Enhancing Instance-level Image Classification with Set-level Labels

Renyu Zhang, Aly A. Khan, Yuxin Chen, Robert L. Grossman

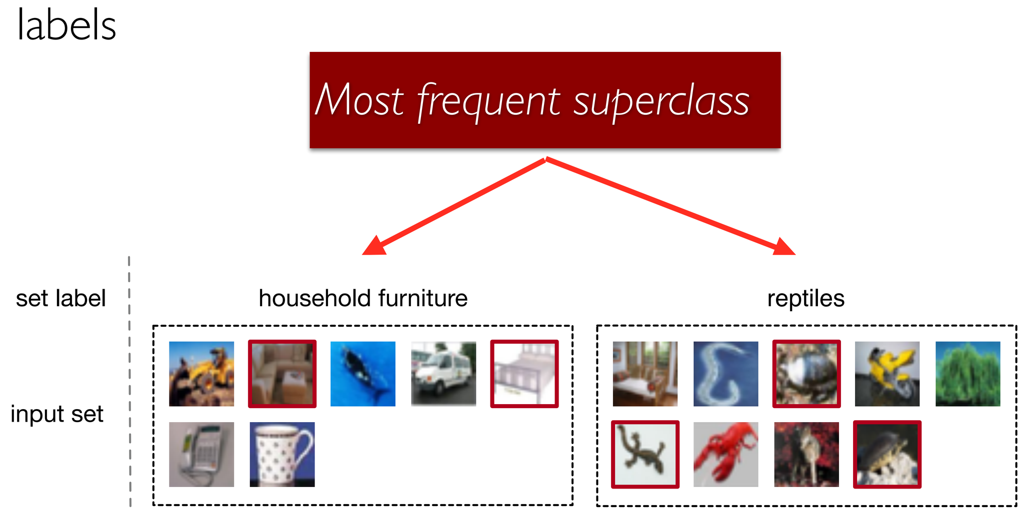
THE UNIVERSITY OF CHICAGO

1 Summary

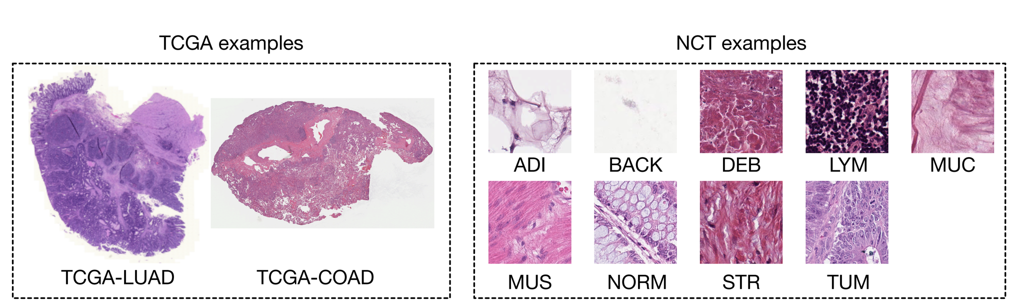
- **FACILE**: an supervised learning algorithm that leverage set-level labels to improve instance-level image classification.
- A theoretical analysis of the proposed method, including recognition of conditions for fast excess risk.
- Experimental studies on two distinct categories of datasets: natural image datasets and histopathology image datasets.

2 Motivation

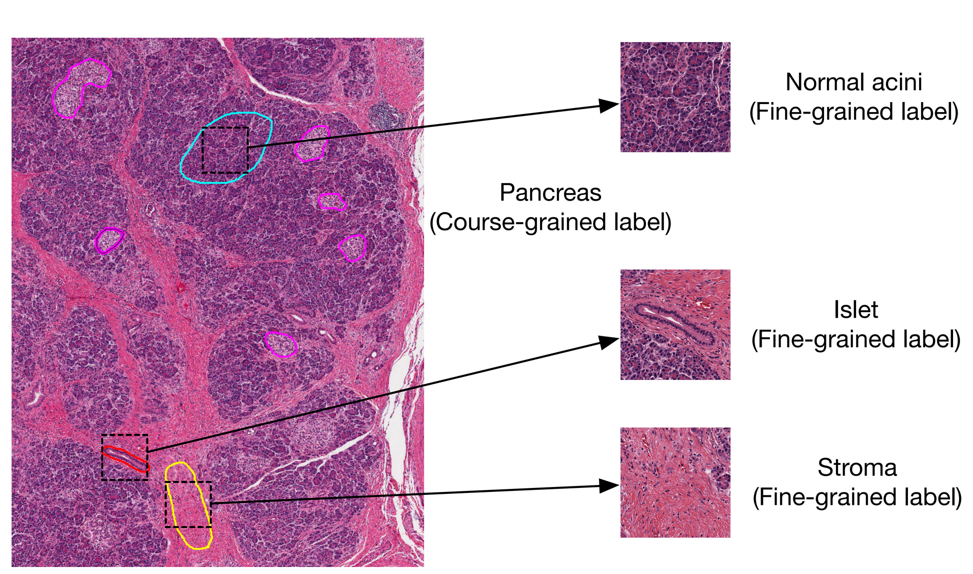
From coarse-grained set-level labels to instance-level labels



