Vision-Language Models are Strong Noisy Label Detectors

**Tong Wei, Hao-Tian Li, Chun-Shu Li, Jiang-Xin Shi,
Yu-Feng Li, Min-Ling Zhang**

Southeast University, Nanjing, China
Nanjing University, Nanjing, China

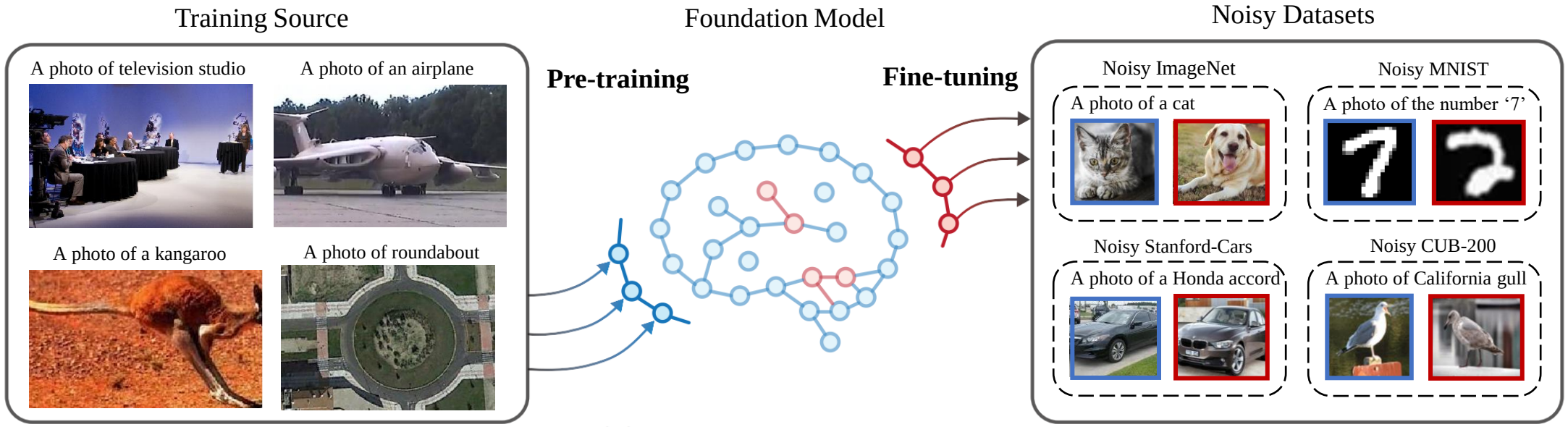
Outline

- Background
- The Proposed Approach
- Experiments
- Conclusion

Background

Fine-Tuning Vision-Language Models

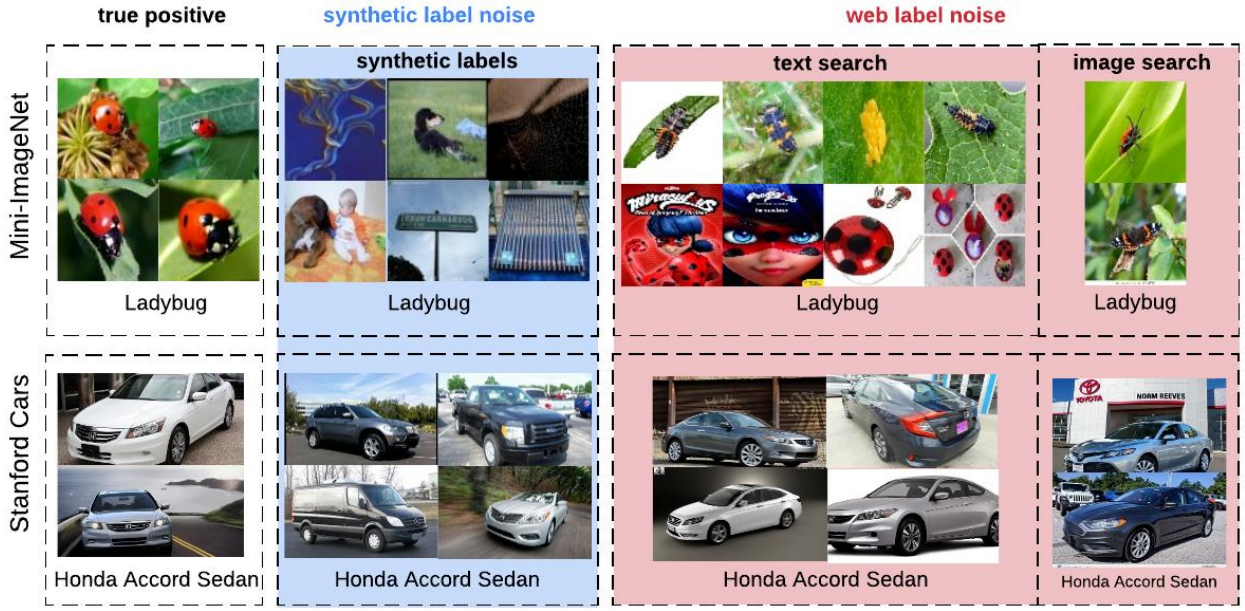
- Vision-Language models such as CLIP have gained widespread adoption in various classification tasks.
- Despite its good zero-shot performance, fine-tuning becomes necessary when the data distribution of downstream tasks significantly deviates from the CLIP training source.



