Cloud Computing and Data Science
https://ojs.wiserpub.com/index.php/CCDS/

UNIVERSAL WISER
PUBLISHER

Research Article

GCI-ViTAL: Gradual Confidence Improvement with Vision Transformers for Active Learning on Label Noise

Moseli Mots' oehli[*], **Kyungim Baek**

Department of Information and Computer Sciences, University of Hawai'i at Manoa, Honolulu, HI, 96822, United States
E-mail: moselim@hawaii.edu

Received: 28 May 2025; **Revised:** 17 June 2025; **Accepted:** 20 June 2025

Abstract: Active Learning (AL) aims to train accurate classifiers while minimizing labeling costs by strategically selecting informative samples for annotation. This study focuses on image classification tasks, comparing AL methods on the CIFAR10, CIFAR100, Food101, and Chest X-ray datasets under varying label noise rates. We investigate the impact of the model architecture by comparing Convolutional Neural Networks (CNNs) and Vision Transformer (ViT)-based models. We propose a novel deep AL algorithm, Gradual Confidence Improvement with Vision Transformers for Active Learning (GCI-ViTAL), designed to be robust to label noise. GCI-ViTAL utilizes prediction entropy and the Frobenius norm of last-layer attention vectors compared to class-centric clean set attention vectors. Our method identifies uncertain and semantically divergent samples from typical images in their assigned class. This allows GCI-ViTAL to select informative data points even in the presence of label noise while flagging potentially mislabeled candidates. Label smoothing is applied to train a model that is not overly confident about potentially noisy labels. We evaluate GCI-ViTAL under varying levels of symmetric label noise and compare it to five other AL strategies. Our results show that using ViTs leads to better performance over CNNs across all AL strategies, particularly in noisy label settings. Additionally, using the semantic information of images as label grounding leads to a more robust model under label noise. Notably, we skip extensive hyperparameter tuning, providing an out-of-the-box comparison that helps practitioners select AL models and strategies without an exhaustive literature review on real-world vision model tuning.

***Keywords*:** deep active learning, vision transformer, label noise, image classification

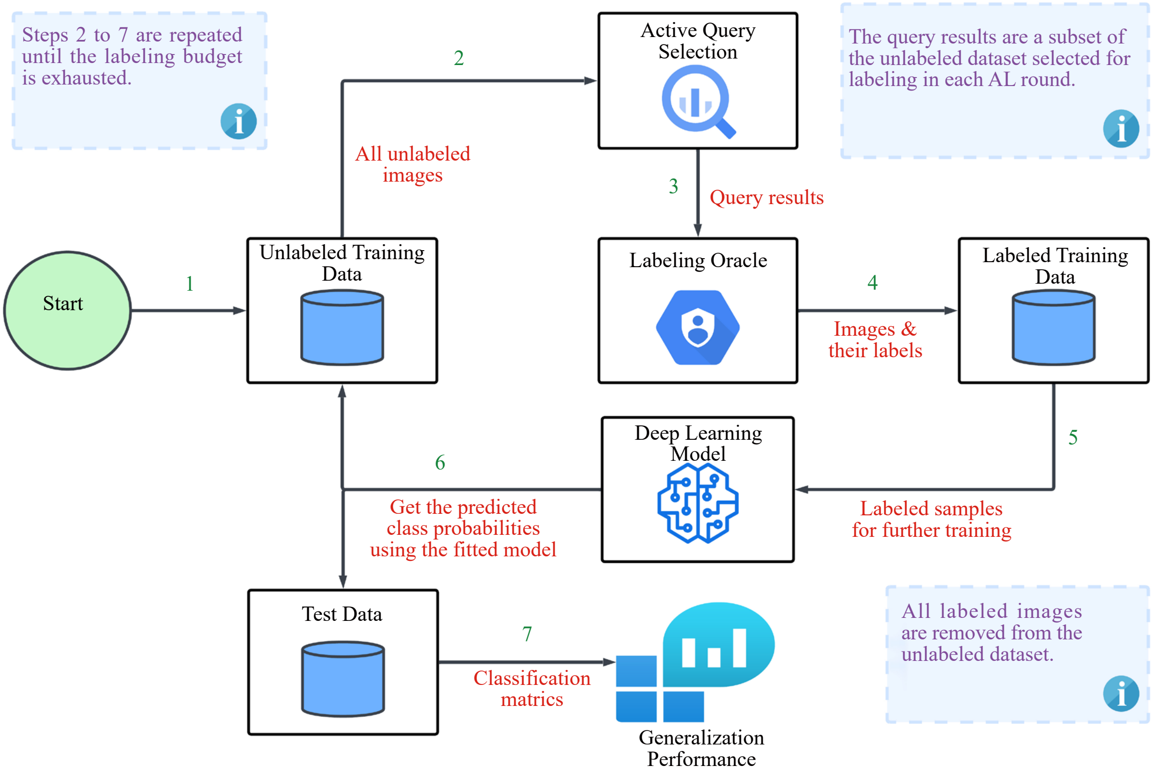
Abbreviation

DL	Deep Learning
AL	Active Learning
DAL	Deep Active Learning
MHSA	Multi-Head Self-Attention
CNN	Convolutional Neural Network
ViT	Vision Transformer
MLP	Multi-Layer Perceptron

DOI: https://doi.org/10.37256/ccds.6220257319

1. Introduction

While most works in the literature often compare results to baseline AL algorithms such as random selection or simple entropy-based selection, a large number of hyper-parameter tuning performed during training are often omitted, resulting in significant performance differences among authors for the same CNN architecture, AL algorithm [1], and label noise rate [2]. This not only raises questions about the credibility of reported state-of-the-art results but can also delay progress in developing AL schemes that are robust to label noise and achieve performances comparable to models trained on clean labels. The work on the non-active training of DL models in the presence of label noise, as well as the training of DL models on noise-free datasets in an AL setting, are well addressed in the literature. However, the intersection of these niches has a long way to go [3]. AL algorithms seek to train an optimal model with minimal training data that is labeled iteratively.



Steps 2 to 7 are repeated until the labeling budget is exhausted.
2
Active Query Selection
The query results are a subset of the unlabeled dataset selected for labeling in each AL round.
All unlabeled images
3
Query results
Start
1
Unlabeled Training Data
Labeling Oracle
4
Images & their labels
Labeled Training Data
5
Deep Learning Model
6
Get the predicted class probabilities using the fitted model
Labeled samples for further training
Test Data
7
Classification matrics
Generalization Performance
All labeled images are removed from the unlabeled dataset.

Figure 1. The Deep Active Learning with label noise framework

Most AL methods seek to explore diverse training examples or focus on samples that the DL algorithm is uncertain about. AL in the presence of label noise is a particularly challenging topic since training DL models with a higher concentration of incorrect labels presents problems for the back-propagation algorithm's ability to converge as demonstrated in [2]. It has also been shown that without sufficient training data, large models can memorize the noisy labels, and fail to generalize to the test set [4-8]. Figure 1 depicts the basic AL framework for training DL image classifiers in the presence of label noise.

In this work, we compile a unified view of existing works on AL for image classification with label noise. We are particularly interested in the fine-tuning CNN and ViT models pre-trained on the ImageNet-1k dataset. We explore different DL model architectures and AL strategies on different datasets while varying the label noise rates up to 60%. We re-implement the commonly used baseline AL strategies namely: random query, maximum entropy, margin-based selection, model delta, and hybrid uncertainty sampling with diversity.

This work seeks to address the following:

- In the realm of AL, where the emphasis is often on query selection, and in image classification with label noise, which is typically trained without AL, reliable benchmarks for AL algorithms with label noise are scarce. To address this, we train DL models with various AL algorithms across noise levels and report results on standard datasets, including CIFAR10, CIFAR100, Food101, and Chest X-ray images (pneumonia).
- Given that the ViT-based models currently outperform CNNs in image classification on CIFAR10 [9, 10], and most competitive results on CIFAR100, Food101, Chest X-ray images (pneumonia), and other classification datasets [10-12], how do the ViTs compare to CNN-based models in an AL setting with Label Noise (ALLN), and what can be done to improve on ViT learners in this setting?
- Lastly, we propose an AL scheme customized for the properties of the transformer network to improve AL in the presence of label noise. We also provide new insights based on using ViTs for AL under label noise and propose avenues to advance this work.

2. Related works

In this section, we highlight the current state of the literature on AL, as well as AL with label noise. We also highlight the use of the ViT for image classification and briefly discuss works that employ the ViT for AL and label noise settings.

2.1 *Deep active learning*

In most supervised Machine Learning (ML) applications, there is an initial data annotation cost in both money and time. In some domains and tasks, datasets are inherently difficult to label for various reasons. In other cases, the cost of hiring expert annotators is high, such as is the case in medical imaging [13, 14], or the cost of producing the samples is high, such as is the case in experimental physics where observations come from costly telescopes or particle accelerators with limited access [15-17]. This challenges the real-world use of ML systems, especially as the unlabeled dataset sizes increase.

Although much progress has been made in improving Self-Supervised Learning (SSL) methods to leverage large unlabeled datasets to extract quality image embeddings [18-22] to be used in downstream tasks with little labeled data, these methods still fall short when applied directly to noisy labels-labeled datasets in an AL setting [3]. AL is an ML paradigm, as shown in Figure 1, which seeks to address the issues related to training ML models within a labeling budget, letting the learning algorithms iteratively select a subset L^m of size m, from a larger unlabeled dataset U^n of size $n : m \leq n$, to be labeled by an oracle O for training [23-27]. However, the oracle may not always provide the correct label [28-30]. The AL mantra under label noise can be stated as follows: Train an ML model on a significantly smaller dataset that may contain p% label noise, with little to no drop in test performance, while staying within a pre-determined labeling budget B.

2.2 *DAL with label noise*

Most work in literature uses random query, uncertainty sampling, and entropy-based sampling as baseline algorithms in comparing more complex methods for ALLN such as [28, 31], where a mixture of information gain and uncertainty is used for query selection. Other works in the literature focus on reducing the labeling budget by using a mixture of weak and strong annotators as well as annotator abstention in the case of uncertainty [28, 32-33]. Despite these works posting promising results in budget optimization, there is little to no improvement in terms ofthe robustness to label noise and improved query selection.

In [34], Huang et al. show that DAL is viable with oracle epiphany, that is, the oracle is allowed to abstain from labeling samples they are unsure of until later in the DAL cycle, once they have seen enough examples to provide a more confident label. While their method is more realistic and leads to better performance, abstention may not always be possible in a fast-paced sector where lots of data is generated on a daily basis and requires urgent labeling. In [35], Yan et al. utilize abstention in DAL under label noise. A key difference in their work is that the algorithm need not be aware of either the abstention or noise rate. While there are obvious merits to the methods above, the solutions rely too heavily on the specific setup of the AL cycle and the human annotators to be trusted in the general setting. For these and more reasons, our work focuses on a more algorithmic approach to robustness and query selection.

Recent work [36] deploys a more data-centric approach to label noise for active label cleaning by ranking samplesfor label correctness and labeling difficulty. In [37], the authors propose a data-driven self-adapting DAL strategy that selects potentially noisy labels for correction in a manner that automatically avoids class imbalance in thelabeled dataset with no prior knowledge of the class distributions. A thorough study of the literature [1, 3, 38] raises questions on the superiority and robustness of more complex DAL methods. The authors state that due to a lack of standardized benchmark settings in overlapping niches, performances by baseline models and AL queries in noisy labels tend to be understated. The use of grid search and related methods for hyper-parameter tuning can also aid in findingthe idealmodel parameters that show the superiority of a proposed AL model over the baseline methods. It is for this reason we opt to avoid any extensive hyper-parameter optimization that could skew the results of this study, and thus we report a realistic out-of-the-box benchmark in AL under label noise.

2.3 *Vision transformer for image classification*

In [9], Kolesnikov et al. present the Vision Transformer (ViT) as a CNN replacement for image classification tasks. They show that, in the extensive dataset regime, ViTs achieve higher classification accuracy, are more computationally efficient, and show no signs of saturation compared to CNNs such as ResNet and EfficientNet on increasingly larger datasets. The standard ViT architecture takes 16 by 16 patches from an image, flattens them, and applies a linear projection onto a higher dimensional space equal to that of the original text-based transformer. The patches are then marked for where in the image they were extracted, and so a second input to the transformer is the 1D positional embedding of each patch as shown in Figure 2. The spatial correlations of the patches are implicitly learned through their positional embeddings and self-attention vectors. The positional embeddings ensure the model learns both the relationships between pixels in the patches as well as the local and global 1D proximity representations of the tokens (image patches). Below we give an overview of the main transformer encoder and the attention mechanism within the ViT architecture.

Transformer Encoder: The transformer encoder block in a ViT consists of Multi-Head Self-Attention (MHSA) and Multi-Layer Perceptron (MLP) layer with normalization layers before and after MHSA each performing specificoperations ontheinput tokens or subsequent later outputs. MHSA attempts to learn diverse features to capture semantical and contextual complexity in image patches while learning all these features in parallel for a given input. The parameterized MLP pools and aggregates the learned high-dimensional features from all the last layer attention heads, and compresses them into a lower-dimensional representation for a downstream task such as image classification, regression, and more. Figure 3 illustrates the transformer encoder block architecture.

Self-Attention: Given a sequence of input vectors, theself-attention mechanism calculates the similarity between the vectors. In the case of ViTs, starting with a sequence of m patch embeddings of size e, $X = [x_1, x_2, ..., x_m]$, where $x_i \in R^e$ and so $X \in R^{m\times e}$, the goal of the self-attention mechanism is to learn a representation of each input token, expressed as a vector that is a weighted sum of the other token representations, where the weighting is based on how similar the tokens are to each other contextually. This similarity is measured by taking the dot product of any two embedding vectors, representing the cosine similarity between two vectors. To achieve this, three parameterized matrices are used: the Query matrix $W^Q \in R^{e\times e_q}$, the key matrix $W^K \in R^{e\times e_k}$, and the value matrix $W^V \in R^{e\times e_v}$, such that $e_q = e_k$.