CIFAR-100



Whole slide image (WSI) examples from TCGA and patches from NCT dataset are in the lower row.



Hierarchy of course- and fine-grained labels for histopathology images.

3 Notation

- Coarse-grained dataset: $\mathcal{D}_m^{cg} = \{(s_i, w_i)\}_{i=1}^m$
  - s_i: set of instances $\{x_j\}_{j=1}^a$
  - w_i: set-level coarse-grained label
  - ℓ^{fg}: loss on fine-grained labels.
- Fine-grained dataset $\mathcal{D}_n^{fg} = \{(x_i, y_i)\}_{i=1}^n$,
  - y_i: set-level coarse-grained labels.
  - ℓ^{cg}: pretraining loss on coarse-grained labels.
- Instance feature maps $e \in \mathcal{E}$, set-input functions $g \in \mathcal{G}$, and fine-grained label predictors $f \in \mathcal{F}$. The corresponding set-input feature map of an instance feature map e is defined as ϕ^e.

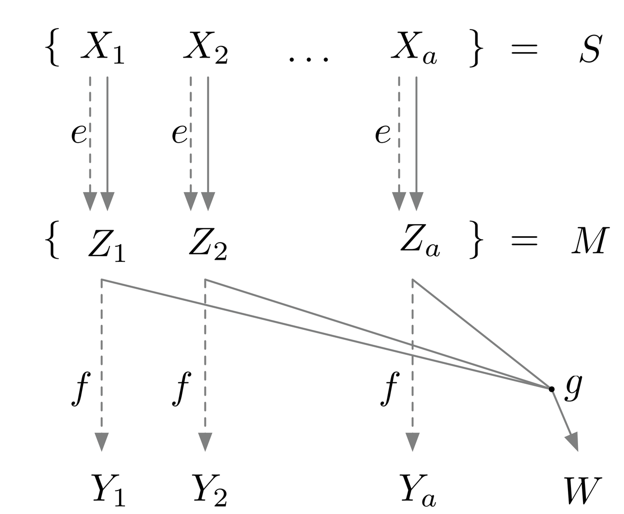
Problem Statement

- Our primary goal is to learn an instance-level predictor $\hat{f} \circ \hat{e}$ that achieves low **excess risk**:

$$\mathbb{E}_{P_{X,Y}}\left[\ell^{fg}(\hat{f} \circ \hat{e}(X), Y) - \ell^{fg}\left(f^* \circ e^*(X), Y\right)\right]$$

Where $e^* \in \arg\min_{e \in \mathcal{E}} \mathbb{E}_{P_{S,W}} \ell_e^{cg}(S, W)$ and

$$f^* \in \arg\min_{f \in \mathcal{F}} \mathbb{E}_{P_{X,Y}}\left[\ell^{fg}\left(f \circ e^*(X), Y\right)\right].$$



Schema of the model. The dotted lines represent the flow of fine-grained data, and the solid lines denote the flow of coarse-grained labels

5 Theoretical Analysis

We denote the underlying distribution of $\mathcal{D}_m^{cg}$ as $P_{S,W}$ and the underlying distribution of $\mathcal{D}_n^{fg}$ as $P_{X,Y}$. We assume the joint distribution of Z and Y is $P_{Z,Y}$.

Definition 1. (Coarse-grained learning; pretraining) Let $\text{Rate}_m(\ell^{cg}, P_{S,W}, \mathcal{E})$ be the excess risk rate of $\mathcal{A}_m(\ell^{cg}, P_{S,W}, \mathcal{E})$.

