南京航空航天大學
Nanjing University of Aeronautics and Astronautics

Fine-tuning Convolutional Neural Networks for Biomedical Image Analysis: Actively and Incrementally*

Zongwei Zhou[1], Jae Shin[1], Lei Zhang[1], Suryakanth Gurudu[2], Michael Gotway[2], and Jianming Liang[1]

[1]Arizona State University

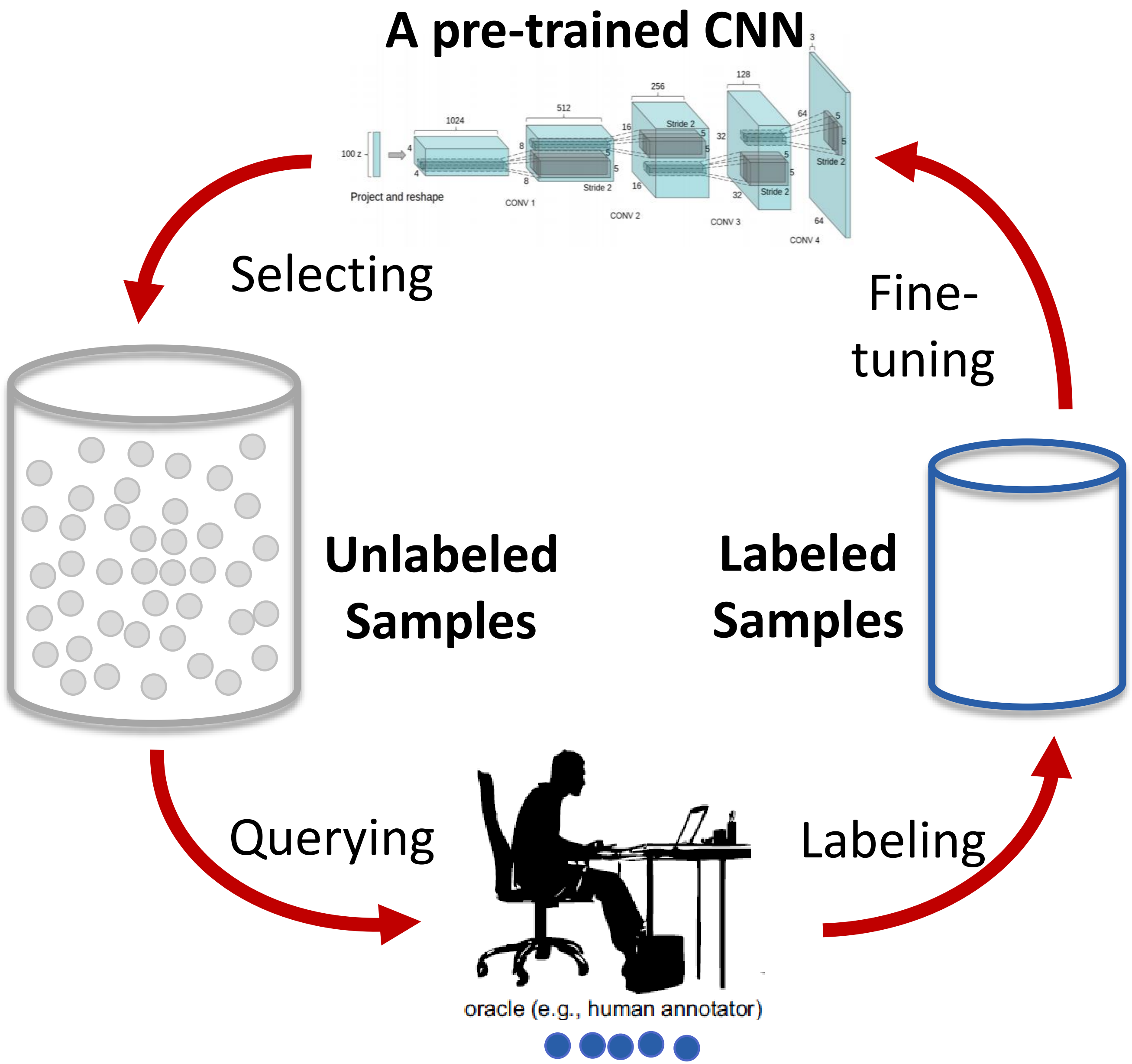
{zongweiz,sejong,lei.zhang.10,jianming.liang}@asu.edu