Background

Learning from Noisy labels: the given label varies from true class

- Fine-tuning CLIP necessitates perfectly labeled datasets which may not be readily available in many real-world tasks.
- To mitigate the negative impact of noisy labels, researchers have proposed various approaches for learning with noisy labels.
- However, the exploration of this problem in the context of CLIP adaptation remains limited.



The Model Fine-tuning Paradigms

- **Full fine-tuning (FFT):** modifies all model parameters.
- **Parameter-efficient fine-tuning (PEFT):** modifies a few extra parameters, such as LoRA and VPT.

What is the most effective paradigm for vision-language model adaptation with noisy data?

心於至善

https://research.google/blog/understanding-deep-learning-on-controlled-noisy-labels/

Outline

- Background
- The Proposed Approach
- Experiments
- Conclusion

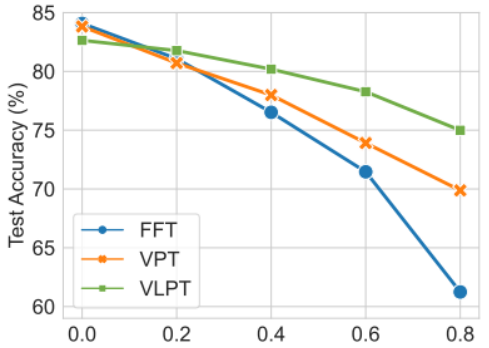
Initial Findings

We utilize three fine-tuning approaches to adapt CLIP on both noisy and clean datasets

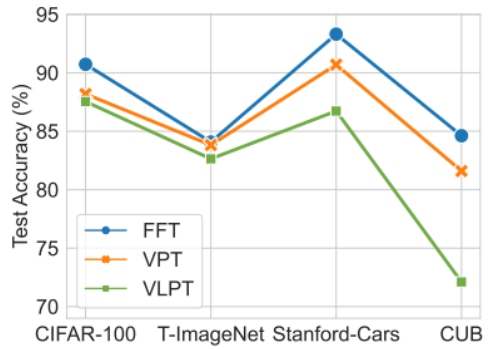
- **FFT:** full fine-tuning for visual encoder and an additional linear head for classification
- **VPT:** visual prompt tuning for visual encoder and an additional linear head for classification
- **VLPT:** prompt tuning for both visual and textual encoder, with the learned textual prompts for classification

Initial Findings

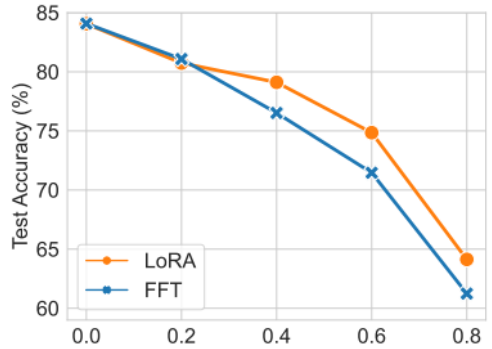
a) PEFT benefits visual representation learning under massive noisy labels, i.e., *figure (a), (c) and (d).*
b) Textual classifier is more robust to noisy labels than linear classifier, i.e., *figure (a).*
c) FFT enhances visual recognition on clean datasets, i.e., *figure (b).*