Definition 2. (Fine-grained learning; downstream task learning) Let $\text{Rate}_n(\ell^{fg}, P_{Z,Y}, \mathcal{F})$ be the excess risk rate of $\mathcal{A}_n(\ell^{fg}, P_{Z,Y}, \mathcal{F})$.

Definition 3. We say that f is L-Lipschitz relative to $\mathcal{E}$ if for all $s \in \mathcal{S}, x \in s, y \in \mathcal{Y}$, and $e, e' \in \mathcal{E}$,

$$|\ell^{fg}(f \circ e(x), y) - \ell^{fg}(f \circ e'(x), y)| \le L\ell^{cg}(g_e \circ \phi^e(s), g_{e'} \circ \phi^{e'}(s))$$

The function class $\mathcal{F}$ is L-Lipschitz relative to $\mathcal{E}$, if every $f \in \mathcal{F}$ is L-Lipschitz relative to $\mathcal{E}$.

Theorem 4. Assume that $\text{Rate}_m\left(\ell^{cg}, P_{S,W}, \mathcal{E}\right) = \mathcal{O}\left(1/m^\alpha\right)$ and growth rate $m = \Omega\left(n^\beta\right)$, we obtain **excess risk bound** with probability at least $1 - \delta$ by

$$\mathcal{O}\left(\frac{d\alpha\beta \log RL'n + \log\frac{1}{\delta}}{n} + \frac{B + 2L}{n^{\alpha\beta}}\right)$$

$\mathcal{F}$ is L'-Lipschitz in its d-dimensional parameters in the l_2 norm

$\mathcal{F}$ is contained in the Euclidean ball of radius R

ℓ^{fg} is bounded by $B > 0$

$\mathcal{F}$ is L-Lipschitz relative to $\mathcal{E}$

Conditions of fast excess risk rate (i.e., $\alpha\beta \ge 1$):
(1) Larger α: better generalization performance on the pretraining task.
(2) Larger β: larger growth rate of coarse-grained labels.

6 Experiments: CIFAR-100

Pretrain with unique superclass number

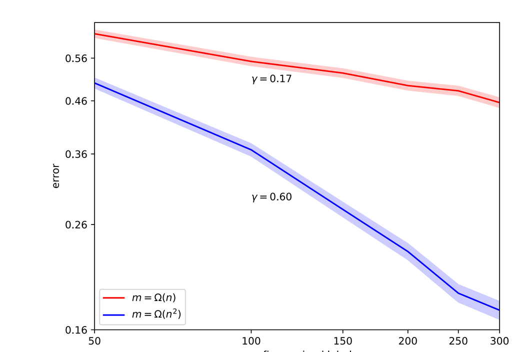
- Input-sets: we generate input sets by sampling between 6 and 10 images from CIFAR-100 training data
- Targets: the unique superclass number of the input sets
- Downstream task: few-shot testing on 100 classes of test set
- Fine-grained learning: nearest centroid (NC); logistic regression (LR); ridge classifier (RC)

pretrain method	NC	LR	RC
SimCLR	76.07 ± 0.97	75.88 ± 1.01	75.50 ± 1.02
SimSiam	78.15 ± 0.93	79.44 ± 0.92	79.03 ± 0.95
FSP-Patch	N/A	N/A	N/A
FACILE-SupCon	N/A	N/A	N/A
FACILE-FSP	$\mathbf{86.25 \pm 0.79}$	$\mathbf{85.42 \pm 0.82}$	$\mathbf{85.84 \pm 0.81}$

Pretrain with Most Frequent Superclass

- Targets: the most frequent superclass of the input sets

pretrain method	NC	LR	RC
SimCLR	75.91 ± 1.00	75.82 ± 1.01	75.91 ± 1.02
SimSiam	78.80 ± 0.93	79.44 ± 0.95	79.43 ± 0.93
FSP-Patch	73.21 ± 0.97	73.92 ± 0.98	73.40 ± 0.98
FACILE-SupCon	79.54 ± 0.92	79.54 ± 0.96	79.12 ± 0.95
FACILE-FSP	$\mathbf{82.04 \pm 0.84}$	$\mathbf{81.70 \pm 0.91}$	$\mathbf{81.75 \pm 0.90}$



We can improve excess risk by increasing growth rate of coarse-grained labels and maintain log-linear relationship.

Generalization error (with two growth rates) of FACILE-FSP on CIFAR-100 test dataset as a function of the number of coarse-grained labels m.