[2]Mayo Clinic

{gurudu.suryakanth,gotway.michael}@mayo.edu

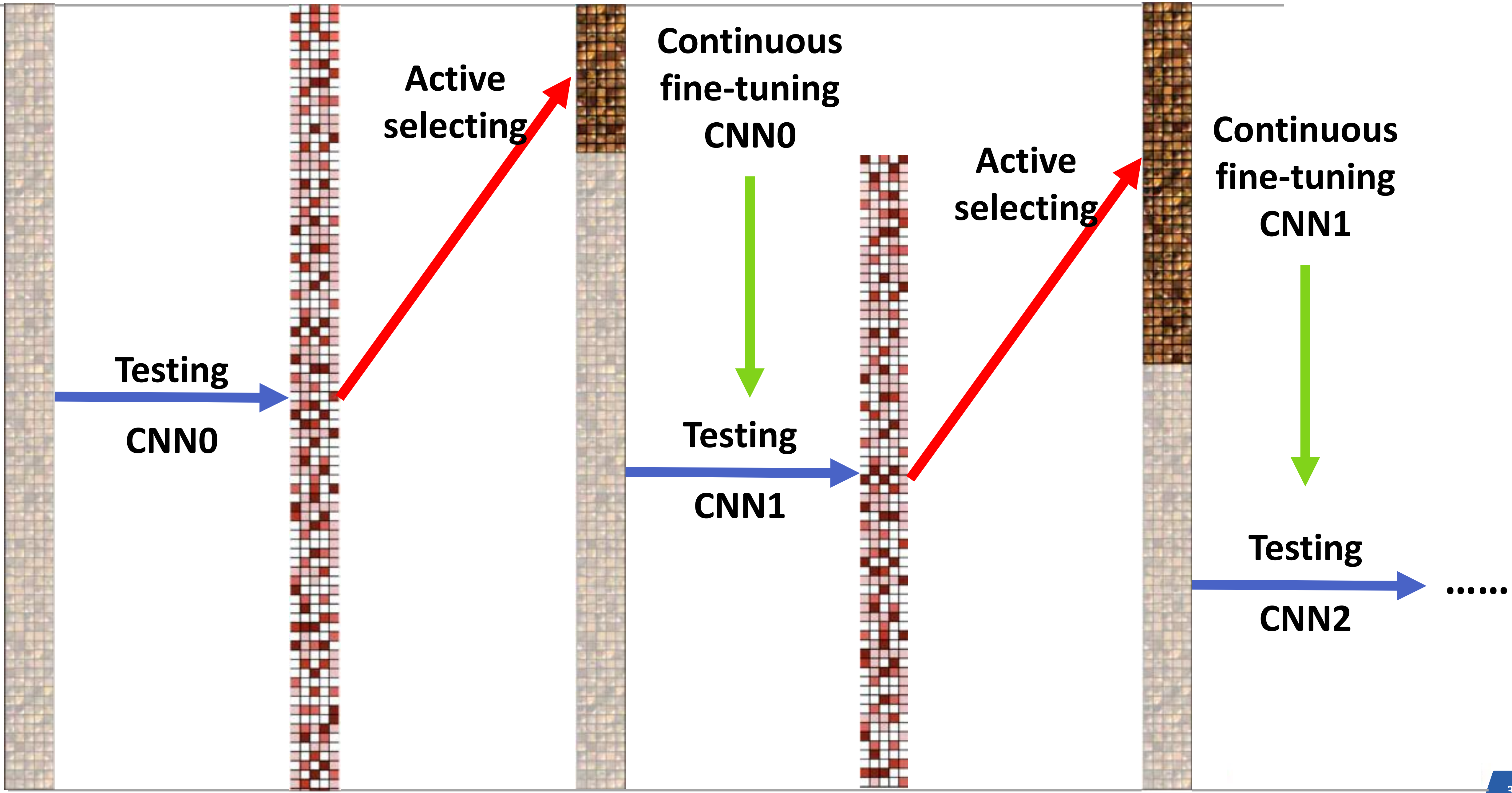
CVPR2017

Framework

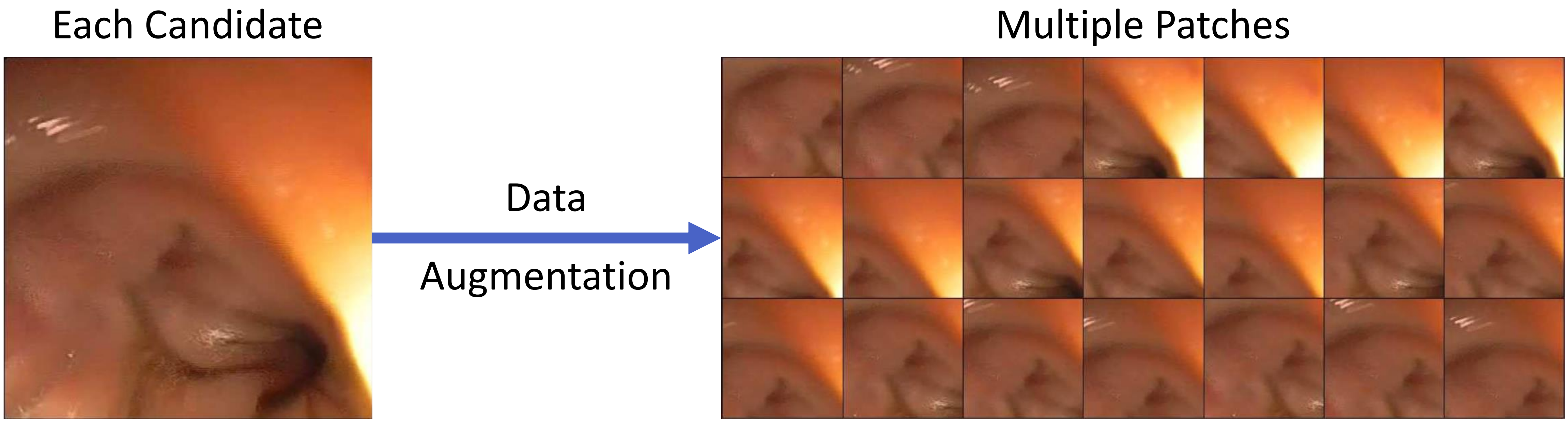


- Starting with a completely empty labeled dataset and a pre-trained CNN.
- The pre-trained CNN seeks worthy samples for annotation.
- Then incorporating newly annotated samples in each iteration to fine-tune the CNN continuously.

An Illustration



How to design query strategy ?



- These patches generated from the same candidate share the same label, and are naturally expected to have similar predictions by current CNN.
- Their entropy and diversity provide a useful indicator to the power of a candidate in elevating the performance of the current CNN.

Active candidate selection

➢ Uncertainty

- Higher uncertainty values denote higher degrees of information.

- Assuming the prediction of patch x_i^j by the current CNN is p_i^j, we define its entropy as:

$$e_i^j = -\sum_{k=1}^{|Y|} p_i^{j,k} \log p_i^{j,k}$$

- Entropy e_i^j denotes the information furnished by patch x_i^j of candidate C_i in the unlabeled pool.

Active candidate selection

➤ Diversity

- Higher diversity values denote higher degrees of prediction inconsistency among the patches within a candidate.

- Diversity between patches x_i^j and x_i^l of candidate C_i as:

$$d_i(j,l) = \sum_{k=1}^{|Y|} (p_i^{j,k} - p_i^{l,k}) \log(\frac{p_i^{j,k}}{p_i^{l,k}})$$

- Diversity $d_i(j,l)$ estimates the amount of information overlap between patches x_i^j and x_i^l of candidate C_i.

How to combine uncertainty and diversity ?

