eq. (2).

$$\tau_{\text{intra}} = \frac{\text{Trace}(M)}{C}, \quad \tau_{\text{inter}} = \frac{\text{Sum}(M) - \text{Trace}(M)}{C}, \quad \tau = \frac{\tau_{\text{inter}}}{\tau_{\text{intra}} + \tau_{\text{inter}}}$$

where a larger $\tau$ indicates a better feature representation space.
A.2. Algorithm of Our Framework

**Algorithm 1 AL with Label-irrelevant Patches Augmentation**

**Input:** Initial Labeled/Unlabeled Dataset $D^\text{lab}_0/D^\text{unl}_0$, AL Rounds $K$, Off-the-shelf Encoder/Decoder $f_{\text{enc}}/f_{\text{dec}}$, Adapter $g$.

1: $r \leftarrow a \mod b$
2: for $k = 0$ to $K$
3: for $(x, y) \in D^\text{lab}_k$
4: $\text{Cor} = \text{Localize}(x, y; f_{\text{enc}}, f_{\text{dec}}, g)$
5: $\bar{x}_p = M_{\text{Cor}} \odot x_p$ /*Localize the Label-irrelevant Patches*/
6: $\hat{x} = A(x)$ /*Argument the Label-irrelevant Patches*/
7: $\beta_x = (1 - M_{\text{Cor}})^T \text{Cor}$
8: $\hat{y} = \beta_x \cdot \frac{1}{|Y|} y + (1 - \beta_x) \cdot y$ /*Instance-adaptive Label Smoothing*/
9: $D^\text{lab}_{k+1} = D^\text{lab}_k \cup \{\hat{x}, \hat{y}\}$
10: end for
11: Train $g$ on $D^\text{lab}_k$
12: $B = \alpha(D^\text{unl}_k, [f_{\text{enc}}, f_{\text{dec}}, g])$ /*Acquire Unlabeled Samples*/
13: $D^\text{lab}_{k+1} = D^\text{lab}_k \cup B$
14: $D^\text{unl}_{k+1} = D^\text{unl}_{k+1} \setminus B$
15: end for

**Output:** Adapter $g$

The overall algorithm of our efficient AL framework with label-irrelevant augmentation is shown in algorithm 1, where $\alpha(D^\text{unl}_k, [f_{\text{enc}}, f_{\text{dec}}, g])$ denotes selecting batch $B$ from $D^\text{unl}_k$ based on the acquisition function $\alpha$.

A.3. Details of Label-Irrelevant Patches Localization Methods

- **LastAttentionMap:** This method generates attention maps using the last layer of the network. These maps are useful in visualizing and understanding where the network is focusing its attention while making predictions.

- **CosineAttentionMap:** This technique generates attention maps using cosine similarity, which measures the cosine of the angle between two vectors. It can highlight the most influential regions in the input for the output prediction.

- **Saliency:** Saliency maps are a common approach in visualizing and interpreting neural networks. These maps show the gradient of the output with respect to the input image, giving an indication of which pixels contribute most to the network’s decision.

- **DeepLIFT:** DeepLIFT is a method for computing the contributions of inputs to outputs, given a neural network. It assigns contribution scores by comparing the activation of each neuron to its activations and computing the differences. This helps to identify which parts of the input are important for prediction.

The results of the ablation study for these four different localization methods are presented in Table 3.

A.4. Label-Irrelevant Patches Augmentation Methods

Many augmentation methods have been used for image classification tasks. We compared our augmentation methods with the following state-of-the-art methods:

- **AutoAug:** AutoAug first adopts a search phase that uses reinforcement learning to do the choose best operations of augmentation.
• **RandAug**: RandAug reduces the complexity of the augmentation process by removing the research phase, which reduces the parameter space to a fixed uniform possibility for every operation of transformation and a universal magnitude parameter.

• **NormalAug**: NormalAug applies a combination of random horizontal and vertical flips to the images.

We also plugged some other DA methods, which we call extended patch augmentation methods:

• **ColorPatchAug**: Color jitter is applied to label irrelevant patches.

• **RandPatchAug**: RandAug is applied to the label irrelevant patches.

• **ZeroMaskPatchAug**: Mask the label irrelevant patches with zeros.

### A.5. Details of AL Strategies

We conducted our experiments with different acquisition functions $\alpha(x, M_k)$ utilized as the query strategies in active learning as follows:

• **LeastConfidence**: Least Confidence method chooses samples whose predicted labels the current model is least certain with. The acquisition function for Least Confidence method can be denoted as $\alpha_{\text{LeastConfidence}}(x, M_k) = -\max_{\hat{y}} p_{M_k}(\hat{y}|x)$, where $\hat{y}$ is the prediction with the highest probability.

• **Margin**: Margin Sampling Looks into the first and second most likely predicted labels of unlabeled samples, and selects those where the difference in probability between the top two predicted labels is relatively small. The acquisition function for Margin Sampling can be denoted as $\alpha_{\text{Margin}}(x, M_k) = -(p_{M_k}(\hat{y}_1|x) - p_{M_k}(\hat{y}_2|x))$, where $\hat{y}_1, \hat{y}_2$ is the two most likely labels of $x$.

• **Entropy**: Maximum Entropy Sampling selects samples with the most significant entropy loss. The acquisition function for Maximum Entropy Sampling can be denoted as $\alpha_{\text{MaximumEntropy}}(x, M_k) = -\sum_c p_{M_k}(y = c|x) \log p_{M_k}(y = c|x)$, where $c$ denotes the classes.

