Semi-Supervised Noise Adaptation

Transferring Knowledge from Noise Domain

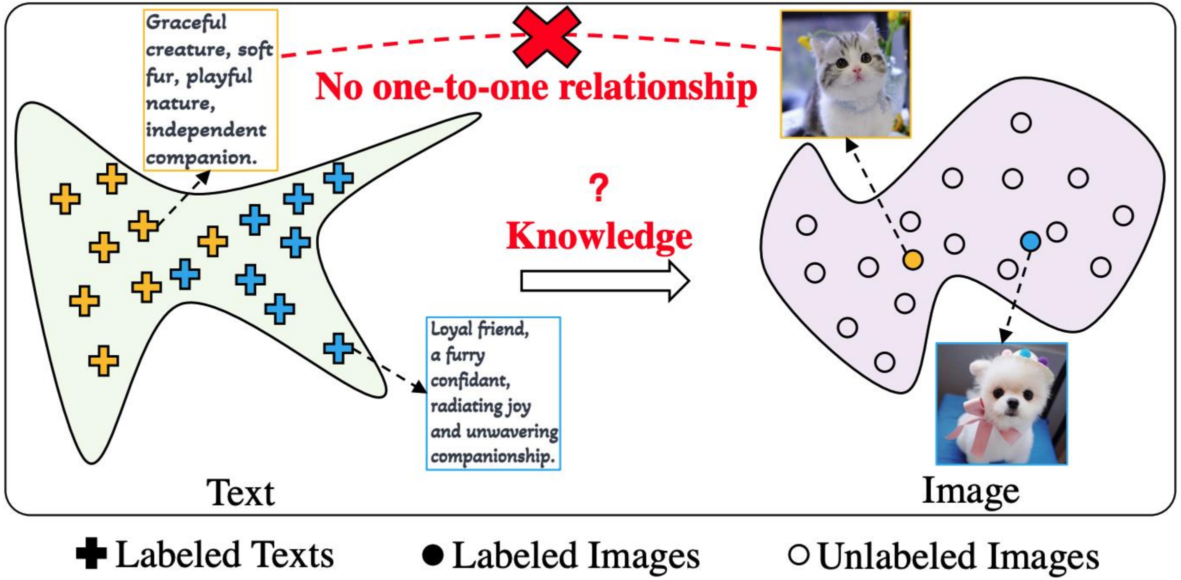
Yuan Yao[1,2] **,** Jin Song[3] , Huixia Li[4] , Tongtong Yuan[5] , Jiaqi Wu[6] , Yu Zhang[7]

[1]Teleinfo, CAICT, [2]Guangming Lab, [3]NJUPT, [4]BJTU , [5]BJUT, [6]THU, [7]SUSTech

2026.06

Background: Cross-modality Transfer Learning

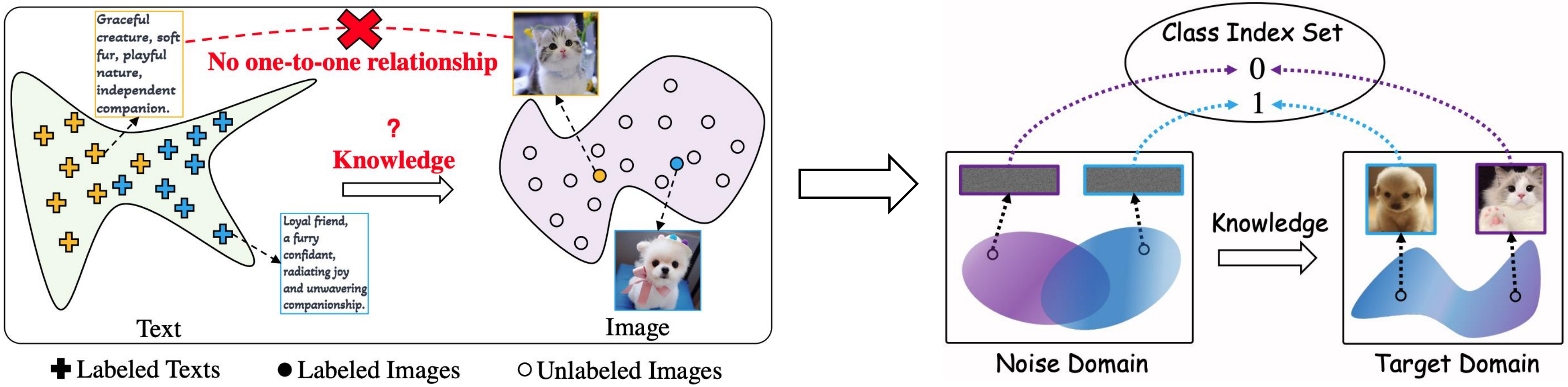
Cross-modality (Heterogeneous) transfer learning leverages a label-rich source domain from a **distinct modality** (e.g., text) to facilitate learning in a label-scare target domain (e.g., image).