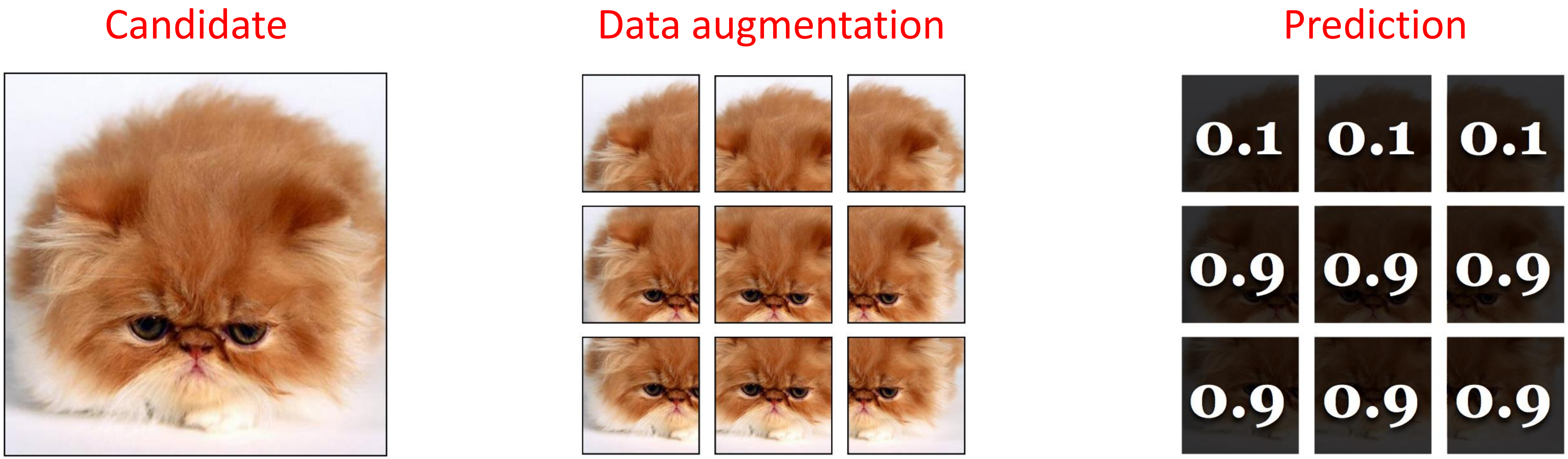
- By definition, all the entries in e_i^j and $d_i(j,l)$ are non-negative.

- Further, $d_i(j,j) = 0, \forall j$

- We combine e_i^j and $d_i(j,l)$ into a signal matrix R_i for each candidate C_i:

$$R_i(j,l) = \begin{cases} \lambda_1 e_i^j & \text{if} \quad j = l, \\ \lambda_2 d_i(j,l) & \text{otherwise} \end{cases}$$

Problem of data augmentation



- Data augmentation inevitably generates hard samples for some candidates injecting noisy labels.

Handling noisy labels via majority selection

Majority selection

- We compute entropy and diversity by selecting only a portion of the patches of each candidate according to the predictions by the current CNN.

- For each candidate C_i compute the average probabilistic predictions of all of its patches:

$$a_i = \frac{1}{m} \sum_{j=1}^{m} p_i^j$$

$$\begin{cases} \text{Select the top } \alpha \text{ percent patches, if } a_i > 0.5 \\ \text{Select the bottom } \alpha \text{ percent patches, otherwise} \end{cases}$$

Algorithm

Algorithm 1: Active incremental fine-tuning method.

Input:
$\mathcal{U} = \{\mathcal{C}_i\}, i \in [1, n]$ $\{\mathcal{U}$ contains n candidates$\}$
$\mathcal{C}_i = \{x_i^j\}, j \in [1, m]$ $\{\mathcal{C}_i$ has m patches$\}$
$\mathcal{M}_0$: pre-trained CNN
b: batch size
α: patch selection ratio

Output:
$\mathcal{L}$: labeled candidates
$\mathcal{M}_t$: fine-tuned CNN model at Iteration t

Functions:
$p \leftarrow P(\mathcal{C}, \mathcal{M})$ $\{$outputs of $\mathcal{M}$ given $\forall x \in \mathcal{C}\}$
$\mathcal{M}_t \leftarrow F(\mathcal{L}, \mathcal{M}_{t-1})$ $\{$fine-tune $\mathcal{M}_{t-1}$ with $\mathcal{L}\}$
$a \leftarrow mean(p_i)$ $\{a = \frac{1}{m}\sum_{j=1}^{m} p_i^j\}$

Initialize:
$\mathcal{L} \leftarrow \varnothing$, t$\leftarrow$ 1

repeat
  for *each* $\mathcal{C}_i \in \mathcal{U}$ **do**
    $p_i \leftarrow P(\mathcal{C}_i, \mathcal{M}_{t-1})$
    if $mean(p_i) > 0.5$ **then**
      $\mathcal{S}_i' \leftarrow$ top α percent of the patches of $\mathcal{C}_i$
    else
      $\mathcal{S}_i' \leftarrow$ bottom α percent of the patches of $\mathcal{C}_i$
    end
    Build matrix R_i using Eq. 3 for $\mathcal{S}_i'$
  end
  Sort $\mathcal{U}$ according to the numerical sum of R_i
  Query labels for top b candidates, yielding $\mathcal{Q}$
  $\mathcal{L} \leftarrow \mathcal{L} \bigcup \mathcal{Q}$; $\quad \mathcal{U} \leftarrow \mathcal{U} \setminus \mathcal{Q}$
  $\mathcal{M}_t \leftarrow F(\mathcal{L}, \mathcal{M}_{t-1})$; $t \leftarrow t + 1$
until *classification performance is satisfactory*;