• **Coreset**: CorrSet methods select the most representative samples by selecting a set of center points and minimizing the distance from any point in the dataset to its closest center point. This is equivalent to minimizing the difference between the average loss calculated over the selected center points and the average loss calculated over the entire dataset. The acquisition function for Coreset can be denoted as $\alpha_{\text{Coreset}}(x, M_k) = \max_{x_i \in D_{\text{lab}}} d(M_k(x), M_k(x_j))$, where $d(\cdot, \cdot)$ is distance metric and $M_k(x)$ denotes the representation of $x$ encoded by $M_k$.

### B. Additional Experimental Results

#### B.1. More Ablation Studies

**Results for Many-shot AL.** The results of Many-shot AL are shown in Table 6 and Fig. 8. Compared to Few-shot AL, our DA method delivers more modest enhancements, but it continues to outshine existing DA techniques.

**More Results for Few-shot AL.** More results for different AL strategies under the setting of Few-shot AL are presented in Table 7 and Fig. 9. It shows the effectiveness of our method for boosting the performances for various AL strategies consistently.

**Effectiveness of Instance-adaptive Label Smoothing.** The ablation study of Instance-adaptive Label Smoothing (IaLS) are shown in Table 8 conducted on ViT-B (MAE), which demonstrate the effectiveness of IaLS.

**Structure of Adapter $g$.** We also explore different structures for the adapter $g$ shown in Table 9 where the ResAdapter is designed as adding a skip connection between $g_{\text{enc}}$ and $g_{\text{cls}}$. 


Table 6: Results of the comparison between our proposed label-irrelevant patch augmentation methods and other DA methods across 6 datasets for medical image classification by utilizing Least Confidence and Margin as the AL strategy, under the setting of Many-shot AL.

B.2. Visualization Studies

The visualization results of our proposed label-irrelevant DA methods are shown in Fig. [10] and Fig. [11].
Figure 8: The AUBC results of comparing existing AL model structures (ResNet) and our efficient model structures (ViT-B + Adapter) via various AL strategies, including Least Confidence and Margin. Mean here denotes the averaged values of results among 6 different datasets. The results are produced under the setting of Many-shot AL.

Figure 9: The AUBC results of comparing existing AL model structures (ResNet) and our efficient model structures (ViT-B + Adapter) via AL strategies as Coreset and Entropy under the setting of Few-shot AL. It can be viewed as the extension for Fig. 6.

Table 7: More Results for the comparison study for DA methods under the setting of Few-shot AL. It can be viewed as the extension for Table 1.
Table 8: Results of the Ablation Study on Instance-adaptive Label Smoothing (IaLS) conducted on ViT-B (MAE)

<table>
<thead>
<tr>
<th>Model</th>
<th>Least Confidence With LaLS</th>
<th>Least Confidence Without LaLS</th>
<th>Margin With LaLS</th>
<th>Margin Without LaLS</th>
<th>Mean</th>
<th>1/L1</th>
</tr>
</thead>
<tbody>
<tr>
<td>BloodMNIST</td>
<td>0.6270</td>
<td>0.6040</td>
<td>0.6470</td>
<td>0.5970</td>
<td>0.6200</td>
<td>0.5970</td>
</tr>
<tr>
<td>DermaMNIST</td>
<td>0.6920</td>
<td>0.6570</td>
<td>0.6880</td>
<td>0.6540</td>
<td>0.6570</td>
<td>0.6540</td>
</tr>
<tr>
<td>PneumoniaMNIST</td>
<td>0.8330</td>
<td>0.8230</td>
<td>0.8330</td>
<td>0.8230</td>
<td>0.8530</td>
<td>0.8230</td>
</tr>
<tr>
<td>OrganAMNIST</td>
<td>0.6840</td>
<td>0.6930</td>
<td>0.7220</td>
<td>0.7260</td>
<td>0.6750</td>
<td>0.7100</td>
</tr>
<tr>
<td>OrganCMNIST</td>
<td>0.6390</td>
<td>0.6530</td>
<td>0.7000</td>
<td>0.6810</td>
<td>0.6470</td>
<td>0.6740</td>
</tr>
<tr>
<td>OrganSMNIST</td>
<td>0.5170</td>
<td>0.5410</td>
<td>0.5650</td>
<td>0.5620</td>
<td>0.5720</td>
<td>0.5460</td>
</tr>
<tr>
<td>Mean</td>
<td>0.6833</td>
<td>0.6618</td>
<td>0.6725</td>
<td>0.6778</td>
<td>0.6065</td>
<td>0.6775</td>
</tr>
</tbody>
</table>

Table 9: Results for Different Model Structures of the Adapter \( g \).

<table>
<thead>
<tr>
<th>Model</th>
<th>Least Confidence</th>
<th>Margin</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adapter + SubstitutivePatchAug</td>
<td>0.6513</td>
<td>0.6430</td>
<td>0.6492</td>
</tr>
<tr>
<td>ResAdapter + SubstitutivePatchAug</td>
<td>0.6410</td>
<td>0.6410</td>
<td>0.6446</td>
</tr>
<tr>
<td>Adapter + SubstitutivePatchAug</td>
<td>0.6117</td>
<td>0.6020</td>
<td>0.6140</td>
</tr>
<tr>
<td>ResAdapter + SubstitutivePatchAug</td>
<td>0.6093</td>
<td>0.5970</td>
<td>0.6040</td>
</tr>
</tbody>
</table>

Figure 10: Fig. 10b: Visualization Results for the Query Set \( Q \) when applying SubstitutivePatchAug.
Fig. 10b: Visualization Results for SubstitutivePatchAug among different \( \lambda \).
Figure 11: Visualization Results for Augmenting Label-irrelevant Patches with Different Methods