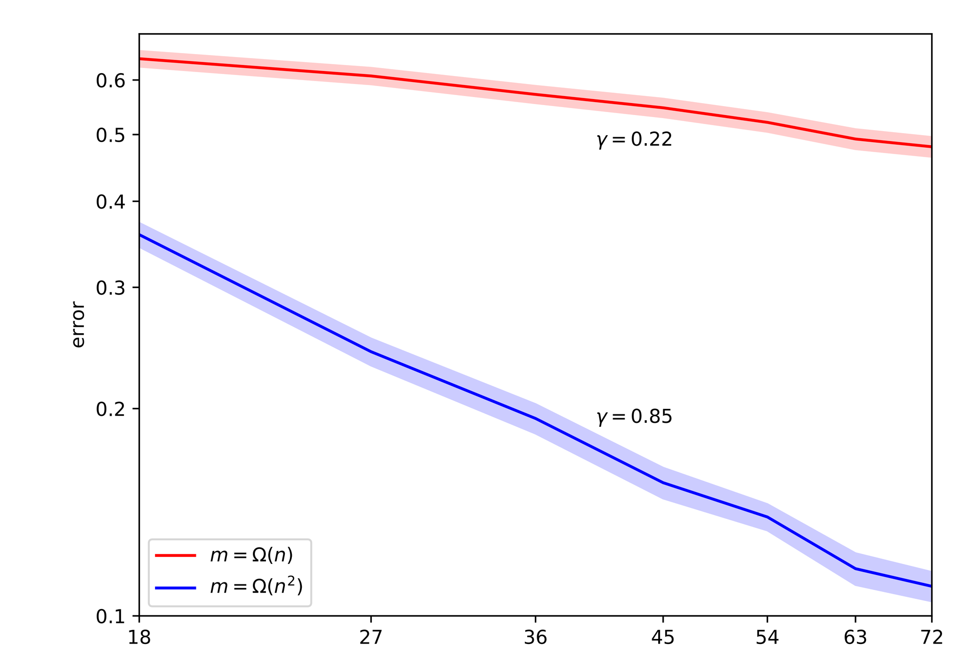
7 Experiments: WSIs

Evaluation on LC, PAIP, and NCT

We first fine-tune fully-connected layer appended to ViT-B/14 from DINO V2 on TCGA patches with size 224×224 at 20× magnification. After the models are trained, we test the feature map in these models on LC, PAIP, and NCT.