Handling noisy labels via majority selection

Active candidate selection

Continuous fine-tuning

Experiments

Six variants of AIFT

- AIFT Diversity$^{1/4}$
- AIFT Diversity
- AIFT Entropy$^{1/4}$
- AIFT Entropy
- AIFT (Entropy + Diversity)$^{1/4}$
- AIFT (Entropy + Diversity)

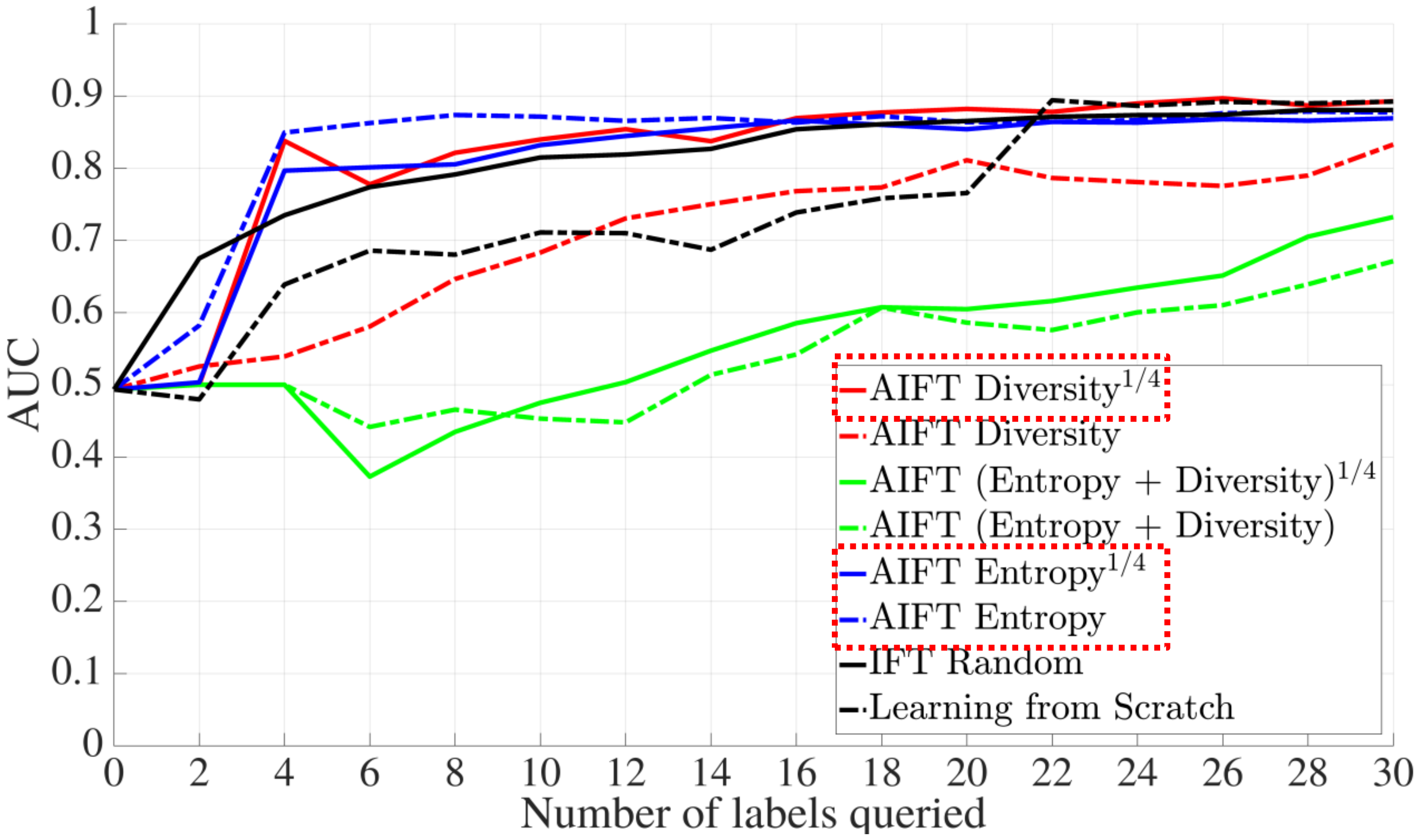
Two compared methods

- IFT Random
- Learning from Scratch

Three applications

- Colonoscopy Frame Classification
- Polyp Detection
- Pulmonary Embolism Detection