To calculate the attention weights, the input sequence X is first projected into the query, key, and value representations through matrix multiplication to get $Q = XW^Q$, $K = XW^K$, and $V = XW^V$. Applying a softmax operation to the normalized dot product of the query and key projections of the input produces the attention weights. The final output of the attention layer is a weighted sum of the value representations, where the weights are the attention weights:

$$\mathbf{Z} = \text{softmax}\left(\frac{\mathbf{QK}^T}{\sqrt{e_q}}\right)\mathbf{V} \tag{1}$$

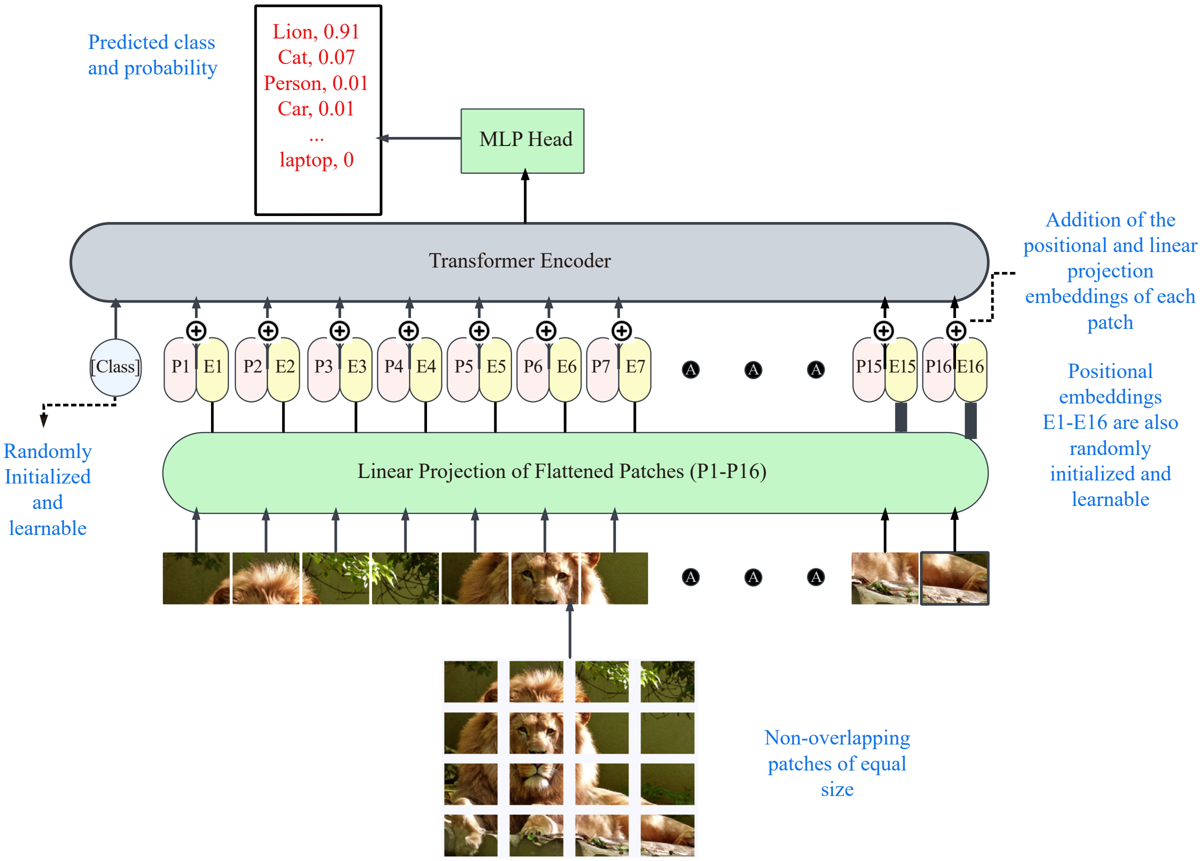

Predicted class and probability
Lion, 0.91
Cat, 0.07
Person, 0.01
Car, 0.01
...
laptop, 0
MLP Head
Transformer Encoder
Addition of the positional and linear projection embeddings of each patch
[Class]
P1 E1 P2 E2 P3 E3 P4 E4 P5 E5 P6 E6 P7 E7 P15 E15 P16 E16
Positional embeddings E1-E16 are also randomly initialized and learnable
Randomly Initialized and learnable
Linear Projection of Flattened Patches (P1-P16)
Non-overlapping patches of equal size

Figure 2. ViT model architecture overview (Adopted from [9])

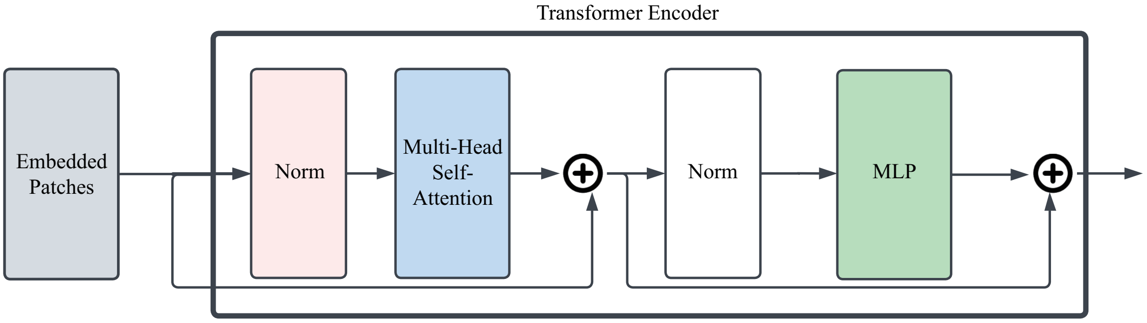


Figure 3. The transformer encoder block (Adopted from [9])

Multi-Head Self-Attention (MHSA): Extends self-attention capabilities by allowing the model to use multiple self-attention heads with different learned parameters in parallel over the same input. Learning using one attention would force the model to compress all learned information about an input image into one fixed-length vector. Using multiple attention heads allows the model to learn and express semantic and contextual complexity over several specializing attention heads simultaneously. A ViT can have one or more MHSA encoders with residual connections to maintain lower-level feature information throughout the network for context in higher-level attention blocks. In this work, we focus on the last layer of MHSA heads as well as the predicted class probabilities after the MLP is trained for classification.

Figure 2 is an illustration of the ViT architecture, showing patch and positional embeddings, as well as the transformer encoder adapted from [9]. Other ViT variations that achieve performances comparable to and in some cases better than CNNs on image classification tasks include Swin Transformer [39], Transformer in Transformer (TNT) [40], DaViT-G [41], and Multiscale Vision Transformers (MViT) [42]. The Swin transformer adapts a similar structure as the ViT but uses shifted overlapping patches instead of rigid non-overlapping patches.

The authors of [39] argue and demonstrate that using shifted window patches instead of non-overlapping patches limits the self-attention operations to locally related patches of an image. This reduces the complexity of the self-attention operations from quadratic to linear on the input. This architectural choice makes the Swin transformer more efficient than the original transformer while performing comparably in terms of top-1 classification accuracy. That said, the performance of ViTs and their variants such as the Swin transformer compared to CNNs has not received much attention in the noisy label domain or AL. For these reasons, we included both the base ViT and the Swin transformer in our experiments.

2.4 *Vision transformers in active learning*

Previous works that adopt the ViTs for image classification in the AL domain include [43, 44]. Caramalau et al. [43] introduce a novel AL query strategy that combines CNN layers for local dependencies and ViTs to capture non-local dependencies while jointly minimizing a task-aware objective. They achieve state-of-the-art performance on most AL-based benchmarks. Their method however suffers from scalability limitations due to ViT's large parameter space and potential batch size restrictions in training. A similar conclusion is reached by He et al. [44]. Their work demonstrates that, while ViTs produce informative and task-aware AL queries on CIFAR10 and 100, they are considerably larger than CNNs in terms of model parameters for them to be a viable replacement in DAL with the existing hardware in terms of training time.

The work of Rotman and Reichart [45] compares different DAL methods on different text classification datasets using transformer-based models. While their work is not focused on image classification or the vision transformer, they demonstrate that transformer-based models tend to lead to inconsistent and poor results in the AL setting when using basic AL strategies. They show that query selection based on a transformer learner sometimes leads to the selection of clusters of neighboring outliers that destabilize training. More recent studies have explored ViT robustness in medical image classification under label noise [46], while [47] adopt active learning to vision-language models to demonstrate even stronger noisy-label detection.

In this work, we introduce a novel DAL algorithm, GCI-ViTAL, that takes advantage of the ViT's ability to capture complex local and global dependencies, like [43]. We define the C-Core attention vectors to help reduce the computational complexity for comparisons between labeled and unlabeled samples per AL circle. Like [44], we make use of the C-Core attention vectors as an informative and noise-aware approach to sample selection.

3. Query based on ViT patch similarity

Most AL approaches rely solely on the predicted probabilities from the trained model to form their query strategy. Label noise leads to a confused learner that outputs uncalibrated probability estimates of the samples, so selecting samples based on the predicted probabilities alone leads to the selection of non-optimal samples for each iteration and further corruption of the model through unstable gradients from incorrectly labeled samples. To address this, we use last-layer ViT attention vectors for query selection, detecting mislabeled samples by comparing labels of images with similar attention representations. We start with a pre-trained ViT and fine-tune it using a small set of images with accurate labels, a common practice in many AL algorithms. Subsequently, we extract attention vectors for all the images in the initial set with accurate labels, and we use these to create core attention representations (centroids) for each class by aggregating the attention vectors of images belonging to the same class. This results in C-Core attention representation vectors that have been trained exclusively using accurate data for a C-class classification problem. The purpose of these core attention vectors is to help identify potentially incorrect labels during the AL process and reduce their impact on the model's training. In each AL cycle, unlabeled images go through the fine-tuned ViT model, and their last layer's attention vectors are collected. Our AL strategy selects samples by combining prediction uncertainty with C-Core attention-based diversity to ensure balanced class representation.

3.1 *Handling label noise through gradual class-centric confidence improvement*

During training, an oracle receives a batch of K samples to label, and before retraining the model with these labeled samples, a portion of them that deviate too much from the C-Core attention vector of the assigned class have their class assignment probabilities changed. We explore handling these examples in two ways, first by assigning the samples to the class dictated by the C-Core attention vectors, or, secondly by label smoothing. With label smoothing, we change the class probabilities assigned by the oracle for a sample so that it is not one-hot-encoded, but rather we introduce a positive probability for another class. Since we have the C-Core class centroids, we smooth the label by assigning a positive probability to the closest class centroid from the current sample in the attention vector space. For example, say sample x_j was selected for labeling, and the noisy oracle assigns it to class C_4 out of 10 classes. This means the probability distribution is given by [0, 0, 0, 1, 0, 0, 0, 0, 0, 0]. However, if we suspect this labeling to be incorrect based on the sample x_j being far from the core attention centroid of class C_4, we alter the probability distribution in such a way that we express less confidence in class C_4 being the correct label. If the closest centroid in attention vector space is that of class C_9, we change the probability distribution $p(y|x_j)$ to be $[0, 0, 0, (1\text{-}\epsilon), 0, 0, 0, 0, \epsilon, 0]$, where $\epsilon \in [0, 1]$ controls how much to trust the oracle as opposed to the C-Core attention vectors that compare images semantically.

Once the initial model is trained on the clean set, future batches for labeling are selected in a way that promotes class-centric confidence. We first calculate prediction uncertainty for each class based on the model's confidence in its predictions. The AL strategy selects samples with the highest distance to their core attention class centroid, measured by the Frobenious norm. We then rank these from highest to lowest based on the Frobenious distance and take the top-K for labeling. As the model improves its performance on the validation set, we shift towards selecting samples with high uncertainty and gradually decreasing reliance on the class centroid distance to the image representation, ultimately focusing on refining the model's understanding of each class.

The selection based on the distance from the class centroid ensures we select samples that have underlying attention maps that are not similar to most seen in the clean training set for that class, which are effectively semantically hard examples. We want these to be sent to the oracle for labeling. Our proposed AL strategy for image classification in the presence of label noise by leveraging attention vectors is summarized in Figures 4 and 5. A detailed description of the steps in Figure 5 is provided in Section 4.2.

In summary:

1. Most AL strategies rely solely on model-predicted probabilities and the oracle's labels, which may lead to suboptimal sample selection and unstable gradients due to label noise.

2. Our approach uses the last-layer attention vectors (C-Core attention vectors) of the ViT, to identify potential images with semantic similarity that have opposing oracle labels. These are potentially incorrectly labeled. We reduce the effects of label noise through label smoothing by combining prediction uncertainty and the distance to the most likely C-Core attention vector.

3. We employ a Gradual Class-Centric Confidence Improvement strategy, initially selecting samples with a high Frobenius norm concerning their supposed class centroid attention maps for low-confidence classes. Gradually, we shift our strategy towards samples with a low Frobenius norm to better understand the classes. This approach helps in the selection of semantically challenging examples and reduces the impact of label noise.

In the rest of this section, we provide the theoretical formulation of the problem, the algorithm, and the theoretical justification for our approach.

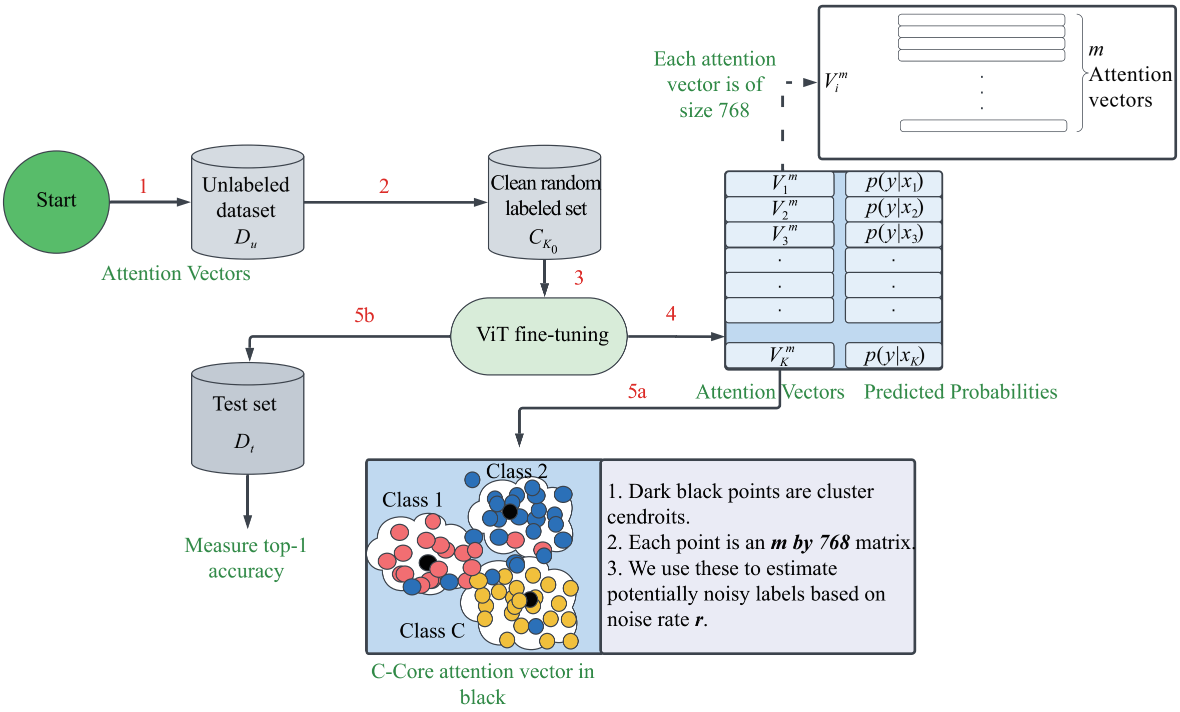

Each attention vector is of size 768
V_i^m
m Attention vectors
Start
1
Unlabeled dataset D_u
Attention Vectors
2
Clean random labeled set C_{K_0}
3
ViT fine-tuning
4
5b
5a
V_1^m
V_2^m
V_3^m
V_K^m
$p(y|x_1)$
$p(y|x_2)$
$p(y|x_3)$
$p(y|x_K)$
Attention Vectors
Predicted Probabilities
Test set D_t
Measure top-1 accuracy
Class 1
Class 2
Class C
1. Dark black points are cluster cendroits.
2. Each point is an ***m by 768*** matrix.
3. We use these to estimate potentially noisy labels based on noise rate ***r***.
C-Core attention vector in black