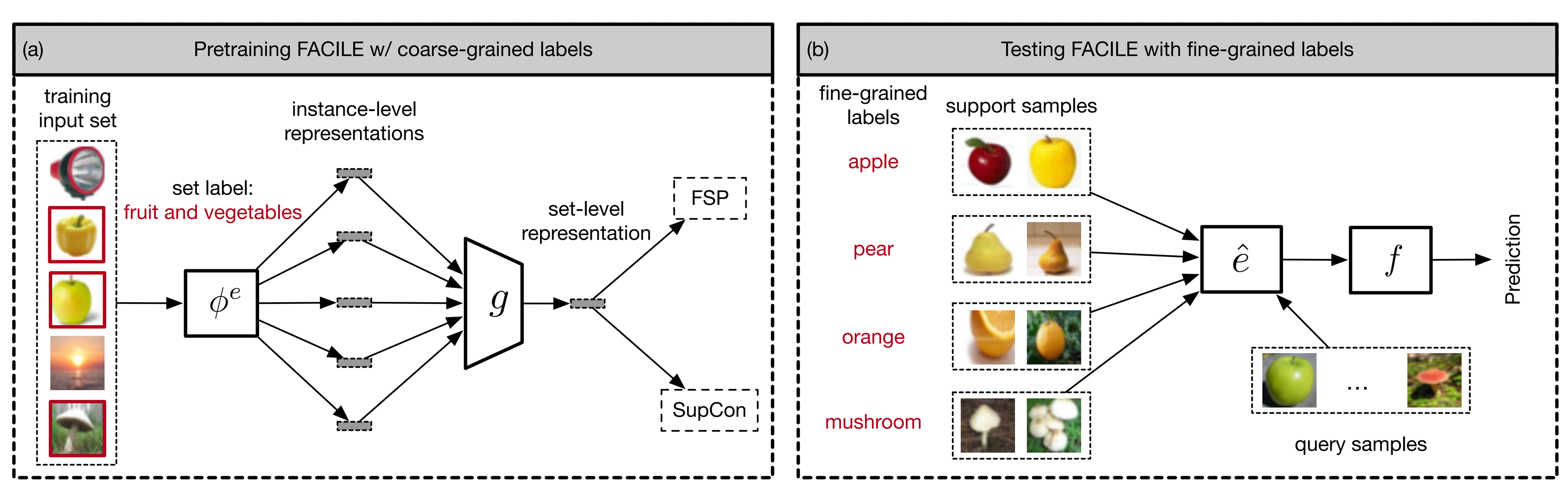
pretraining method	NC	LR	RC	LR+LA	RC+LA
1-shot 5-way test on LC dataset					
DINO V2 (ViT-B/14)	44.82 ± 1.41	47.51 ± 1.39	47.63 ± 1.38	47.36 ± 1.39	48.88 ± 1.44
SimSiam	48.79 ± 1.37	49.43 ± 1.35	48.43 ± 1.36	49.38 ± 1.34	49.50 ± 1.34
SimCLR	50.47 ± 1.31	50.52 ± 1.33	50.44 ± 1.32	51.66 ± 1.32	51.78 ± 1.38
FSP-Patch	49.73 ± 1.41	53.59 ± 1.38	53.07 ± 1.41	51.79 ± 1.40	51.27 ± 1.43
FACILE-SupCon	$\mathbf{56.24 \pm 1.43}$	$\mathbf{56.51 \pm 1.41}$	$\mathbf{55.95 \pm 1.42}$	$\mathbf{56.29 \pm 1.43}$	54.07 ± 1.44
FACILE-FSP	55.67 ± 1.40	56.26 ± 1.36	55.83 ± 1.35	56.01 ± 1.38	$\mathbf{55.35 \pm 1.40}$
5-shot 5-way test on LC dataset					
DINO V2 (ViT-B/14)	66.12 ± 0.98	64.71 ± 1.12	66.36 ± 1.10	72.95 ± 0.93	75.11 ± 0.91
SimSiam	67.51 ± 0.96	64.99 ± 1.05	65.39 ± 1.05	70.30 ± 0.93	71.19 ± 0.93
SimCLR	70.10 ± 0.92	69.28 ± 0.96	69.18 ± 0.97	72.99 ± 0.92	72.91 ± 0.94
FSP-Patch	71.97 ± 0.96	71.11 ± 1.04	71.19 ± 1.03	73.96 ± 0.94	73.20 ± 0.96
FACILE-SupCon	75.58 ± 0.88	74.26 ± 0.94	73.20 ± 0.95	75.81 ± 0.90	74.34 ± 0.96
FACILE-FSP	$\mathbf{75.86 \pm 0.86}$	$\mathbf{74.64 \pm 0.89}$	$\mathbf{74.12 \pm 0.93}$	$\mathbf{76.17 \pm 0.88}$	$\mathbf{75.08 \pm 0.95}$
1-shot 3-way test on PAIP dataset					
DINO V2 (ViT-B/14)	41.51 ± 1.27	44.37 ± 1.26	44.28 ± 1.25	42.43 ± 1.27	42.78 ± 1.27
SimSiam	49.42 ± 1.28	48.07 ± 1.35	48.44 ± 1.36	48.76 ± 1.33	46.48 ± 1.37
SimCLR	48.60 ± 1.19	48.76 ± 1.25	47.98 ± 1.26	48.94 ± 1.23	47.20 ± 1.26
FSP-Patch	46.09 ± 1.17	47.44 ± 1.18	48.09 ± 1.19	46.76 ± 1.18	43.68 ± 1.22
FACILE-SupCon	$\mathbf{51.97 \pm 1.18}$	$\mathbf{52.25 \pm 1.22}$	$\mathbf{51.80 \pm 1.22}$	51.36 ± 1.22	$\mathbf{50.24 \pm 1.23}$
FACILE-FSP	51.34 ± 1.16	51.18 ± 1.19	51.51 ± 1.19	$\mathbf{51.50 \pm 1.16}$	49.77 ± 1.22
5-shot 3-way test on PAIP dataset					
DINO V2 (ViT-B/14)	57.59 ± 1.07	58.19 ± 1.10	59.37 ± 1.07	61.84 ± 0.85	60.81 ± 0.86
SimSiam	61.56 ± 0.97	62.52 ± 1.01	62.81 ± 1.01	64.40 ± 0.86	62.44 ± 0.93
SimCLR	62.20 ± 0.93	61.78 ± 0.99	63.20 ± 0.97	63.38 ± 0.86	63.03 ± 0.88
FSP-Patch	63.77 ± 0.88	63.85 ± 0.94	63.85 ± 0.93	63.61 ± 0.85	60.91 ± 0.87
FACILE-SupCon	$\mathbf{67.16 \pm 0.84}$	67.29 ± 0.89	66.88 ± 0.90	$\mathbf{67.61 \pm 0.85}$	$\mathbf{66.34 \pm 0.84}$
FACILE-FSP	67.14 ± 0.85	$\mathbf{67.67 \pm 0.84}$	$\mathbf{67.54 \pm 0.86}$	67.12 ± 0.81	66.05 ± 0.83
1-shot 9-way test on NCT dataset					
DINO V2 (ViT-B/14)	56.03 ± 1.62	59.11 ± 1.57	60.13 ± 1.55	58.71 ± 1.57	59.06 ± 1.55
SimSiam	62.60 ± 1.45	61.89 ± 1.50	61.90 ± 1.51	62.27 ± 1.47	61.05 ± 1.44
SimCLR	65.43 ± 1.43	64.18 ± 1.44	64.15 ± 1.46	64.83 ± 1.43	62.69 ± 1.38
FSP-Patch	65.22 ± 1.49	65.93 ± 1.41	65.94 ± 1.40	65.26 ± 1.45	62.66 ± 1.46
FACILE-SupCon	71.55 ± 1.36	70.36 ± 1.37	70.52 ± 1.35	71.05 ± 1.35	$\mathbf{68.85 \pm 1.40}$
FACILE-FSP	$\mathbf{72.05 \pm 1.34}$	$\mathbf{70.70 \pm 1.35}$	$\mathbf{70.77 \pm 1.34}$	$\mathbf{71.14 \pm 1.34}$	68.03 ± 1.40
5-shot 9-way test on NCT dataset					
DINO V2 (ViT-B/14)	76.85 ± 0.98	76.51 ± 1.02	78.67 ± 0.94	82.20 ± 0.82	82.75 ± 0.83
SimSiam	80.81 ± 0.85	80.06 ± 0.87	81.55 ± 0.85	83.18 ± 0.80	82.39 ± 0.83
SimCLR	82.87 ± 0.80	81.91 ± 0.82	82.86 ± 0.80	83.92 ± 0.77	82.89 ± 0.79
FSP-Patch	83.63 ± 0.83	83.49 ± 0.80	84.34 ± 0.78	85.32 ± 0.75	83.03 ± 0.79
FACILE-SupCon	87.74 ± 0.64	87.00 ± 0.64	87.38 ± 0.62	87.82 ± 0.63	86.15 ± 0.69
FACILE-FSP	$\mathbf{87.93 \pm 0.65}$	$\mathbf{87.52 \pm 0.65}$	$\mathbf{87.72 \pm 0.62}$	$\mathbf{88.01 \pm 0.64}$	$\mathbf{86.46 \pm 0.70}$