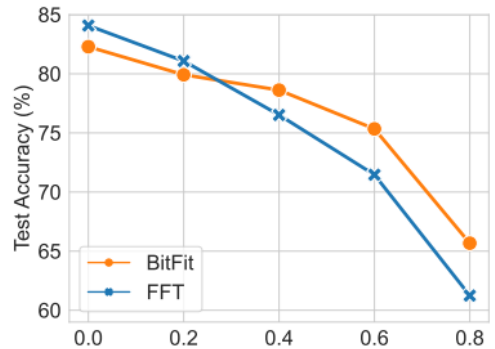
(a) Noisy Tiny-ImageNet



(b) Clean Datasets



(c) LoRA

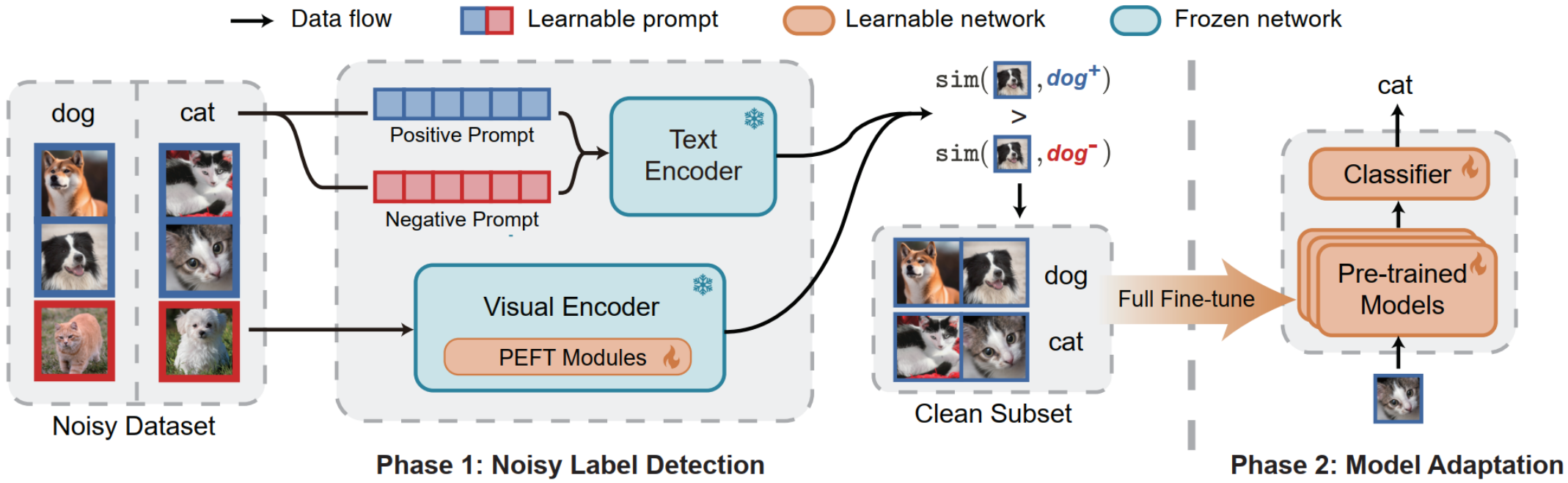


(d) BitFit

心於至善

The Proposed Approach

The Denoising Fine-tuning Framework (DeFT)



The Proposed Approach

The Denoising Fine-tuning Framework

Phase1: Noisy Label Detection

- Previous methods only use the image modality for sample selection and relies heavily on either the estimated noise ratio or the threshold, can we utilize the multimodal information in CLIP to enhance noise detection?
- The robustness of parameter-efficient fine-tuning and textual classifier to label noise has been empirically demonstrated, can we harness this property to better identify noisy samples?