Figure 4. GCI-ViTAL stage-1 selection schematic

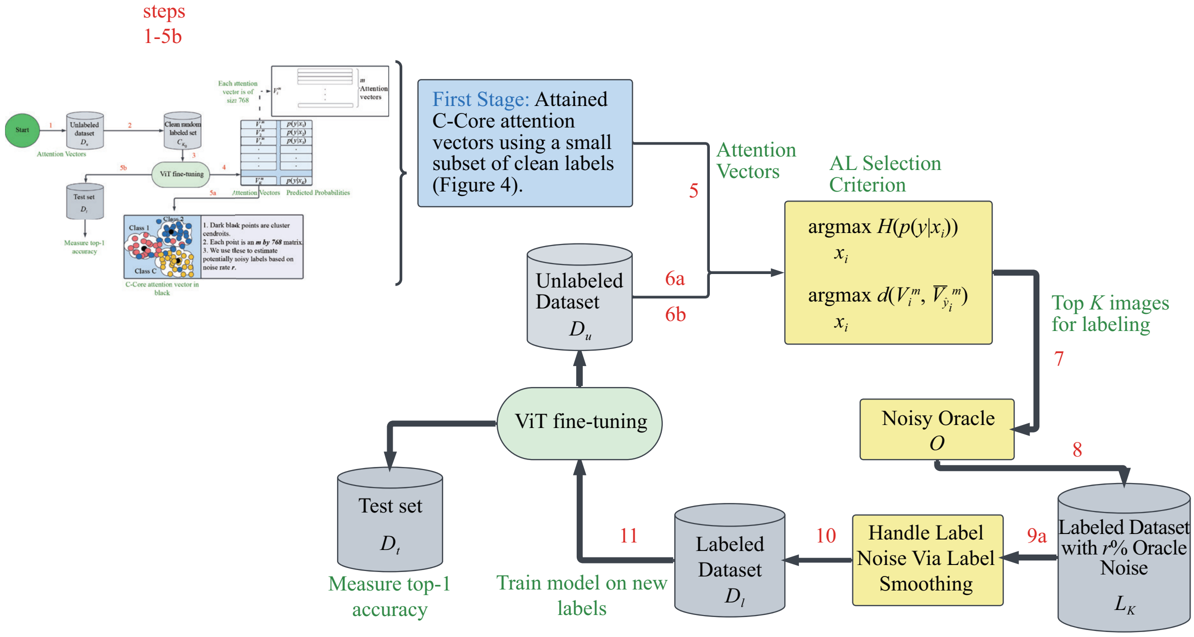

steps 1-5b
First Stage: Attained C-Core attention vectors using a small subset of clean labels (Figure 4).
5
Attention Vectors
AL Selection Criterion
argmax $H(p(y|x_i))$ x_i
argmax $d(V_i^m, \bar{V}_{\hat{y}_i}^m)$ x_i
6a
6b
Unlabeled Dataset D_u
Top K images for labeling
7
ViT fine-tuning
Noisy Oracle O
8
Test set D_t
Measure top-1 accuracy
11
Train model on new labels
Labeled Dataset D_l
10
Handle Label Noise Via Label Smoothing
9a
Labeled Dataset with r% Oracle Noise L_K

Figure 5. The GCI-ViTAL framework: C-Core-based selectio-with label-smoothed noise handling

3.2 *Joint entropy-attention active learning*

Starting with: (1) An initial set of K_0 labeled images, referred to as the “clean labeled set”, consisting of randomly selected images. This set is denoted as C_{K_0} and should be large enough to ensure the representation of at least a few examples from each of the C classes. (2) A base Vision Transformer (ViT) denoted by ViT_b. The ViT has been pre-trained on ImageNet-1k [48], and has the fully connected layers changed to suit the number of classes *C*. (3) Iteratively fine-tune the ViT on a collection of *K* images with the highest prediction entropy $H(p(\hat{y}|x_i))$ and ViT hidden layer attention heads distance to their supposed centroid attention heads, i.e., the images which jointly maximize Equations 2 and 3:

$$\underset{x_i}{\operatorname{argmax}} H\left(p\left(\hat{y}|x_i\right)\right) = \underset{x_i}{\operatorname{argmax}}\left[-\sum_{y=1}^{C} p\left(y|x_i\right) \log p\left(\hat{y}|x_i\right)\right] \tag{2}$$

$$\underset{x_i}{\operatorname{argmax}}\, d\left(V_i^m,\ \overline{V}_{\hat{y}_i}^m\right) \tag{3}$$

where *d* is a suitable distance measure, *C* is the number of classes, and $\hat{y}_i$ is the predicted class of x_i. We also have: the *m* attention vectors V_i^m produced by running image x_i through the fine-tuned model and

$$\overline{V}_{\widehat{y_i}}^m = \frac{\sum_{x_j \in C_{K_0}} z_j^{\hat{y}_i} V_j^m}{\sum_{x_j \in C_{K_0}} z_j^{\hat{y}_i}} \tag{4}$$

where

$$z_j^{\hat{y}_i} = \begin{cases} 1 & \text{if } x_j \in \text{class } \hat{y}_i \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

is an entry in an indicator vector, where each entry is 1 if the image x_j belongs to class $\hat{y}_i$. Consequently, the denominator in Equation 4 represents the number of images in the initial clean labeled set that belong to class $\hat{y}_i$.

Equation 4 represents each of the C-Core attention maps or centroids for all images in the clean labeled random set belonging to the same class. Note that these class centroids are produced only after fine-tuning on this labeled set with no label noise. These represent our best estimate of the truth at this point before oracle noise influences our AL strategy. V_i^m and $\overline{V}_{\hat{y}_i}^m$ are each a collection of $m = 12$ vectors in the case of ViT_b, for each of the 12 attention heads, of embedding size of 768.

We will use the notation V_i^m and leave out the 768 dimension going forward, however, each $V_i^m \in \mathbb{R}^{m\times 768}$ will be treated as a matrix containing attention vectors. Calculating the distance between matrices is typically done by using the Frobenius norm (F_n) [49], the spectral distance (S_d) [50], or the Kullback-Leibler divergence (D_{KL}) [51]. F_n is calculated by summing the squares of the differences between all the elements of the two matrices and then taking the square root of the sum, similar to the L_2 distance between vectors. S_d distance quantifies how different two matrices *A* and *B* are based on the largest eigenvalue of the difference matrix between them $||\mathrm{A} - \mathrm{B}||_2$. D_{KL} treats matrices as probability distributions and measures the amount of work to be done in transforming one distribution into another. Similar to the works in [52-55] we argue the Frobenius norm between the latent representations of input images or text can improve discriminability. Unlike these works, we however do not use the Frobenious norm as a parameterized regularizer in training the network, but rather as part of our AL query strategy. Since we are comparing *m* × 768 centroid matrices to *m* × 768 image attention maps, the Frobenius norm also becomes a natural choice in calculating the distance in the semantically aware representations learned by self-attention. The norm is given by:

$$d\left(V_i^m,\ \overline{V}_{\hat{y}_i}^m\right)_{F_n} = \sqrt{\sum_{j=1}^{m}\sum_{k=1}^{768}\left(V_i^{jk} - \overline{V}_{\hat{y}_i}^{jk}\right)^2} \tag{6}$$

To have a combined objective for optimization, we scale the distance to be in the same range as the entropy. For a two-class image classification problem, the entropy is between 0 and 1, but for a multiclass problem, the entropy is between 0 and log(*C*), where *C* is the number of classes. An easy way to achieve this for a loss with range [0, ∞) is to first rescale values to the range [0, 1] and then scale all outputs to log(C). This would then yield a final range: [0, log(C)]. For image x_i, we calculate F_n, the distance to each clean set cluster centroid to get vector distances to all the C classes. We then apply a softmax function that converts this into a probability distribution over the classes. This is to say, based on features alone, which class is most semantically similar to the image in question? We then multiply this probability vector by log(C) to standardize it to the magnitude of the entropy selection criterion. We introduce a weight $\lambda \in [0, 1]$, that balances and controls the influence of entropy versus class-based feature similarity. The final objective is given by:

$$\underset{x_i}{\operatorname{argmax}}\left[\lambda H\left(p\left(\hat{y}|x_i\right)\right)+(1-\lambda)\ \operatorname{softmax}\left(d\left(V_i^m,\ \bar{V}_{\hat{y}_i}^m\right)F_n\right)\log(C)\right]. \tag{7}$$

Combining equations 2 and 3, through 7 we get the following:

$$\underset{x_i}{\operatorname{argmax}}\left[-\lambda\sum_{y=1}^{C}p\left(y|x_i\right)\log p\left(\hat{y}|x_i\right)+(1-\lambda)\frac{e^{d\left(V_i^m,\ \bar{V}_{\hat{\vartheta}_i}^m\right)_{Fn}}}{\sum_{y=1}^{C}e^{d\left(V_i^m,\ \bar{V}_y^m\right)_{Fn}}}\log(C)\right] \tag{8}$$

Algorithm 1 below shows the pseudo-code for GCI-ViTAL, the AL query strategy we propose in this work in the presence of label noise.

Algorithm 1 GCI-ViTAL: An AL strategy for handling label noise using class prediction entropy and the ViT final layer attention head vectors to identify candidate images of which the most recently trained model is uncertain, and the images seem to be semantically different from typical images in their supposed class label based on the extracted ViT attention head vectors.

```
Require:
    1. A pre-trained ViT model $ViT_b$
    2. A small initial random set $C_{K_0}$ of $K_0$ images with accurate labels from all classes
    3. A set of unlabeled images $D_u$
    4. A labelling oracle $O_r$ with a know n-symmetric label noise rate $r$
    5. A number of samples to label, $K$, per labeling cycle
    6. A weight $\lambda \in [0, 1]$ to balance the influence of entropy versus class-based feature similarity
Ensure: Current labelling budget $B_t \geq 0$
Randomly initialize $D_l = C_{K_0}$
Fine-tune $ViT_b$ using labeled set $D_l$
Extract attention vectors for all images in $D_l$
Create the C-Core attention representation vectors (centroids) for each class in $D_l$
while $D_u \neq \varnothing$ do
    Select $K$ samples from $D_u$ using the following strategy:
    Calculate prediction entropy for each class based on the model's confidence in its predictions
    Calculate the distance to each C-Core attention vector to get a distance vector to all the $C$ classes
    Apply a softmax function to convert the distances to a probability distribution over the classes
    Calculate the final objective for each sample:
        $\underset{x_i}{\operatorname{argmax}}\left[\lambda H\left(p\left(\hat{y}|x_i\right)\right)+(1-\lambda)\ \operatorname{softmax}\left(d\left(V_i^m,\ \bar{V}_{\hat{y}_i}^m\right)F_n\right)\log(C)\right]$   (9)
    Select the top-$K$ samples with the highest final objective
    Send the $K$ selected samples to the oracle for labeling
    Calculate the disagreement between the oracle's labels and the C-Core attention-based class assignment for each of the $K$ samples
```

```
        Apply label smoothing on all potentially noisy labels based on the C-Core attention vectors
        Update $D_l$ with the labeled samples
        Fine-tune $ViT_b$ using $D_l$
        Measure top-1 accuracy on the test set
        Update $D_u$ by removing the $K$ selected samples
    end while
    return
```

3.3 *Theoretical analysis*

In this section, we investigate the theoretical relationship between the predicted class probability distribution entropy, the Frobenius norm of the final layer ViT attention heads of the potential AL query candidates, and the C-Core attention head vectors. We then analyze the relationship between our strategy and increased label noise.

3.3.1 *Attention, entropy, and the frobenius norm*

Starting with the self-attention head outputs of samples in the clean initial random sample:

$$\mathbf{Z} = \text{softmax}\left(\frac{\mathbf{QK}^T}{\sqrt{e_q}}\right)\mathbf{V} \tag{10}$$

(Self-attention: measures how strongly every image patch attends to the others.)

For a C class image classification problem, we first calculate the C-Core attention vectors as the mean attention vector representation of all images in each class. These are synonymous with cluster centroids. The Frobenius norm component in the query strategy in Equation 8 measures the disparity between the attention heads of a given image and the C-Core attention vector of the predicted class. By doing this, part of the query selection strategy prioritizes data points where the attention heads in the final layer of the model capture information that is distinct from the C-Core attention vectors of the predicted class. Such images represent novel and unseen variations within a class based only on the attention mechanism. The entropy component of the query strategy on the other hand only uses the predicted class probabilities in ranking samples. Entropy quantifies the uncertainty in a model's predictions for a given image. Images with high prediction entropy are often not well represented in the dataset and thus selecting these samples provides more information. The combination of entropy and the Frobenius norm thus strikes a good exploration and exploitation balance of the image space. This leads to queries that gradually build a model trained on diverse data samples based on the semantical representation of the images through the use of the C-Core attention vector representation of samples, while gradually building up the model's confidence around the decision boundaries between classes based on both semantics as well as filtered label information from the oracle.

3.3.2 *Label smoothing using the C-Core attention-vectors to reduce the effects of label noise*

In this section, we show how label smoothing potentially incorrect oracle labels helps reduce the adverse effects of incorrect labels during training. Mathematically, we can express label smoothing as a form of regularization or penaltybased learning. In the case of multi-class image classification using the cross-entropy loss, consider the following scenario: For a single training example x_i with predicted probabilities $p(\hat{y}|x_i)$, the cross-entropy loss without label smoothing is given by:

$$L = -\lambda \sum_{y=1}^{C} p\left(y|x_i\right) \log p\left(\hat{y}|x_i\right) \tag{11}$$

(Cross-entropy loss is high when the model disagrees with the oracle's label.)

where $p(y|x_i)$ is the one-hot encoded label from the oracle, and $p(y|x_i)$ is the predicted probability distribution by the model. With label smoothing, the cross-entropy loss is given by:

$$L = -\lambda\left[\sum_{y=1}^{C}(1-\epsilon)p\left(y|x_i\right)\log p\left(\hat{y}|x_i\right)+\frac{\epsilon}{C}\sum_{y=1}^{C}\log p\left(\hat{y}|x_i\right)\right] \tag{12}$$

(Label smoothing: spreads an ϵ fraction of confidence evenly across all classes.)
where ϵ is the smoothing parameter.