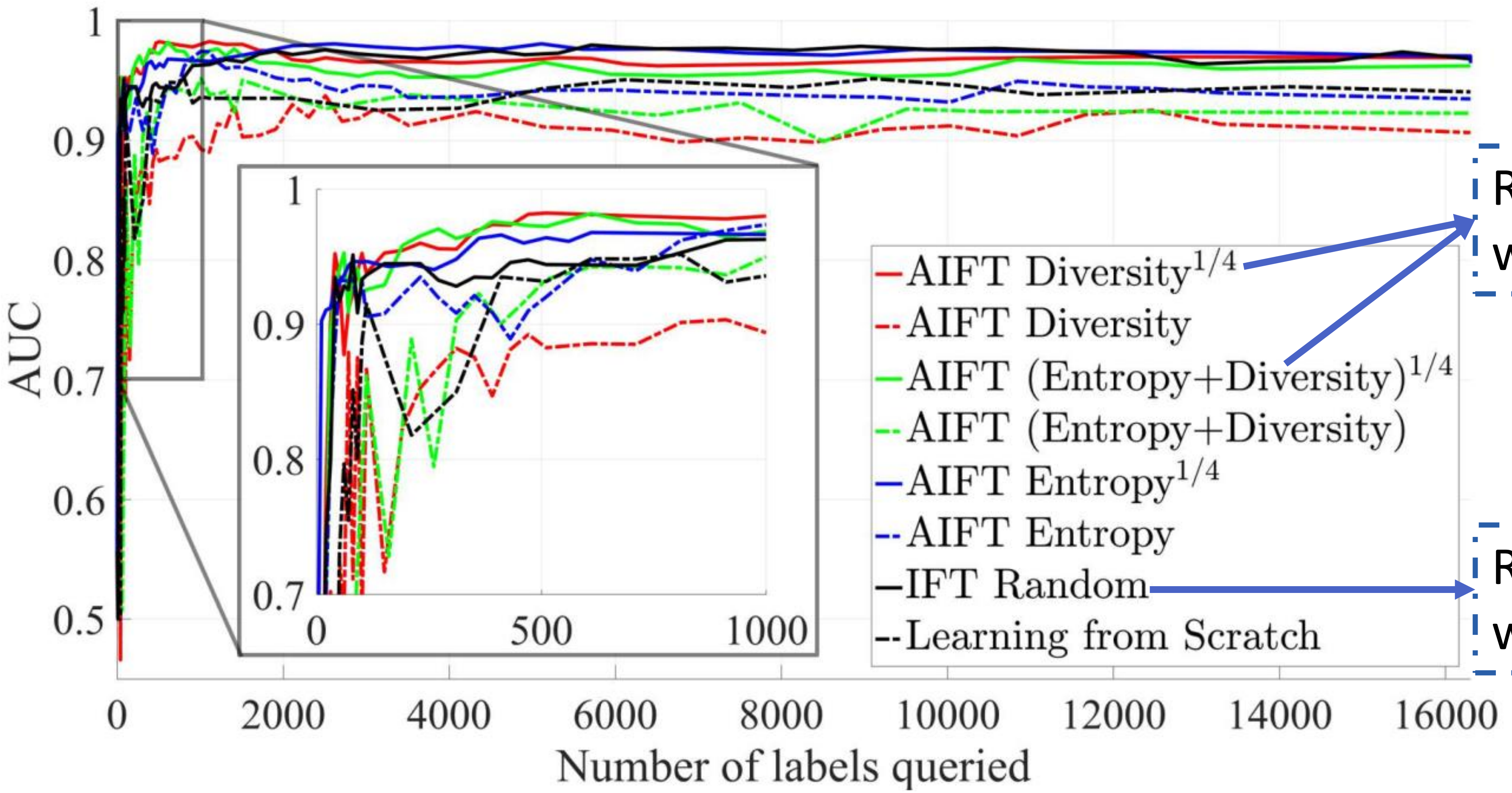
Colonoscopy Frame Classification



- Data: 4000 frames from 6 videos
- Training set: 2000 frames
- Test set: 2000 frames
- Patch: 21 patches for each frame
- Label: informative or non-informative