Identifying Noisy Labels with Dual Prompts

(1) Design a class-specific pair of *positive* and *negative* prompts for the textual encoder as $prompt_k^+$ and $prompt_k^-$:

$$prompt_k^+ = [V]_1^+[V]_2^+[V]_3^+ \dots [V]_M^+[CLS]_k$$

$$prompt_k^- = [V]_1^-[V]_2^-[V]_3^- \dots [V]_M^-[CLS]_k$$

(2) The negative prompt serves as a learnable sample-dependent threshold to induce clean subsets D_{clean}:

$$D_{clean} = \{(\boldsymbol{x}_i, y_i) | \text{sim}(\boldsymbol{I}_i, \boldsymbol{T}_k^+) > \varphi_i, y_i = k\}$$

$$\varphi_i = \text{sim}(\boldsymbol{I}_i, \boldsymbol{T}_k^-)$$

Optimization for Noisy Label Detector

(3) Formulating the clean probability of the i-th image:

$$p_{ik}^{clean} = \frac{\exp(\text{sim}(\boldsymbol{I}_i, \boldsymbol{T}_k^+)/\tau)}{\exp(\text{sim}(\boldsymbol{I}_i, \boldsymbol{T}_k^+)/\tau) + \exp(\text{sim}(\boldsymbol{I}_i, \boldsymbol{T}_k^-)/\tau)}$$

(4) Optimize the parameters of dual prompts in textual encoder and harness PEFT for the adaptation of visual encoder:

$$L_{dp} = \frac{1}{N}\sum_{i=1}^{N} l_{nll}(\boldsymbol{p}_i^{clean}, \hat{y}) + l_{nll}(1 - \boldsymbol{p}_i^{clean}, \bar{y})$$

$$L_{sim} = -\frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{\exp(\text{sim}(\boldsymbol{I}_i, \boldsymbol{T}_i^+)/\tau)}{\sum_{k=1}^{K}\exp(\text{sim}(\boldsymbol{I}_i, \boldsymbol{T}_k^+)/\tau)}\right)$$

$$L = L_{dp} + L_{sim}$$

The Proposed Approach

The Denoising Fine-tuning Framework

Phase2: Model Adaptation

- Although the learned positive textual prompt can be readily employed for classification, its performance may be suboptimal on curated clean datasets, as demonstrated in our previous finding.
- With the selected clean samples after phase1,the second phase can be applied universally to a wide range of pre-trained models, regardless of their backbones.

Model Adaptation using Clean Data

- Learn an additional linear head for classification in the model adaptation phase.
- Remove the PEFT modules in visual encoder and fully fine-tune the pre-trained model.

$$l_{ce} = -\log\left(\frac{\exp(z_y)}{\sum_{k=1}^{K}\exp(z_k)}\right)$$

Algorithm 1: The Proposed DEFT Framework

Input: training dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)_{i=1}^N\}$, PEFT parameters ω, pre-trained parameters θ, warm-up epoch T_0, PEFT epoch T_1 and FFT epoch T_2.
// Phase1: Learning Noisy Label Detector with PEFT
for $t = 1, 2, ..., T_0$ **do**
  Warm-up the pre-trained model on noisy dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)_{i=1}^N\}$
end
for $t = T_0 + 1, ..., T_1$ **do**
  Construct the clean subset $\mathcal{D}^{\text{clean}}$ by Eq. (4) and Eq. (5)
  Compute the total loss $\mathcal{L} = \mathcal{L}_{dp} + \mathcal{L}_{sim}$ by Eq. (7) and Eq. (8)
  Update current model parameters $\omega_t = \text{SGD}(\mathcal{D}^{\text{clean}}, \mathcal{L}, \omega_{t-1})$
end
// Phase2: Adapting Model on Clean Data with FFT
for $t = 1, 2, ..., T_2$ **do**
  Compute the CE loss ℓ_{ce} for samples in the clean subset $\mathcal{D}^{\text{clean}}$
  Update current model parameters $\theta_t = \text{SGD}(\mathcal{D}^{\text{clean}}, \ell_{ce}, \theta_{t-1})$
end