The term $-\frac{\epsilon}{C}\sum_{y=1}^{C}\log p\left(\hat{y}|x_i\right)$ in the loss is a regularization term that penalizes the model for being too confident in noisy label settings, thus forcing it to learn more robust decision boundaries between classes. However, in this case, ϵ is shared between the other $C - 1$ classes to distribute uncertainty amongst them equally. In our case, since we have the C-Core attention vector per class, we only label smooth by assigning the C-Core selected class a non-zero probability. This means that each sample that the oracle assigns to a class other than that we would assign using the Frobenious distance between the image's attention vectors and the C-Core vectors is label smoothed to reflect lower than 100% confidence in the oracle's label, and exactly ϵ% confidence in the class based on the distance to the C-Core attention vectors. Mathematically, we add an indicator variable to Equation 12 so that all other classes that are not the C-Core prediction do not get their zero probability adjusted. The cross-entropy loss based on C-Core Frobenious label smoothing is given by:

$$L = -\lambda\left[\sum_{y=1}^{C}(1-\epsilon)p\left(y|x_i\right)\log p\left(\hat{y}|x_i\right)+I\left(\hat{y}=\hat{y}_{cc}\right)\epsilon\sum_{y=1}^{C}\log p\left(\hat{y}|x_i\right)\right] \tag{13}$$

(C-Core variant: that same ϵ mass moves only toward the nearest centroid class.)
where

$$I\left(\hat{y}=\hat{y}_{cc}\right)=\begin{cases}1 & \text{if } \hat{y}=\hat{y}_{cc}\\ 0 & \text{otherwise}\end{cases} \tag{14}$$

and $\hat{y}_{cc}$ represents the class that would be assigned based purely on the Frobenious norm using the C-Core attention vectors. Focusing on the terms of Equation 11, we see the loss is large when the disparity between $p(\hat{y}|x_i)$ and $p(y|x_i)$ is large. Let us consider two cases, a perfect model (one that can fully predict the underlying ground truth) and a random model (one that assigns random labels for any input). All other scenarios lead to a model contained in the search space of models we seek to optimize. Assuming the perfect model produces $p(\hat{y}|x_i)$ approximately equal to the ground truth label distribution, then the loss changes based on the noise rate of the oracle. At noise rate $r = 0$, the loss $L \approx 0$, and as $r \rightarrow 1 : L \rightarrow \infty$.

Comparatively, the smoothed loss in Equation 13 first reduces the confidence placed on the oracle's labels by a small percentage ϵ, and then encourages reliance on the C-Core attention vectors by reducing the additional loss term by a factor of 1 - ϵ whenever the predicted class by the model agrees with the C-Core attention vector based assignment. Assuming a perfect model again, in this case at noise rate $r = 0$ the first term in the loss $L \approx 0$, while the second term is always $\epsilon \log p(\hat{y} = \hat{y}_{cc}|x_i)$ due to the indicator variable I. Looking at this differently, the smoothed loss uses both the oracle's labels and the C-Core vectors as the ground. As $r \rightarrow 1$, a larger proportion of the oracle's labels is incorrect, meaning the first term of the loss will explode to infinity as is the case with cross-entropy, while the second term is independent of the noise rate and heavily contributes to the loss for deviations from the C-Core attention vectors based assignment. This means at high label noise rates, the smoothed loss leads to weight updates with reduced influence from noisy labels and thus provides more robust learning.

Now assuming a random model, for low noise rates $r = 0$, the oracle's labels $p(y|x_i)$ are correct, and $p(\hat{y}|x_i)$ is a vector with entries $\frac{1}{C}$. The cross-entropy loss $L \rightarrow -\log\frac{1}{C}$, and as $r \rightarrow 1 : L \rightarrow -\log\frac{1}{C}$. The smoothed loss has the same properties as the cross-entropy loss in low- and high-label noise rate settings when the model randomly assigns labels. If we start with a weak model trained on a clean random set, label smoothing yields no gains under either low or high noise. On the other hand, starting with a strong model and thus representative C-Core attention vectors, we have shown that label smoothing reduces the adverse effects of significant label noise rates using our custom loss function. The next section presents the experimental setup, the trained models, the AL strategies tested, the data sets used, and the training configurations.

4. Experimental setup

In this section, we describe the experimental setup that forms a high-dimensional grid of different configurations in AL for image classification in the presence of label noise. We vary several DL architectures (4), AL algorithms (6), benchmark datasets (4), and the Oracle label noise rates (4). Two of the DL models used in this work are CNN-based while the other two are ViT-based. All CNN-based models are trained in an AL setting with 5 of the 6 AL strategies and all datasets over all the 4 label noise rates. The 6th AL strategy is GCI-ViTAL and is unique to ViTs. The ViT models in this work are trained under all six AL strategies, all datasets, and label noise settings.

4.1 *Deep learning models*

For all four DL architectures in this work, the weights are transferred from the pre-training of the model on the ImageNet-1k dataset [48]. We then fine-tune the fully connected layer for classification. The CNN-based models used in this work are: ResNet34 [56] and VGG19 [48, 57], chosen for their popularity in image classification benchmarks as well as good performance. The ViT-based models of choice in this work are the base ViT with 14 non-overlapping 16 by 16 patches and 12 attention heads, and the Swin transformer [39], which implements overlapping shifted patches as opposed to a grid of rigid patches. In Table 1 we show the model sizes in megabytes, as well as the number of frozen and trainable parameters used in our experiments.

Table 1. Brief description of the models

Model	Size (MB)	Trainable parameters	Frozen parameters	Trainable/Frozen	Feature Extractor
ResNet34	44.7	5,130	11.2 M	0.0458%	CNN
VGG19	548	40,970	139.6 M	0.0294%	CNN
ViT base	330	7,690	85.7 M	0.0089%	Transformer
Swin transformer	336	10,250	86.91 M	0.0118%	Transformer

4.2 *Active learning algorithms*

We compare the following standard DAL query strategies: random query, information entropy-based selection, margin sampling, hybrid uncertainty and diversity, model delta, and ours, GCI-ViTAL. Due to the large number of experiments as well as training time, we limit AL strategies to the above six. We briefly explain each method below:

- **Random Query:** This is the simplest query strategy, while also relatively effective. This AL strategy simply selects candidate images for labeling with equal probability from the unlabeled dataset. It does not take into account any information about the unlabeled data except its size, so it is likely to select samples that are not very informative for the model.
- **Information Entropy-Based Selection:** This query strategy selects samples with the highest information entropy, which is a measure of uncertainty. The information entropy is high when the model is not confident about the predicted class of a sample. Selecting samples with high information entropy can help the model learn class boundaries, hence improving its performance.

- **Margin Sampling:** This query strategy selects samples with the smallest margin between the predicted scores of the two most likely classes. The margin is a measure of the confidence of the model in its prediction. Selecting samples with small margins helps label samples that better define decision boundaries of very similar classes.
- **Hybrid Uncertainty and Diversity:** This query strategy selects samples in a way that balances uncertainty and diversity from the labeled data. The uncertainty is measured by the information entropy of the sample, and diversity is measured by the distance of the sample to the labeled data in pixel space.
- **Model Delta:** This query strategy selects samples that are most likely to change the model's predictions if they were labeled. The model delta is calculated by comparing the predicted scores of two consecutive model states, that is the previously trained model at time *t*-1 and the model at time *t*. Samples with the highest change in the class probability distribution are selected for labeling.
- **GCI-ViTAL (Ours):** This strategy uses the final transformer block's output to pick samples that the model struggles with and that deviate semantically from their supposed class average. This potentially means the points are a special case, most likely close to a decision boundary, and thus worth obtaining a label for from the oracle. This strategy has the advantage of considering the semantical similarities of images while making sample selections, and not only relying on the trained model, which is prone to adverse effects due to label noise.

4.3 *Datasets*

Due to the lack of benchmarks on DAL with label noise using multiple datasets for image classification, we run experiments on the following datasets: Chest X-ray Images (Pneumonia) [58], Food101 [59], CIFAR10 and CIFAR100 [60]. Table 2 below is a brief summary of each dataset. Figures 6 and 7 present sample image examples from each dataset.

Table 2. Summary of core dataset attributes

Dataset	Year	# Images	Classes	Image Sizes	Short description
Chest X-ray (Pneumonia)	2018	5,863	2	224 × 224	Chest X-ray images used for pneumonia detection.
CIFAR10	2009	60,000	10	32 × 32	Small color images of objects from 10 common object classes.
CIFAR100	2009	60,000	100	32 × 32	Extending CIFAR10, 100 classes grouped into 20 super-classes like vehicles, animals, and flowers, with finer sub-classes.
Food101	2004	101,000	101	512 × 512	Images of food items categorized into 101 different classes.

For all the datasets, we use 66%, 17%, 17% train, validation, and test splits as pre-split in the CIFAR 10 and 100 datasets. All of these data splits are kept mutually exclusive to each other. The process for injecting label noise and simulating an oracle is as follows: First, we put aside a clean test set that is the same size as what is commonly used for each dataset. We also draw a clean random sample, without replacement, of 1,024 images and their labels from the training set (independent of the number of classes), that will be used as an initial clean set for AL as is customary. The training set labels are then corrupted with four levels of noise $r \in \{0, 0.2, 0.4, 0.6\}$, by replacing the actual label of a sample with a randomly selected class, where each class has an equal probability of being selected. Once class-independent label noise has been injected, we use the number of training samples for validation that correspond to 17% of the total dataset. On the AL training scheme, no fixed labeling budget is set; we train and evaluate the models on the test set after each batch of AL selection and oracle labeling until the entire training set has been used. This has the advantage of allowing for analysis of model performance in both low-budget and high-budget AL regimes in the context of a noisy oracle. This means we never stop the AL algorithm based on the labeling budget in developing the best algorithm. We, however, monitor test performance across labeling budget ticks to be able to compare models and AL query strategies for different datasets and noise labels on varying budgets.

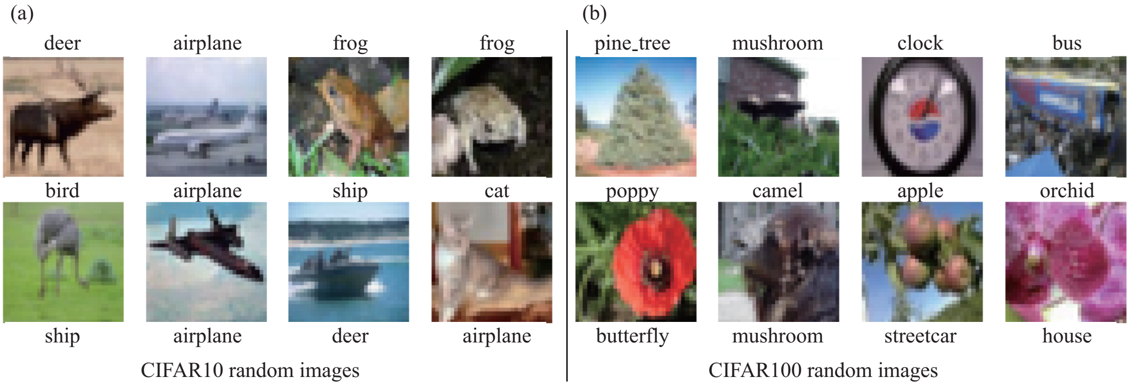


Figure 6. Random images from CIFAR10 (first four columns) and CIFAR100, both (32 × 32)

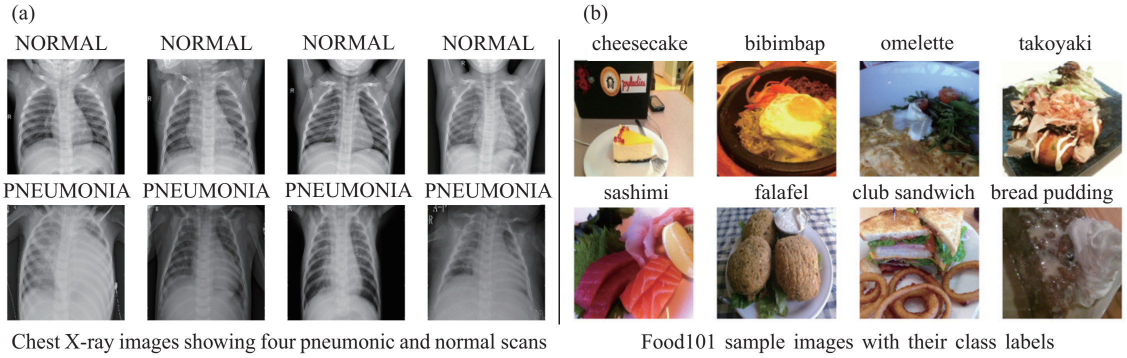


Figure 7. Random images from the Chest X-ray dataset (224 × 224), (first four columns), and the more complex Food101 dataset (512 × 512)

4.4 *Training configuration*

While the training procedure is straightforward in a theoretical framework, the implementation details contain non-trivial components. For one, training in TensorFlow allows for more control of the training loop and thus makes it easier to incorporate AL and noise injection. However, this comes with a disadvantage when it comes to training efficiency on a Graphics Processing Unit (GPU). While PyTorch makes training on a GPU simple with data loaders, the process complicates the AL portion of the cycle as the indices that come out of the data loader do not directly correspond to the indices in the dataset as all operations take and spit out tensors of size *batch_size*, each represented by one index value. We incur extra compute in remapping the indices to match the AL format. We make use of PyTorch's distributed data parallelism and train on two NVIDIA RTX A4000 and two RTX A5000 GPUs, each with approximately 16 Gigabytes of Random Access Memory (RAM), based on availability on one compute node.

In training both CNN and ViT models, a learning rate scheduler is used, that reduces the learning rate by 10% after 10 epochs of no improvement on the validation set. We also train with the early stopping of tolerance 5 epochs, meaning we halt training if the model hits a running 5 epoch validation loss maximum. This indicates that the model is beginning to overfit the training data. In all datasets and AL strategies, training stopped after a median of 15 epochs (CIFAR10: 13 ± 2, CIFAR100: 15 ± 1, Chest X-ray: 14 ± 3, Food101: 15 ± 2). We select samples individually and not the best batch for AL algorithms that are not batch-based. We use the cross-entropy loss across all models except on the GCI-ViTAL AL strategy where we use a C-Core attention vector label smoothed cross-entropy loss. We use the Adam optimizer for ResNet and ViT. Stochastic gradient descent is the recommended optimizer for Visual Geometry Group 19-layer network (VGG19) based on the original paper and the choice in this work. The image transformations such as random cropping, image resizing, and pixel normalization we use for each model are those used by the authors of the respective models on the datasets we focus on in this work. This is sometimes necessary as the model architecture asserts a specific image size, such as is the case with base ViT, which expects images of size 224. Images were resized to 32 by 32 for the

CNN-based models and resized to 224 by 224 for the ViT models. For both CNN and ViT networks, training halts when all training samples are used, but we monitor the AL result after each labeling round.