Latent augmentation (LA) was originally proposed in Yang et al. (2022) to improve the performance of the few-shot learning system in a simple unsupervised way.



Generalization error on NCT dataset. The FACILE-FSP (ResNet18) trains on TCGA dataset with m coarse-grained labels. We show the error curve with two growth rates of m.

4 FACILE Algorithm



FSP: fully supervised preparing; SupCon: Supervised Contrastive Learning

Algorithm: FACILE algorithm

1. **input:** loss function ℓ^{fg}, ℓ^{cg}, feature map $\mathcal{E}$, predictors $\mathcal{G}, \mathcal{F}$, datasets $\mathcal{D}_m^{cg}, \mathcal{D}_m^{fg}$
2. obtain feature map $\hat{e} \leftarrow \mathcal{A}(\ell^{cg}, \mathcal{D}_m^{cg}, \mathcal{E})$
3. create artificial dataset $\mathcal{D}_n^{fg,aug} = \left\{\left(z_i, y_i\right) : z_i = \hat{e}\left(x_i\right), \left(x_i, y_i\right) \in \mathcal{D}_n^{fg}\right\}_{i=1}^n$
4. obtain fine-grained label predictor $\hat{f} \circ \hat{e}$, where $\hat{f} \leftarrow \mathcal{A}\left(\ell^{fg}, \mathcal{D}_n^{fg,aug}, \mathcal{F}\right)$
5. **output:** $\hat{f} \circ \hat{e}$