Outline

- Background
- The Proposed Approach
- Experiments
- Conclusion

Experiments

Performance for Noisy Label Detection

Method	*Sym.* 0.2		*Sym.* 0.4		*Sym.* 0.6		*Ins.* 0.2		*Ins.* 0.3		*Ins.* 0.4	
	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
CIFAR-100												
Label-match	99.83	63.62	99.61	63.85	99.31	63.52	99.93	63.65	99.85	63.72	99.81	63.69
Small-loss	97.24	96.79	95.68	94.49	92.93	90.68	95.20	95.46	94.00	92.53	90.33	89.85
DEFT (ours)	**99.51**	**97.77**	**98.75**	**97.91**	**97.04**	**97.27**	**98.47**	**97.88**	**96.32**	**97.63**	**94.08**	**95.28**
Δ	↑ 2.27	↑ 0.98	↑ 3.07	↑ 3.42	↑ 4.11	↑ 6.59	↑ 3.27	↑ 2.42	↑ 2.32	↑ 5.10	↑ 3.75	↑ 5.43
Tiny-ImageNet												
Label-match	99.92	60.81	99.83	60.79	99.50	60.66	99.91	60.58	99.84	60.53	99.76	60.47
Small-loss	97.25	**96.93**	95.33	94.48	92.63	90.89	94.74	95.17	93.66	92.35	90.41	89.71
DEFT (ours)	**99.50**	96.00	**98.78**	**95.97**	**97.21**	**95.44**	**99.21**	**96.21**	**97.80**	**95.80**	**95.45**	**95.77**
Δ	↑ 2.25	↓ 0.93	↑ 3.45	↑ 1.49	↑ 4.58	↑ 4.55	↑ 4.47	↑ 1.04	↑ 4.14	↑ 3.45	↑ 5.04	↑ 6.06
Stanford-Cars												
Label-match	99.97	60.34	99.86	60.27	99.70	60.71	99.85	60.34	99.82	60.32	99.80	60.25
Small-loss	96.92	96.56	93.71	93.21	89.46	87.79	96.94	97.78	96.72	95.96	95.25	94.48
DEFT (ours)	**98.72**	**99.56**	**98.98**	**98.56**	**98.58**	**95.62**	**99.02**	**99.09**	**98.96**	**98.15**	**98.75**	**97.71**
Δ	↑ 1.80	↑ 3.00	↑ 5.27	↑ 5.35	↑ 9.12	↑ 7.83	↑ 2.08	↑ 1.31	↑ 2.24	↑ 2.19	↑ 3.50	↑ 3.23
CUB-200-2011												
Label-match	99.92	53.26	99.74	53.13	99.46	53.02	99.96	53.39	99.96	53.32	99.74	53.69
Small-loss	96.74	96.32	93.69	92.84	84.10	82.01	96.91	97.33	96.49	95.59	93.98	93.96
DEFT (ours)	**99.04**	**97.01**	**96.76**	**95.60**	**93.88**	**96.43**	**99.15**	**97.45**	**97.93**	**96.85**	**96.03**	**97.11**
Δ	↑ 2.30	↑ 0.69	↑ 3.07	↑ 2.76	↑ 9.78	↑ 14.42	↑ 2.24	↑ 0.12	↑ 1.44	↑ 1.26	↑ 2.05	↑ 3.15

Table 1: On each dataset, we compare the Precision (%) and Recall (%) of DEFT with CLIP label-match and small-loss to evaluate the clean sample selection performance. Δ is the difference between the performance of DEFT and small-loss.