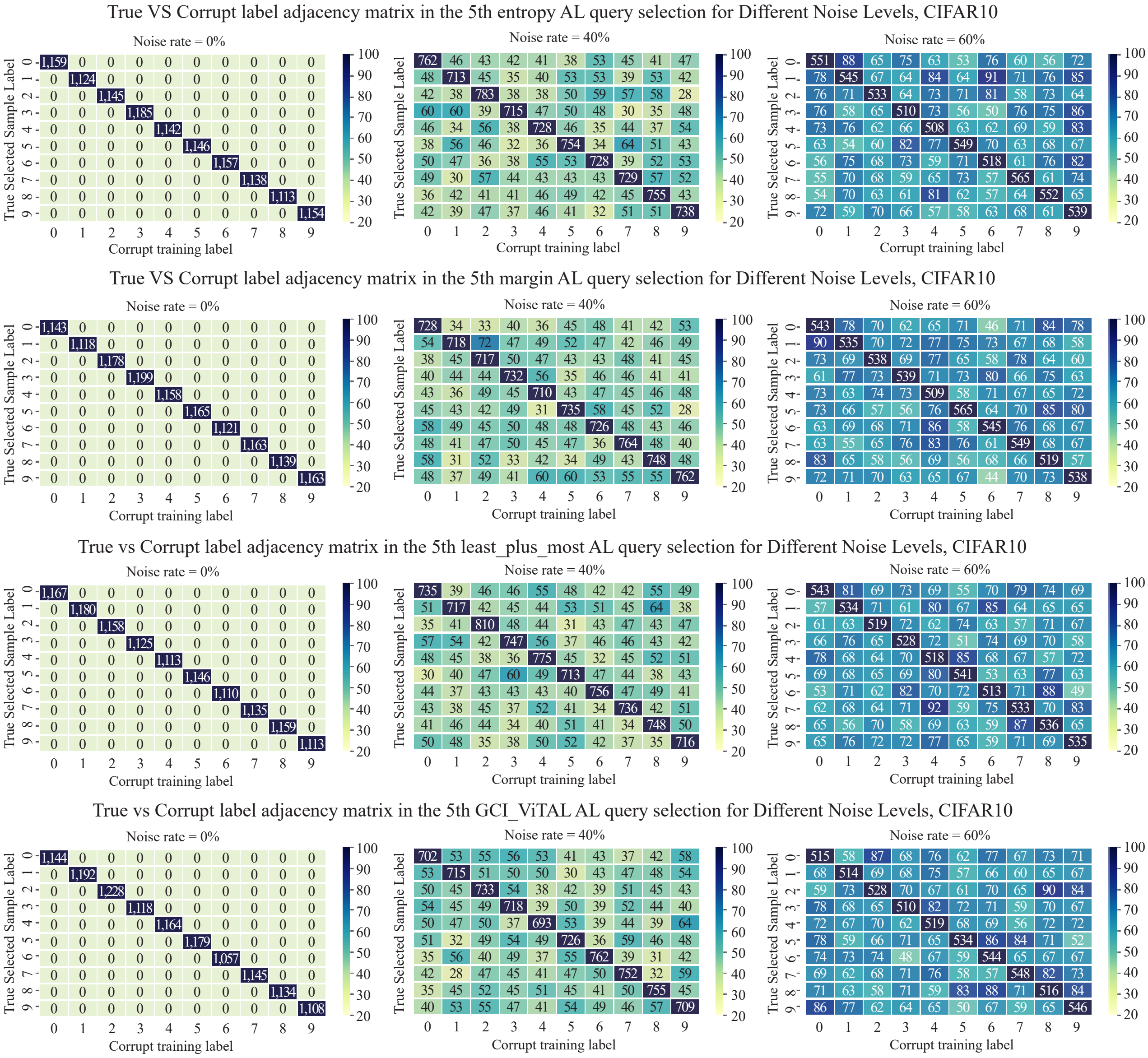


Figure 8. Label adjacency matrices across AL strategies and noise rates (CIFAR10)

Figures 8 and 9 present the adjacency matrix as well as real to corrupt label distribution in a given AL query selection cycle on CIFAR10 and CIFAR100 respectively. These figures demonstrate that we inject the same uniform label noise irrespective of the AL strategy, and thus the difference in performance will be driven mostly by the informativeness of the samples for most AL strategies, but the C-Core smoothed loss function as well in the case of GCI-ViTAL.

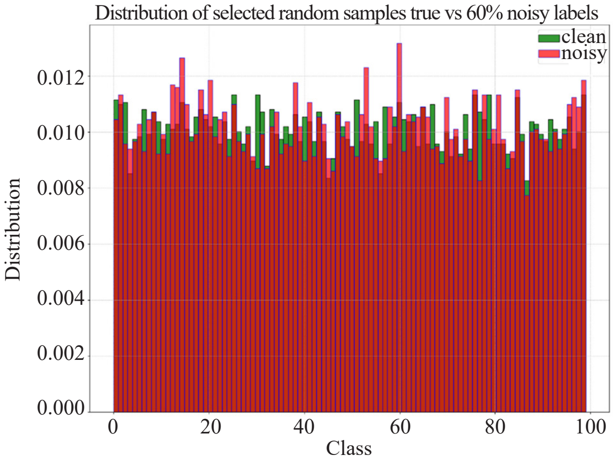


Figure 9. CIFAR100 clean vs. noisy label distribution at 60% noise rate

5. Results

In this section, we present the test results of different DL models trained on CIFAR-10, CIFAR-100, Food-101, and the Chest X-ray (Pneumonia) dataset across four label-noise levels using six DAL algorithms including our proposed GCI-ViTAL. In summary, transformers consistently outperform CNNs, with the gap widening as the number of classes grows. We find that simple **Random Query** performs surprisingly well under label noise, often matching or exceeding uncertainty-based methods. **GCI-ViTAL** achieves the highest generalization performance under high label noise rates (40-60%). Test performance against labeling budget plots can be found in Figures 10-17.

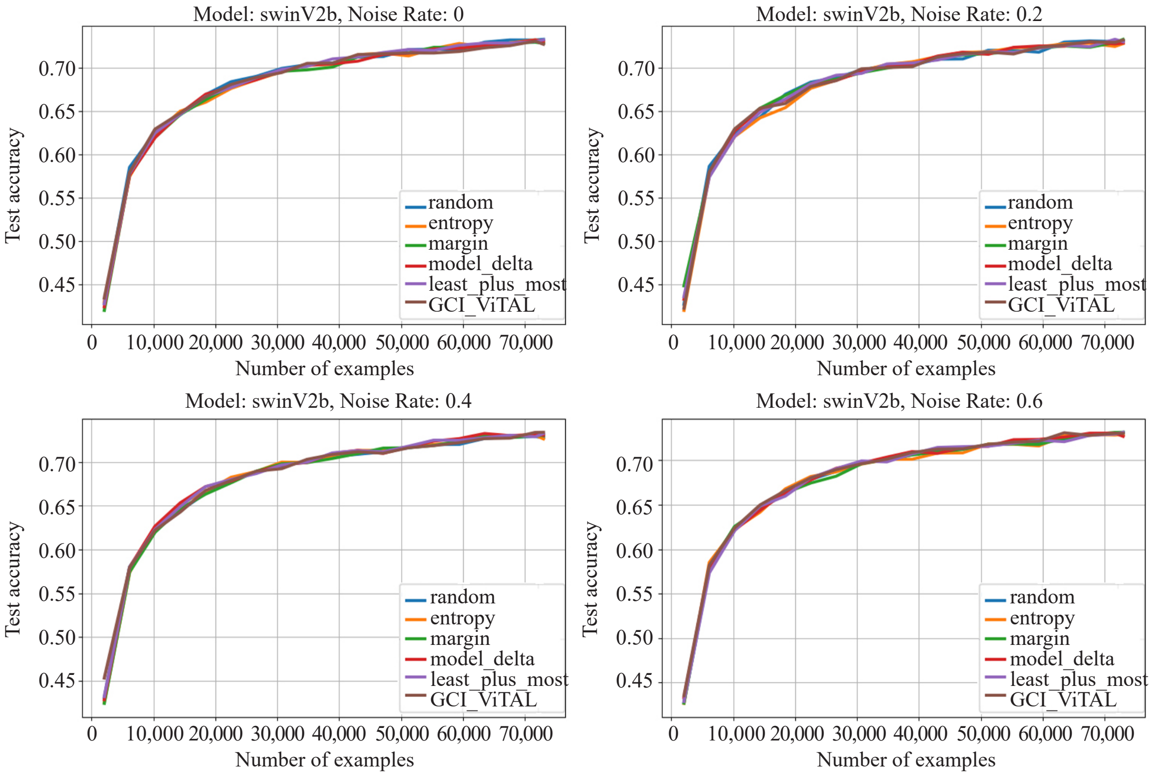


Figure 10. Test performances of Swin transformer on the more complex Food101 dataset

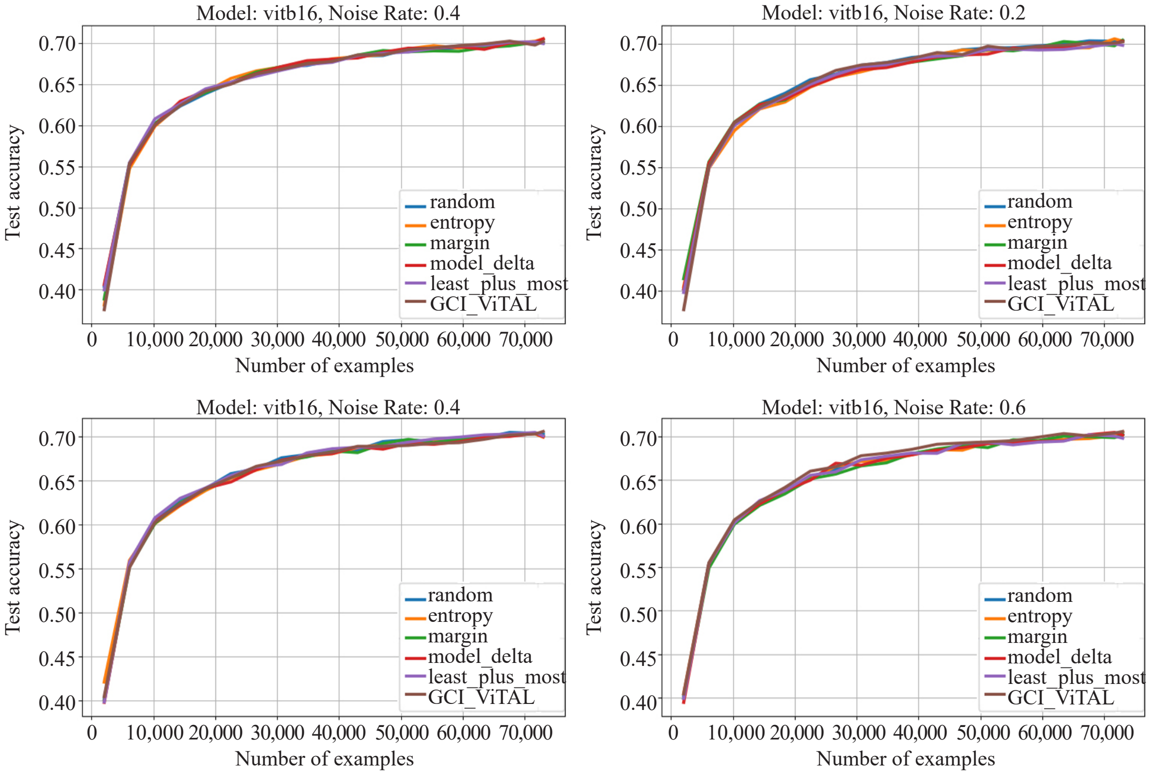


Figure 11. Test performances of ViT transformer on the more complex Food101 dataset