Comparing 8 methods in colonoscopy frame classification

Polyp Detection



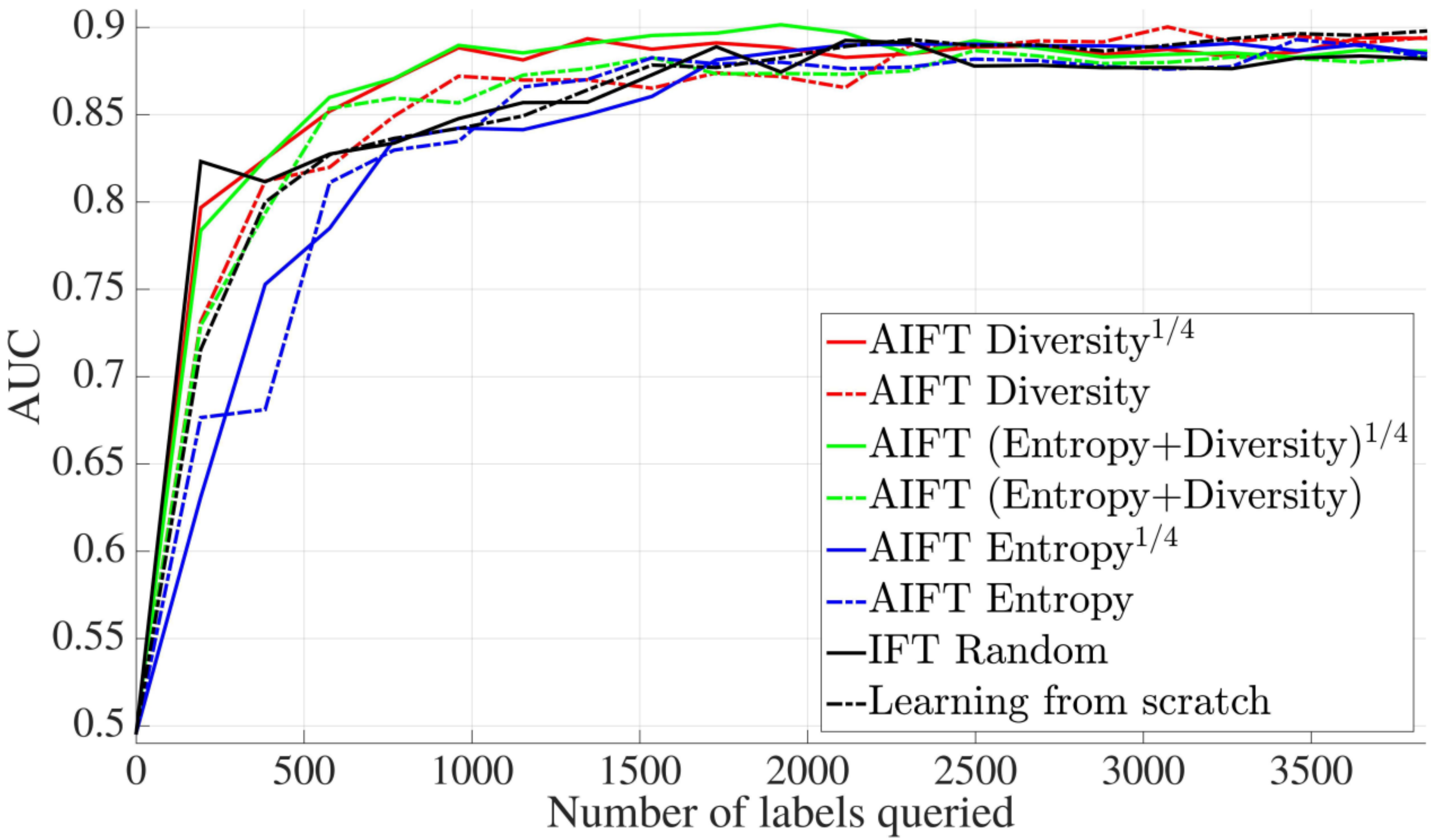
Reach the peak performance with 610 label quires.

Cut 90% annotation cost

Reach the peak performance with 5711 label quires.

It's contributed to the majority selection, which can efficiently select the informative and representative candidates while excluding those with noisy labels.

Pulmonary Embolism Detection



THANKS