Performance for Image Classification

	Method	*Sym.* 0.2	*Sym.* 0.4	*Sym.* 0.6	*Ins.* 0.2	*Ins.* 0.3	*Ins.* 0.4
CIFAR-100							
FFT	CE	86.71 / 86.70	84.06 / 82.60	81.05 / 77.45	87.30 / 87.18	84.60 / 83.64	78.41 / 75.66
	ELR	86.53 / 86.53	83.66 / 83.66	78.34 / 78.34	86.61 / 86.61	85.89 / 85.89	**85.78 / 85.78**
	SCE	86.82 / 86.82	83.84 / 83.84	78.90 / 77.71	86.61 / 86.61	83.99 / 83.20	80.06 / 73.45
	GMM	88.49 / 88.49	87.21 / 87.21	85.22 / 85.20	88.44 / 88.44	87.95 / 87.95	82.14 / 82.11
DEFT	Ours	**89.38 / 89.35**	**88.17 / 88.11**	**85.81 / 85.72**	**89.38 / 89.35**	**88.68 / 88.68**	85.75 / 85.74
Tiny-ImageNet							
FFT	CE	81.77 / 81.08	76.53 / 76.52	73.17 / 71.46	80.75 / 80.71	78.83 / 78.57	74.80 / 74.08
	ELR	79.40 / 79.40	77.13 / 77.13	73.74 / 73.74	79.98 / 79.98	77.13 / 77.13	73.74 / 73.74
	SCE	79.23 / 79.23	76.24 / 76.18	71.76 / 70.62	78.96 / 78.90	77.80 / 77.54	74.47 / 73.25
	GMM	81.91 / 81.88	80.37 / 80.37	43.47 / 43.47	81.84 / 81.79	81.26 / 81.26	79.01 / 79.01
DEFT	Ours	**82.91 / 82.91**	**82.48 / 82.37**	**80.60 / 80.59**	**83.37 / 83.33**	**82.69 / 82.65**	**80.52 / 80.49**
Stanford-Cars							
FFT	CE	89.75 / 89.74	85.10 / 84.89	71.70 / 71.55	89.13 / 89.06	85.94 / 85.92	80.59 / 80.59
	ELR	86.61 / 86.61	76.98 / 76.98	61.58 / 61.58	84.40 / 84.40	83.11 / 83.11	75.97 / 75.84
	SCE	91.11 / 91.11	87.73 / 87.45	79.09 / 79.09	90.34 / 90.34	87.35 / 86.23	83.50 / 80.69
	GMM	90.10 / 90.08	83.14 / 83.10	56.90 / 56.90	88.15 / 88.10	85.39 / 85.33	78.76 / 78.72
DEFT	Ours	**92.13 / 92.12**	**90.75 / 90.75**	**85.72 / 85.45**	**92.19 / 92.15**	**90.77 / 90.77**	**89.74 / 89.68**
CUB-200-2011							
FFT	CE	80.76 / 80.76	73.09 / 72.87	55.42 / 55.21	80.36 / 80.25	75.80 / 75.53	69.62 / 69.62
	ELR	77.70 / 77.70	68.26 / 68.26	50.17 / 49.88	78.32 / 78.32	73.16 / 73.08	63.57 / 63.34
	SCE	82.81 / 82.74	78.12 / 77.87	63.31 / 63.31	81.91 / 81.91	78.31 / 78.03	71.25 / 70.95
	GMM	75.79 / 75.73	64.39 / 64.38	42.84 / 42.84	75.73 / 75.65	69.95 / 69.95	56.13 / 55.80
DEFT	Ours	**83.05 / 83.03**	**79.24 / 79.13**	**73.08 / 73.08**	**82.53 / 82.50**	**81.39 / 81.39**	**79.34 / 79.24**

Table 2: Test accuracy (%) on synthetic datasets with *symmetric* and *instance-dependent* label noise.

Dataset	CE	ELR	SCE	GMM	RoLT	UNICON	LongReMix	ProMix	DEFT (Ours)
CIFAR-100N	72.41	72.83	72.52	76.06	75.91	77.68	73.94	75.97	**79.04**
Clothing1M	69.75	72.14	70.49	70.03	70.46	70.38	70.62	70.71	**72.44**
WebVision	84.64	79.32	82.88	84.88	84.12	84.56	84.96	84.44	**85.12**

Table 3: Test accuracy (%) on datasets with real-world label noise.