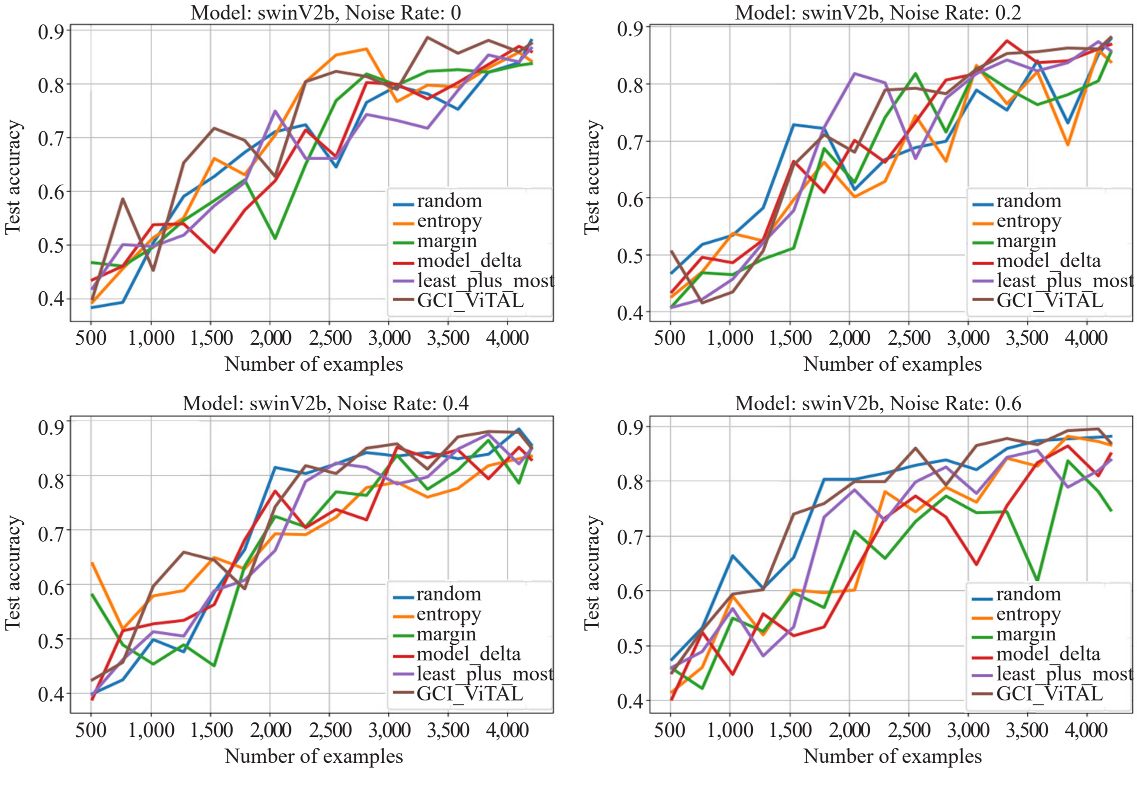


Figure 12. Test performances of Swin transformer on the Chest X-ray dataset

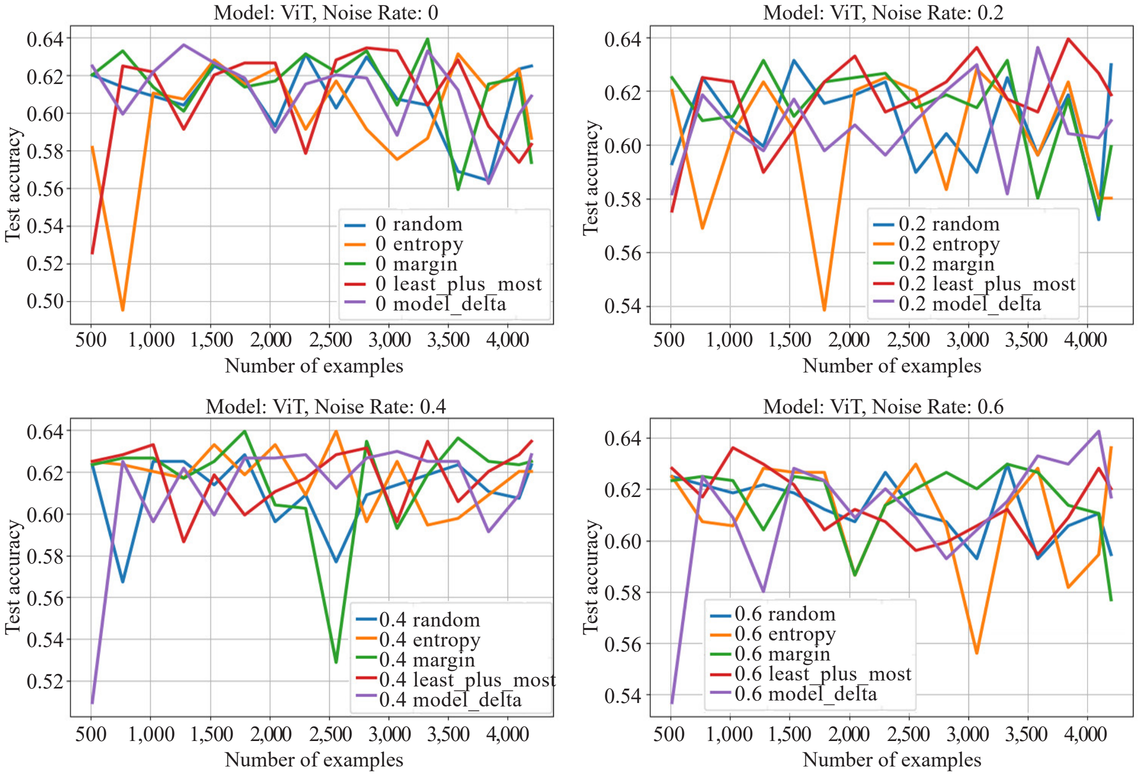


Figure 13. Test performances of ViT transformer on the Chest X-ray dataset

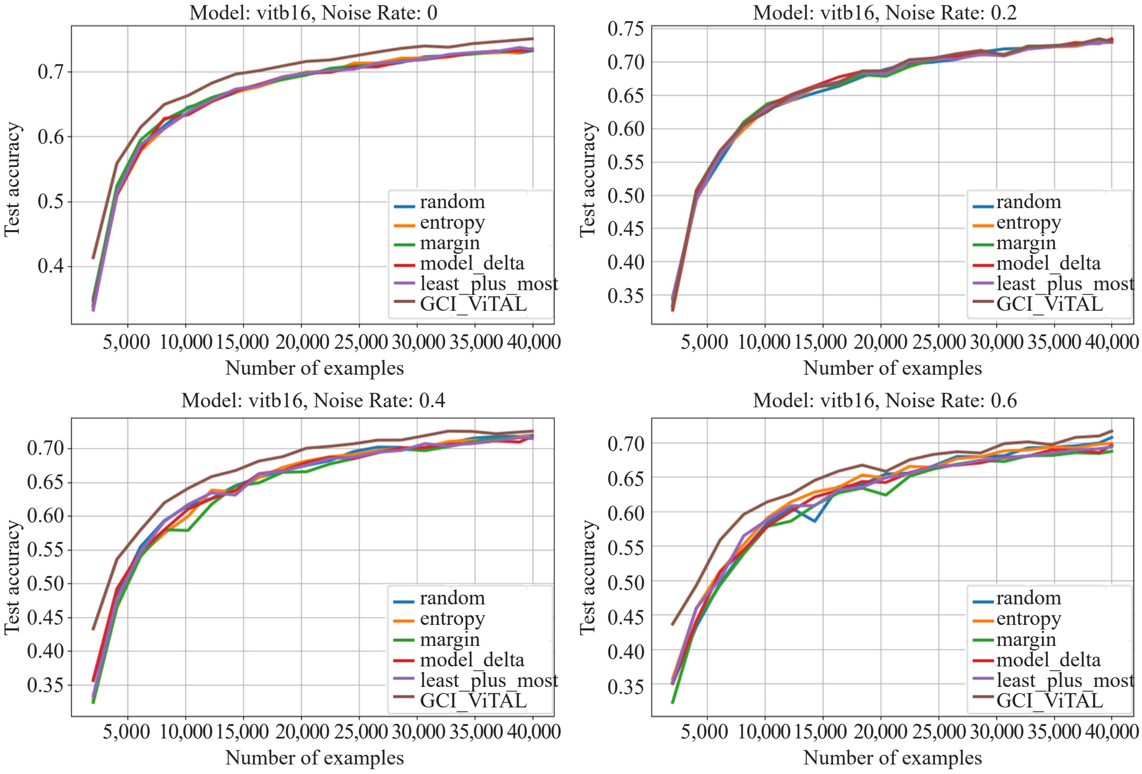


Figure 14. Test performances of ViT on CIFAR100

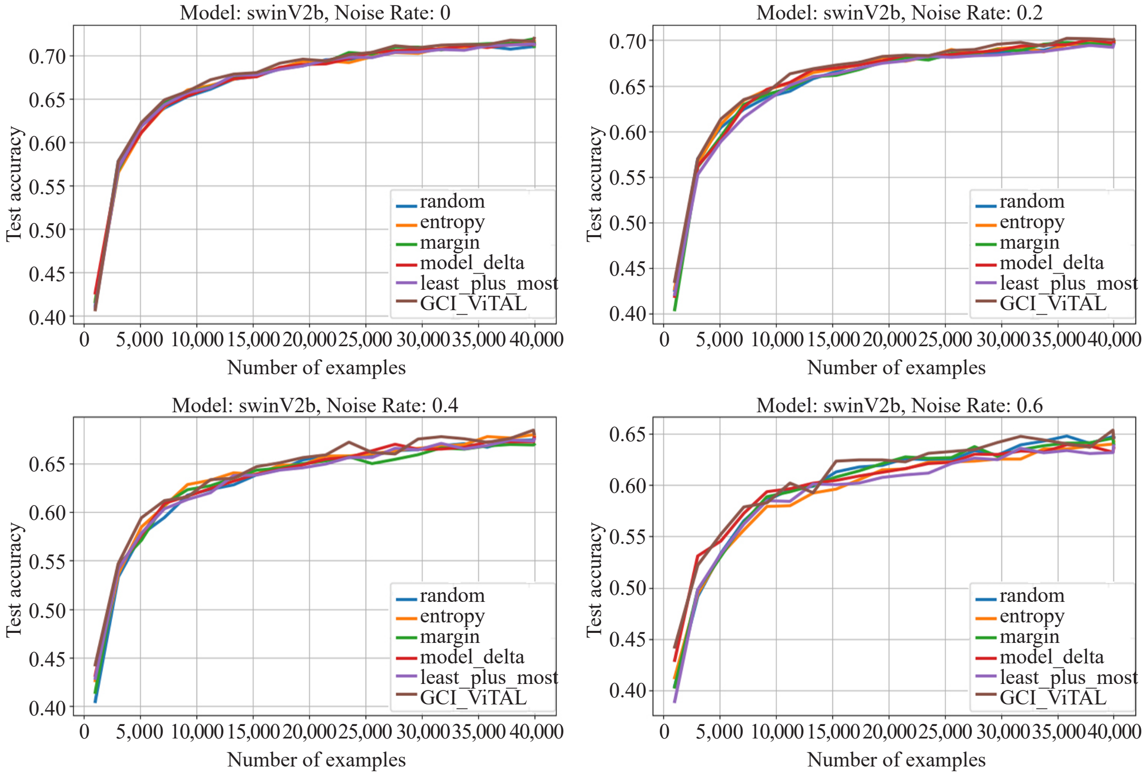


Figure 15. Test performances of Swin transformer on CIFAR100

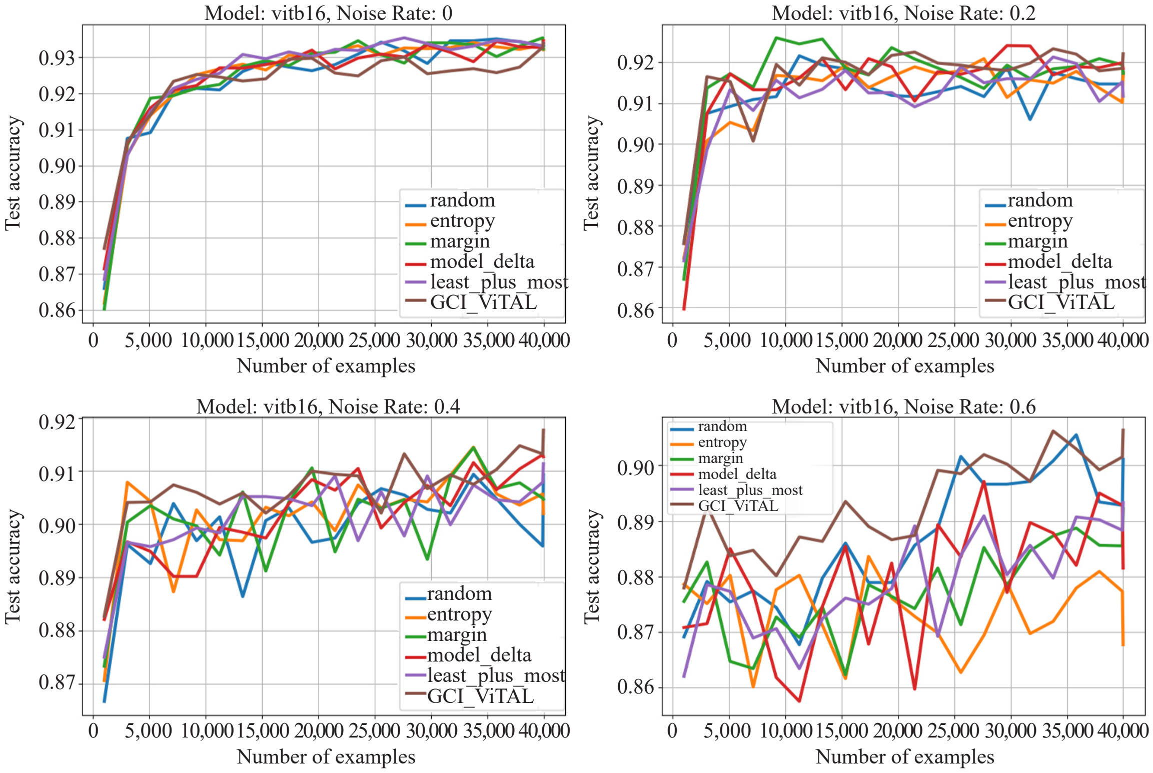


Figure 16. Test results of ViT transformer on CIFAR10 showing each AL strategy under different label noise rates

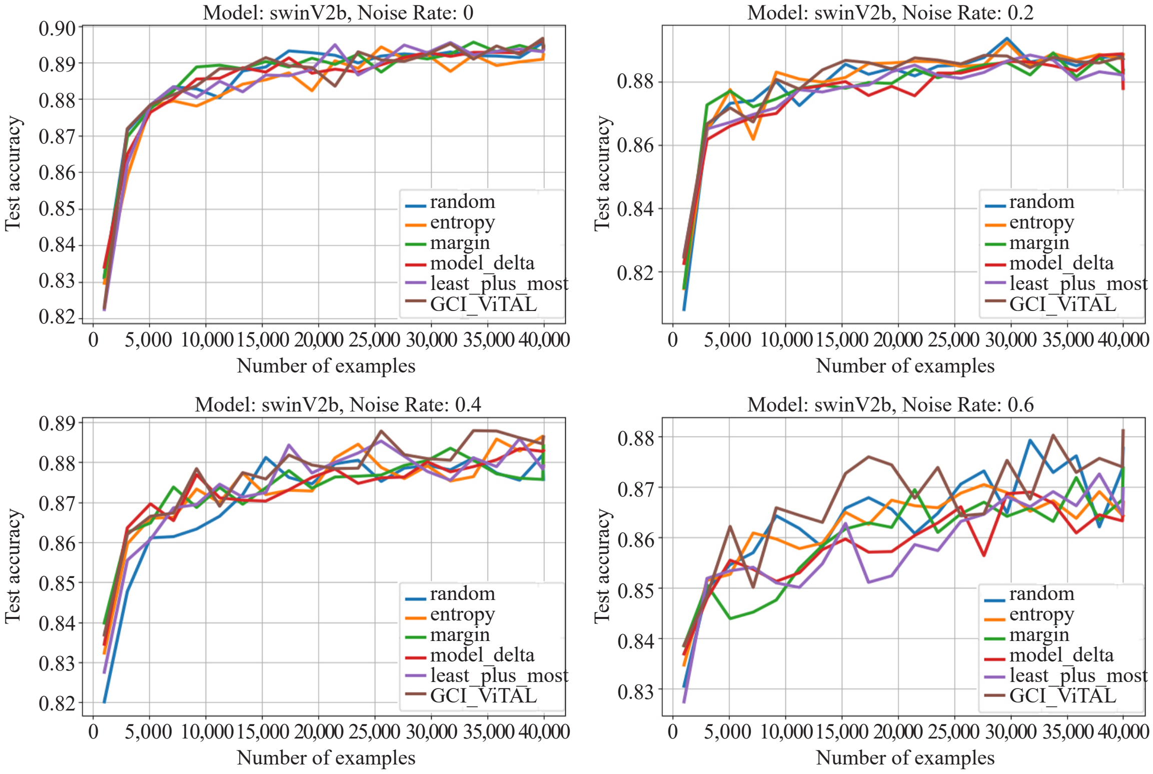


Figure 17. Swin transformer results on CIFAR10

- The first finding in our study that has support in the literature for non-active learning image classification is the superiority of ViTs over CNN-based models in label noise DAL. We observe that for three of the datasets in Tables 3, 4, and 5 in this study, transformer-based models (ViTb16 and SwinV2b16) performed significantly better than CNN-based models (ResNet34 and VGG19) out of the box, across all AL strategies and label noise rates. The performance difference is more pronounced in CIFAR100. CIFAR100 has 10 times more class labels while maintaining the same number of training examples as CIFAR10, while the Chest X-ray dataset presents complexity in the input image since the images are of much higher resolution than CIFAR10 as seen in Table 2, and requires expert opinion to tell positive and negative classes apart. There is very little difference in the performance of the DAL methods on the Food101 dataset across label noise rates.

- With regards to the DAL strategies across datasets, it is worth noting that, contrary to most literature, and in direct support of the findings in the following works [1, 3, 61], we find that random query selection, while simple to implement and computationally efficient, posts comparable or superior classification accuracy results in DAL with label noise compared to margin, entropy, hybrid, and model delta-based selection when out-of-the-box DL models are used without extensive model hyper-parameter optimization. We find this to be especially true for higher label noise rates. Our results show that sample selection that is only guided by a model trained on partially noisy labels selects non-optimal samples compared to random selection. This is not the case for GCI-ViTAL since the C-Core attention component of the AL query strategy does not get updated with noisy information once trained on the clean initial subset of the data, leading to noise resilience.

Table 3. CIFAR10 test accuracy results after 40,000 training samples

Noise rate	Model	Active learning strategy					
		Random	Entropy	Margin	Hybrid	Model delta	GCI-ViTAL
0	VGG19	0.8278	0.8288	0.8273	0.8370	0.8287	-
	ResNet34	0.8132	0.8319	0.8398	0.8348	0.8328	-
	ViTb16	0.9330	0.9320	0.9322	0.9338	0.9345	0.9338
	SwinV2b16	0.8946	0.8922	0.8945	0.8936	0.8936	0.8939
0.2	VGG19	0.7821	0.7564	0.7575	0.7433	0.7137	-
	ResNet34	0.7843	0.7496	0.7426	0.7499	0.7278	-
	ViTb16	0.9176	0.9174	0.9173	0.9117	0.9185	0.9286
	SwinV2b16	0.8828	0.8887	0.8837	0.8808	0.8779	0.8830
0.4	VGG19	0.7406	0.7280	0.6646	0.6651	0.6296	-
	ResNet34	0.7485	0.6691	0.6968	0.6496	0.6009	-
	ViTb16	0.9093	0.9020	0.9047	0.9114	0.9127	0.9152
	SwinV2b16	0.8757	0.8789	0.8841	0.8791	0.8829	0.8874
0.6	VGG19	0.7063	0.6450	0.6737	0.6612	0.4912	-
	ResNet34	0.7041	0.6234	0.6592	0.5575	0.4477	-
	ViTb16	0.9011	0.8677	0.8922	0.8933	0.8815	0.9117
	SwinV2b16	0.8777	0.8643	0.8738	0.8697	0.8670	0.8850

Table 4. CIFAR100 test accuracy after 40,000 training samples

Noise rate	Model	Active learning strategy					
		Random	Entropy	Margin	Hybrid	Model delta	GCI-ViTAL
0	VGG19	0.5691	0.5657	0.5720	0.5675	0.5605	-
	ResNet34	0.5698	0.5621	0.5657	0.5646	0.5637	-
	ViTb16	0.7327	0.7359	0.7357	0.7347	0.7339	0.7504
	SwinV2b16	0.7146	0.7168	0.7106	0.7130	0.7130	0.7349
0.2	VGG19	0.5116	0.4605	0.4588	0.4612	0.4495	-
	ResNet34	0.5135	0.4582	0.4711	0.4867	0.4499	-
	ViTb16	0.7316	0.7287	0.7316	0.7287	0.7341	0.7318
	SwinV2b16	0.6973	0.6978	0.6964	0.6951	0.6986	0.7064
0.4	VGG19	0.4595	0.3911	0.3899	0.3719	0.3400	-
	ResNet34	0.4664	0.3795	0.3679	0.3833	0.3359	-
	ViTb16	0.7194	0.7180	0.7146	0.7160	0.7177	0.7325
	SwinV2b16	0.6771	0.6786	0.6698	0.6749	0.6767	0.6800
0.6	VGG19	0.4084	0.3009	0.2930	0.3246	0.2204	-
	ResNet34	0.3946	0.3097	0.3196	0.3639	0.2210	-
	ViTb16	0.7075	0.6989	0.6873	0.6936	0.6959	0.7215
	SwinV2b16	0.6468	0.6426	0.6403	0.6389	0.6361	0.6441

Table 5. Chest X-ray test accuracy after 40,000 training samples

Noise rate	Model	Active learning strategy					
		Random	Entropy	Margin	Hybrid	Model delta	GCI-ViTAL
0	VGG19	0.6250	0.3189	0.4183	0.2724	0.5128	-
	ResNet	0.4359	0.2115	0.2115	0.3237	0.6138	-
	ViTb16	0.6250	0.5865	0.5737	0.5833	0.6089	0.6109
	SwinV2b16	0.8798	0.8413	0.8365	0.8654	0.8589	0.8620
0.2	VGG19	0.4359	0.2772	0.1778	0.3589	0.5513	-
	ResNet	0.5321	0.4070	0.3429	0.3605	0.6250	-
	ViTb16	0.6298	0.5801	0.5993	0.6186	0.6089	0.6275
	SwinV2b16	0.8782	0.8381	0.8542	0.8574	0.8685	0.8807
0.4	VGG19	0.4439	0.2051	0.5352	0.1939	0.2901	-
	ResNet	0.5465	0.3701	0.6266	0.4535	0.6233	-
	ViTb16	0.6234	0.6201	0.6250	0.6346	0.6282	0.6368
	SwinV2b16	0.8558	0.8349	0.8509	0.8477	0.8285	0.8510
0.6	VGG19	0.5208	0.3766	0.5144	0.5496	0.4936	-
	ResNet	0.3878	0.3766	0.2212	0.3253	0.4262	-
	ViTb16	0.5946	0.6362	0.5769	0.6201	0.6169	0.6273
	SwinV2b16	0.8814	0.8654	0.7468	0.8381	0.8494	0.8661

- Unlike other AL strategies, Model-delta selects samples based on the change in model confidence between consecutive updates. Samples with a large model confidence change are selected for labeling. We find that model delta AL performs better than entropy, margin, and hybrid in no-label noise settings, and tends to perform worse in increasing label noise settings. High label noise causes early discrepancies in predicted probabilities across iterations, risking model collapse. This behavior only persists in CNN-based models, i.e., model delta performs on par with all other AL strategies for zero label noise, but up to 45% worse at 60% label noise rate. Model delta however performs on par with other baseline AL strategies even for high label noise rates when the learner is a transformer-based model. This sheds more light on the robustness and high calibration of transformer-based models.
- GCI-ViTAL's higher generalization performance at high label noise rates comes at a higher computational cost in attaining the samples to be labeled as well as computing its custom loss function that is based on the C-Core attention assignment for each sample. We found that GCI-ViTAL runs as slow as model delta and approximately 2 times slower than entropy-based selection, margin, and hybrid uncertainty & diversity selection. The difference in total runtime increases in the case of GCI-ViTAL since for each new retrained model, the attention maps are computed over the entire unlabeled dataset and stored for computing the Frobenius norm to the C-Core vectors and selecting the next training samples until the labeling budget is exhausted. Table 6 shows the average runtime for each DAL algorithm per AL cycle. The random query strategy has the fastest runtime as expected since it does not depend on the input images nor the predicted probabilities from the currently trained model. However, we see a benefit in using GCI-ViTAL in the high-label noise regime with a large label budget.
- We empirically observe that GCI-ViTAL's robustness under label noise is more pronounced with the increasing

number of classes in the classification problem. On CIFAR100 and Food101, ViT with GCI-ViTAL performs equal or marginally better than all other AL strategies under 40% and 60% label noise as can be seen in Tables 4, 7 and Figures 12, 13, 16, and 17 as opposed to performance on CIFAR10 with 10 classes shown in Figures 10 and 11, as well as in the Chest X-ray dataset with just 2 classes shown in Figures 14, and 15.

We hypothesize this is due to label smoothing guided by the C-Core Frobenius norm in the following manner: As the label noise rate increases up to 60% in a 100-class problem, it means each sample selected for labeling is 60% likely to be labeled erroneously as one of the other 99 classes. This leads to incorrect weight updates and thus large training errors. However, with GCI-ViTAL, the loss function is partially grounded in the small clean label distribution to redistribute the oracle's labeling in such a manner it agrees with the pre-trained model's semantical assignment of images to classes based on the learned representations from clean pre-training. Recall from Section 4.3.2 that while the loss is designed to tackle label noise, the C-Core assignment is not affected by label noise as we only train the MLP layer of the ViT on noisy labels, and not the feature extraction multi-head self-attention layers.

Table 6. CIFAR10 average AL selection times per cycle, excluding training time

Algorithms	Average run time (seconds) per AL cycle
Random query	0.593
Margin sampling	87.42
Entropy-based selection	87.33
Hybrid uncertainty and diversity	87.02
Model delta	164.3
GCI-ViTAL (ours)	163.6

Table 7. Food101 test accuracy after 73,000 training samples for ViTb16 and SwinV2b16

Noise rate	Model	Active learning strategy					
		Random	Entropy	Margin	Hybrid	Model delta	GCI-ViTAL
0	ViTb16	0.7062	0.7053	0.7004	0.6996	0.7055	0.7022
	SwinV2b16	0.7329	0.7327	0.7299	0.7325	0.7326	0.7312
0.2	ViTb16	0.7038	0.7035	0.7046	0.6983	0.7035	0.7025
	SwinV2b16	0.7330	0.7286	0.7331	0.7302	0.7332	0.7324
0.4	ViTb16	0.7034	0.6993	0.7006	0.7006	0.6998	0.7082
	SwinV2b16	0.7326	0.7304	0.7305	0.7325	0.7324	0.7339
0.6	ViTb16	0.7066	0.7055	0.7020	0.6981	0.7017	0.7023
	SwinV2b16	0.7320	0.7293	0.7317	0.7307	0.7312	0.7326

6. Discussions

We find that GCI-ViTAL matches or slightly outperforms standard AL methods at low noise levels and significantly outperforms them under high noise (e.g., 60% on CIFAR-100). This stems from its built-in pseudo-memory of C-Core attention vectors, which both identify outliers and downweight unreliable oracle labels. By leveraging multi-head attention maps pretrained on clean ImageNet data, and avoiding heavy hyperparameter tuning, we ensure a fair and robust comparison under label noise.

As the label noise rate increases from 0% up to 60%, the test accuracy for all models and AL strategies declines considerably for CIFAR10 and CIFAR100, and only marginally for the Chest X-ray and Food101 datasets. Still, the decline in performance is steeper for the CNN-based models as compared to their transformer-based counterparts. One reason for this could be that the pre-training in the transformer captures adequate local dependencies while at the same time learning rich global dependencies between pixels so that fine-tuning on noisy labels has a less detrimental effect on the overall probability outputs of the model. This is not the case with CNN-based models that have an inductive bias to prioritize local dependencies in explaining the differences in class labels, thus making the transformer more robust to label noise than CNN-based models [62-64]. To contrast the two in the face of high label noise, CNNs learn a smoother decision boundary between classes while transformers learn a more complex decision boundary [9, 65-66]. This is because ViTs are larger models with considerably more parameters than CNNs as shown in Table 1, and more parameters allow for a more fine-grained complex decision boundary between classes. The more parameters of transformers allow for the decision boundary to change due to label noise but only change so insignificantly that the change does not lead to the misclassification of many neighboring samples. However we note that the ViT is more robust to label noise in most cases that require complex decision boundaries since it has a smaller trainable-to-frozen parameters ratio, meaning it is forced to condense a lot more of the signal into useful weights, thus it is less likely to overfit to noisy training data as compared to CNN-based models. This leads us to the idea of aiming for larger frozen parameter models for feature extraction and low trainable parameters as future work.

While the use of ViTs in AL with label noise shows promising results, the high computational cost is a disadvantage that needs to be addressed. Possible ways of addressing this include deploying ViTs with fewer layers or reducing model size through methods such as quantization [67, 68] and model distillation [69-71]. We are also interested in exploring how ViT models of different sizes (small, base, and large) are impacted by label noise. This can be especially important in navigating performance and computational complexity trade-offs. The more fundamental question resulting from this work is: in the context of high label noise, at what point do the AL and few-shot learning paradigms converge, what are the factors, and what role can weak or self-supervised learning play in improving generalization and robustness? Last but not least, in the wake of the advances in Large Language Models (LLMs) as well as multi-modal learning, how can we leverage LLMs to produce text-guided AL strategies, guide the oracle through caption generations, and create explainable ViT applications in the label noise domain? These ideas are partially inspired by the following works [72-74].

6.1 *Limitations*

We note that while this work addresses the use of out-of-the-box DL models to give a good reflection of realistic AL baselines, it lacks in providing a comprehensive comparison of the proposed method (GCI-ViTAL) to state-of-the-art AL methods with label noise. The work also does not address or show results of GCI-ViTAL on more complex real-world datasets outside the datasets covered due to time and resource constraints. This study also lacks a comprehensive ablation on the value of λ in Equation 8.

7. Conclusion

In conclusion, this study provides valuable insights into the performance of various deep learning models trained and tested on CIFAR10, CIFAR100, Food101, and the Chest X-ray (Pneumonia) datasets, particularly in the context of AL with label noise. We found that transformer-based models, such as ViT and Swin Transformer, consistently outperformed CNN-based models across all datasets and AL strategies, indicating their superiority in handling complex image classification tasks. We attribute ViT's robustness to their ability to capture both local and global dependencies, resulting in more complex and elastic decision boundaries that are less affected by incorrectly labeled training samples. Furthermore, our results challenge conventional wisdom by showing that random query selection often yields superior classification accuracy in AL with label noise compared to more complex AL strategies that rely solely on uncertainty and diversity metrics. Despite these findings, it is worth noting that transformer-based models incur higher computational costs compared to CNN-based models. GCI-ViTAL, our proposed AL strategy, exhibits higher generalization performance at high label noise rates, albeit with increased computational cost. This strategy leverages the inherent robustness of transformer-based models by making use of the semantic relationships of images without an oracle's labels.

This work opens up several avenues for future research. Exploring methods to mitigate the computational overhead of transformer-based models, such as deploying smaller ViT models or applying model quantization and distillation techniques, could be a promising direction. We also hope to investigate the effect of asymmetric label noise on the different AL strategies using different transformer variants. Additionally, investigating the convergence of AL and few-shot learning paradigms in the context of high-label noise, as well as exploring the role of weak or self-supervised learning to improve generalization in label noise scenarios could provide further insights into enhancing model performance. We are also interested in investigating how ViT model size impacts DAL under label noise. In summary, our study contributes to the current understanding of DL models' behavior under DAL scenarios with label noise and suggests potential avenues for addressing challenges in this domain.

Acknowledgements

We acknowledge technical support and advanced computing resources from the University of Hawaii Information Technology Services-Research Cyberinfrastructure, funded in part by the National Science Foundation CC* awards #2201428 and #2232862. We also thank the anonymous peer reviewers for their insightful feedback and suggestions, which greatly improved the quality of this manuscript.