Experiments

Further Analyses

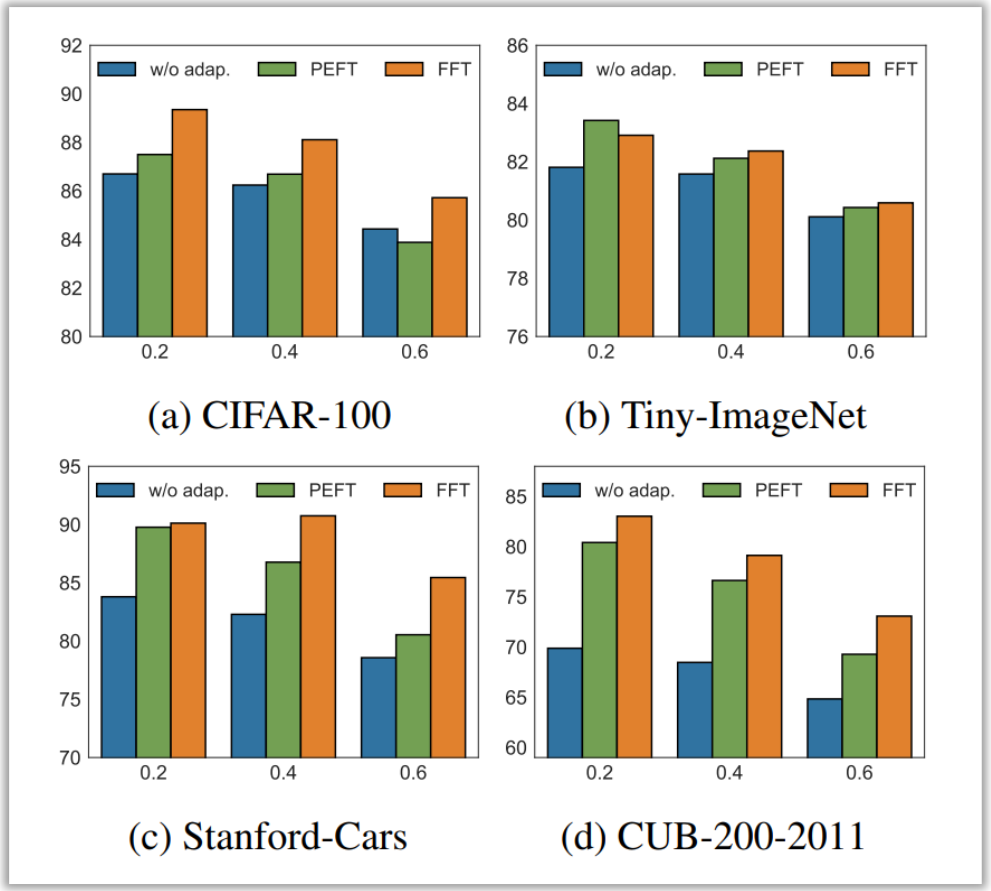
Necessity of Model Adaptation

- We conduct ablation studies and report the test accuracy across varying noise ratios for variants.
- Employing FFT for model adaptation is more effective in mitigating significant domain shifts.

DeFT for Various Pre-trained Models

- DEFT can seamlessly integrate with various pre-trained visual backbones during the model adaptation phase.

Architecture	CE	GCE	ELR	TURN	DEFT (Ours)
ResNet-50 [11]	66.02	66.19	66.19	66.31	**70.82**
MAE-ViT-B [10]	61.31	60.80	61.51	61.96	**65.23**
ViT-B/16 [5]	68.98	69.74	68.73	**70.28**	69.84
ConvNeXt-T [27]	68.80	68.92	68.52	69.53	**71.68**



Outline

- Background
- The Proposed Approach
- Experiments
- Conclusion

Conclusion

- **Methods:** we delve into a new landscape for learning with noisy labels, departing from the classic single-modal toward a multi-modal regime.
- **Versatility:** DeFT is robust to various types of label noise, generalizable to many pre-trained models, and does not require the dynamics of training samples.
- **Effectiveness:** we investigate the effectiveness of DeFT on a wide range of synthetic and real-world datasets, showing its superior performance in both noisy label detection and image classification tasks.
- **Future work:** we hope our work will inspire future research toward multi-modal noisy label detection.

Thanks !

https://github.com/HotanLee/DeFT

liht@seu.edu.cn