Conflict of interest

The authors declare that they have no conflict of interest.

References

[1] Ren P, Xiao Y, Chang X, Huang P, Li Z, Gupta BB, et al. A survey of deep active learning. *ACM Computing Surveys*. 2020; 54(9): 1-40.

[2] Algan G, Ulusoy I. Image classification with deep learning in the presence of noisy labels: A survey. *Knowledge-Based Systems*. 2021; 215: 106771.

[3] Mots'oehli M, Baek K. Deep active learning in the presence of label noise: A survey. *arXiv:2302.11075*. 2023. Available from: https://doi.org/10.48550/arXiv.2302.11075.

[4] Zhang C, Bengio S, Hardt M, Recht B, Vinyals O. Understanding deep learning requires rethinking generalization. *Communications of the ACM*. 2016; 64(3): 107-115.

[5] Malach E, Shalev-Shwartz S. Decoupling "When to Update" from "How to Update". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. Red Hook, NY, USA: Curran Associates Inc.; 2017. p.961-971.

[6] Han B, Yao Q, Yu X, Niu G, Xu M, Hu W, et al. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18)*. Red Hook, NY, USA: Curran Associates Inc.; 2018. p.8536-8546.

[7] Liu S, Niles-Weed J, Razavian N, Fernandez-Granda C. Early-learning regularization prevents memorization of noisy labels. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS'20)*. Red Hook, NY, USA: Curran Associates Inc.; 2020. p.20331-20342.

[8] Liang X, Liu X, Yao L. Review-a survey of learning from noisy labels. *ECS Sensors Plus*. 2022; 1(2): 021401.

[9] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai XH, Unterthiner T, et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. In: *9th International Conference on Learning Representations (ICLR)*. Brookline, MA, USA: JMLR.org; 2021.

[10] Touvron H, Cord M, Sablayrolles A, Synnaeve G, Jégou H. Going deeper with image transformers. In: *2021 IEEE/CVF International Conference on Computer Vision*. USA: IEEE; 2021. p.32-42.

[11] Tal R, Ben-Baruch E, Noy A, Zelnik L. ImageNet-21K pretraining for the masses. In: Vanschoren J, Yeung S. (eds.). *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*. 2021. p.1-12.

[12] Yuan K, Guo S, Liu Z, Zhou A, Yu F, Wu W. Incorporating convolution designs into visual transformers. In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Montreal, QC, Canada: IEEE; 2021. p.559-568.

[13] Górriz M, Carlier A, Faure E, Nieto XG. Cost-effective active learning for melanoma segmentation. *arXiv:1711.09168*. 2017. Available from: https://doi.org/10.48550/arXiv.1711.09168.

[14] Konyushkova K, Sznitman R, Fua P. Learning active learning from data. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Red Hook: Curran Associates Inc.; 2017. p.4228-4238.

[15] Kremer J, Stensbo-Smidt K, Gieseke F, Pedersen KS, Igel C. Big universe, big data: Machine learning and image analysis for astronomy. *IEEE Intelligent Systems*. 2017; 32(2): 16-22.

[16] Carena F, Carena W, Chapeland S, Chibante Barroso V, Costa F, Dénes E, et al. The ALICE data acquisition system. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*. 2014; 741: 130-162.

[17] Gutleber J, Murray S, Orsini L. Towards a homogeneous architecture for high-energy physics data acquisition systems. *Computer Physics Communications*. 2003; 153(2): 155-163.

[18] Zhai X, Oliver A, Kolesnikov A, Beyer L. S4L: Self-supervised semi-supervised learning. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. Seoul: IEEE; 2019. p.1476-1485.

[19] Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In: *Proceedings of the 37th International Conference on Machine Learning*. Brookline, MA, USA: JMLR.org; 2020. p.1597-1607.

[20] Chen X, Xie S, He K. An empirical study of training self-supervised vision transformers. In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Montreal: IEEE; 2021. p.9620-9629.

[21] He K, Chen X, Xie S, Li Y, Dollár P, Girshick R. Masked autoencoders are scalable vision learners. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New Orleans: IEEE; 2022. p.15979-15988.

[22] Assran M, Duval Q, Misra I, Bojanowski P, Vincent P, Rabbat M, et al. Self-supervised learning from images with a joint-embedding predictive architecture. In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Vancouver: IEEE; 2023. p.15619-15629.

[23] Lewis D, Gale W. A sequential algorithm for training text classifiers. In: *Proceedings of the 17th annual international ACM SIGIR Conference on Research and Development in Information Retrieval*. Dublin: Springer; 1994. p.3-12.

[24] Cao X, Tsang I. Bayesian active learning by disagreements: A geometric perspective. *arXiv:2105.02543*. 2021. Available from: https://doi.org/10.48550/arXiv.2105.02543.

[25] Gal Y, Islam R, Ghahramani Z. Deep Bayesian active learning with image data. In: *Proceedings of the 34th International Conference on Machine Learning (ICML'17)*. Brookline, MA, USA: JMLR.org; 2017. p.1183-1192.

[26] Sener O, Savarese S. Active learning for convolutional neural networks: A core-set approach. In: *6th International Conference on Learning Representations (ICLR)*. Vancouver; 2018.

[27] Har-Peled S, Roth D, Zimak D. Maximum margin coresets for active and noise tolerant learning. In: *Proceedings of the 20th International Joint Conference on Artifical Intelligence*. San Francisco: Morgan Kaufmann Publishers Inc.; 2007. p.836-841.

[28] Younesian T, Epema DH, Chen L. Active learning for noisy data streams using weak and strong labelers. *arXiv:2010.14149*. 2020. Available from: https://doi.org/10.48550/arXiv.2010.14149.

[29] Wei J, Zhu Z, Cheng H, Liu T, Niu G, Liu Y. Learning with noisy labels revisited: A study using real-world human annotations. *arXiv:2110.12088*. 2022. Available from: https://doi.org/10.48550/arXiv.2110.12088.

[30] Chao G, Zhang K, Wang X, Chu D. Three-teaching: A three-way decision framework to handle noisy labels. *Applied Soft Computing*. 2024; 154: 111400.

[31] Gupta G, Sahu AK, Lin W. Noisy batch active learning with deterministic annealing. *arXiv:1909.12473*. 2020. Available from: https://doi.org/10.48550/arXiv.1909.12473.

[32] Amin K, DeSalvo G, Rostamizadeh A. Learning with labeling induced abstentions. *Advances in Neural Information Processing Systems*. 2021; 34: 12576-12586.

[33] Younesian T, Zhao Z, Ghiassi A, Birke R, Chen L. QActor: Active learning on noisy labels. *Proceedings of the 13th Asian Conference on Machine Learning*. 2021; 157: 548-563.

[34] Huang T, Lihong L, Vartanian A, Amershi S, Zhu XJ. Active learning with oracle epiphany. In: *Advances in Neural Information Processing Systems.* Red Hook, NY, USA: Curran Associates Inc.; 2016. p.2828-2836.

[35] Yan S, Chaudhuri K, Javidi T. Active learning from imperfect labelers. In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. Red Hook: Curran Associates Inc.; 2016. p.2136-2144.

[36] Bernhardt M, Castro DC, Tanno R, Schwaighofer A, Tezcan K, Monteiro M, et al. Active label cleaning for improved dataset quality under resource constraints. *Nature Communications.* 2022; 13(1): 1161.

[37] Sheng M, Sun Z, Chen T, Pang S, Wang Y, Yao Y. Foster adaptivity and balance in learning with noisy labels. In: *Computer Vision-ECCV 2024*. Cham: Springer Nature Switzerland; 2025. p.217-235.

[38] Li Y, Chen M, Liu Y, He D, Xu Q. An empirical study on the efficacy of deep active learning for image classification. *arXiv:2212.03088*. 2022. Available from: https://doi.org/10.48550/arXiv.2212.03088.

[39] Liu H, Cheng G, Lin W, Yang J, Yang J, Zhang H, et al. Swin transformer: Hierarchical vision transformer using shifted windows. In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV).* Montreal, QC, Canada: IEEE; 2021. p.9992-10002.

[40] Han K, Xiao A, Wu E, Guo J, Xu C, Wang Y. Transformer in transformer. In: *Proceedings of the 35th International Conference on Neural Information Processing Systems*. Red Hook: Curran Associates Inc.; 2021. p.15908-15919.

[41] Ding M, Xiao B, Codella N, Luo P, Wang J, Yuan L. DaViT: dual attention vision transformers. In: *Computer Vision-ECCV 2022*. Cham: Springer Nature Switzerland; 2022. p.74-92.

[42] Fan H, Xiong B, Mangalam K, Li Y, Yan Z, Malik J, et al. Multiscale vision transformers. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Montreal, QC, Canada: IEEE; 2021. p.6804-6815.

[43] Caramalau R, Bhattarai B, Kim T. Visual transformer for task-aware active learning. *arXiv:2106.03801*. 2021. Available from: https://doi.org/10.48550/arXiv.2106.03801.

[44] He KL, Gan C, Li ZY, Rekik I, Yin ZH, Ji W, et al. Transformers in medical image analysis. *Intelligent Medicine*. 2023; 3(1): 59-78.

[45] Rotman G, Reichart R. Multi-task active learning for pre-trained transformer-based models. *Transactions of the Association for Computational Linguistics*. 2022; 10: 1209-1228.

[46] Khanal S, Lee H, Smith J. Evaluating the robustness of vision transformers under label noise in medical image classification. In: *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention*; 2024. p.512-521.

[47] Safaei B, Patel VM. Active learning for vision-language models. In: *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. Tucson, AZ, USA: IEEE; 2025. p.4902-4912.

[48] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*. 2015; 115(3): 211-252.

[49] Frobenius G. Ueber lineare substitutionen und bilineare formen. *Journal Für Die Reine und Angewandte Mathematik*. 1877; 84: 1-63.

[50] Kazakos D, Papantoni-Kazakos P. Spectral distance measures between gaussian processes. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing*. Denver, CO, USA: IEEE; 1980. p.612-613.

[51] Kullback S, Leibler RA. On information and sufficiency. *The Annals of Mathematical Statistics*. 1951; 22(1): 79-86.

[52] Guo PC. A frobenius norm regularization method for convolutional kernel tensors in neural networks. *Computational Intelligence and Neuroscience*. 2022; 2022(1): 3277730.

[53] Wu Y, Inkpen D, El-Roby A. Maximum batch frobenius norm for multi-domain text classification. In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech andSignal Processing (ICASSP)*. Singapore: IEEE; 2022. p.3763-3767.

[54] Qiu D, Makur A, Zheng L. Probabilistic clustering using maximal matrix norm couplings. In: *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. Monticello, IL, USA: IEEE Press; 2018. p.1020-1027.

[55] Patro RN, Subudhi S, Biswal PK. Spectral clustering and spatial frobenius norm based jaya optimization for band selection of hyperspectral images. *IET Image Processing*. 2019; 13(2): 307-315.

[56] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *2016 IEEEConference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE; 2016. p.770-778.

[57] Simonyan K, Zisserman A. Very deep convolutional networks forlarge-scale image recognition. *arXiv:1409.1556.* 2015. Available from: https://doi.org/10.48550/arXiv.1409.1556.

[58] Kermany D, Zhang K, Goldbaum M. Labeled optical coherence tomography (OCT) and chest X-ray images for classification. *Mendeley Data.* 2018; 2(2): 651.

[59] Bossard L, Guillaumin M, Van Gool L. Food-101-mining discriminative components with random forests. In: *European Conference on Computer Vision.* Zurich, Switzerland: Springer; 2014.

[60] Krizhevsky A, Nair V, Hinton G. Learning multiple layers of features from tiny images. *CIFAR-100 (Canadian Institute for Advanced Research)*. 2009.

[61] Althammer S, Zuccon G, Hofstätter S, Verberne S, Hanbury A. Annotating data for fine-tuning a neural ranker? current active learning strategies are not better than random selection. In: *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*. NewYork, NY, USA: Association for Computing Machinery; 2023. p.139-149.

[62] Khanal B, Shrestha P, Amgain S, Khanal B, Bhattarai B, Linte CA. Investigating the robustness of vision transformers against label noise in medical image classification*. Annual International Conference*. 2024; 2024: 1-6.

[63] Caron M, Touvron H, Misra I, Jegou H, Mairal J, Bojanowski P, et al. Emerging properties in self-supervised vision transformers. In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Piscataway, NJ, USA: IEEE; 2021. p.9630-9640.

[64] Hendrycks D, Mazeika M, Kadavath S, Song D. Using self-supervised learning can improve model robustness and uncertainty. In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc.; 2019.

[65] Ramachandran P, Parmar N, Vaswani A, Bello I, Levskaya A, Shlens J. Stand-alone self-attentionin vision models. In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc.; 2019. p.68-80.

[66] Cordonnier JB, Loukas A, Jaggi M. On the relationship between self-attention and convolutional layers. *arXiv:1911.03584*. 2019. Available from: https://doi.org/10.48550/arXiv.1911.03584

[67] Rok B, Azarpeyvand A, Khanteymoori A. A comprehensive survey on model quantization for deep neural networks in image classification. *ACM Transactions on Intelligent Systems and Technology*. 2023; 14(6): 1-50.

[68] Feng K, Chen Z, Gao F, Wang Z, Xu L, Lin W. Post-training quantization for vision transformer in transformed domain. In: *2023 IEEE International Conference on Multimedia and Expo (ICME)*. Los Alamitos, CA, USA: IEEE Computer Society; 2023. p.1457-1462.

[69] Hinton GE, Vinyals O, Dean J. Distilling the knowledge in a neural network. *arXiv:1503.02531*. 2015. Available from: https://doi.org/10.48550/arXiv.1503.02531.

[70] Yang Z, Li Z, Zeng A, Li Z, Yuan C, Li Y. ViTKD: Feature-based knowledge distillation for vision transformers. In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Seattle, WA, USA: IEEE; 2024. p.1379-1388.

[71] Das R, Sanghavi S. Understanding self-distillation in the presence of label noise. In: *Proceedings ofthe 40th International Conference on Machine Learning (ICML'23)*. Brookline, MA, USA: JMLR.org; 2023.

[72] Hu P, Peng X, Zhu H, Zhen L, Lin J. Learning cross-modal retrieval with noisy labels. In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Nashville, TN, USA: IEEE; 2021. p.5399-5409.

[73] Wazzan A, MacNeil S, Souvenir R. Comparing traditional and LLM-based search for image geolocation. In: *Proceedings of the 2024 Conference on Human Information Interaction and Retrieval*. New York, NY, USA: Association for Computing Machinery; 2024. p.291-302.

[74] Chen X, Wang X, Changpinyo S, Piergiovanni A, Padlewski P, Salz D, et al. PaLI: A jointly-scaled multilingual language-image model. In: *Proceedings of the eleventh International Conference onLearning Representations (ICLR 2023)*. Brookline, MA, USA: JMLR.org; 2023.