However, cross-domain paired samples are unavailable, yet transfer remains effective.

Background: A Surprising Observation

Noise can serve as a surrogate source domain and enable positive transfer in a semi-supervised setting **without real source samples** [1].



This observation is important: privacy, confidentiality, and copyright restrict access to usable source samples.

[1] Yao, Y., Zhang, X., Zhang, Y., Jin, J., & Yang, Q. (2025). Noise May Contain Transferable Knowledge: Understanding Semi-supervised Heterogeneous Domain Adaptation from an Empirical Perspective. *arXiv preprint arXiv:2502.13573.*

Background: Limitations of Previous Work

Lack of Theory

[1] lacks a generalization bound analysis explaining why the noise domain improves generalization.

Limited Validation

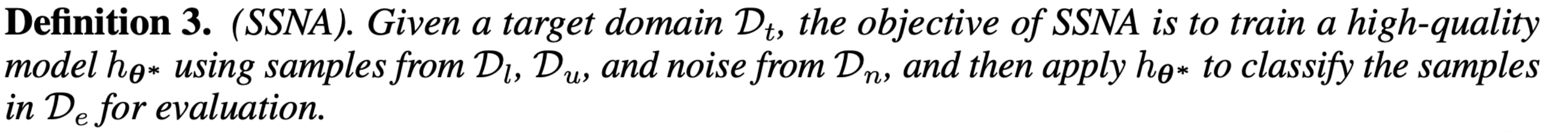
[1] omits standard benchmarks such as CIFAR-10/100 or ImageNet, which may limit the applicability of its findings.

[1] Yao, Y., Zhang, X., Zhang, Y., Jin, J., & Yang, Q. (2025). Noise May Contain Transferable Knowledge: Understanding Semi-supervised Heterogeneous Domain Adaptation from an Empirical Perspective. *arXiv preprint arXiv:2502.13573.*

Problem Formulation: Semi-Supervised Noise Adaptation

Target domain: $\mathcal{D}_t = \mathcal{D}_l$ (few labeled) $\cup$ $\mathcal{D}_u$ (massive unlabeled) $\cup$ $\mathcal{D}_e$ (test)

Noise domain: $\mathcal{D}_n$ (sampled from $\mathbb{R}^p$)



Definition 3. *(SSNA). Given a target domain $\mathcal{D}_t$, the objective of SSNA is to train a high-quality model $h_{\boldsymbol{\theta}*}$ using samples from $\mathcal{D}_l$, $\mathcal{D}_u$, and noise from $\mathcal{D}_n$, and then apply $h_{\boldsymbol{\theta}*}$ to classify the samples in $\mathcal{D}_e$ for evaluation.*

Methodology: A Generalization Bound for SSNA

Theorem 4.1 (Generalization Bound of SSNA). *Let* $\hat{f} = \arg\min_{f\in\mathcal{F}} \hat{\epsilon}_\alpha(f)$ *be the empirical minimizer of* $\hat{\epsilon}_\alpha(f)$, *and let* $f_t^* = \arg\min_{f\in\mathcal{F}} \epsilon_t(f)$ *be the target error minimizer. Then, for any* $\delta \in (0,1)$, *with probability at least* $1-\delta$ *(over the choice of the samples), we have:*

$$\epsilon_t(\hat{f}) \le \epsilon_t(f_t^*) + \mathcal{O}\left(\gamma\sqrt{\frac{d\log m + \log(\frac{1}{\delta})}{m}}\right) + 2(1-\alpha)\left[\frac{1}{2}\overset{\text{Empirical Distributional Discrepancy}}{\boxed{\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathbb{U}_n,\mathbb{U}_t)}} + \mathcal{O}\left(\sqrt{\frac{d\log m' + \log(\frac{1}{\delta})}{m'}}\right)\right.$$
$$\left. + \overset{\text{Empirical Noise Error}}{\boxed{\hat{\epsilon}_n(\hat{f})}} + \underset{\text{Empirical Target Error}}{\boxed{\hat{\epsilon}_t(\hat{f})}} + \mathcal{O}\left(\sqrt{\frac{d\log(\frac{(1-\beta)m}{d}) + \log(\frac{1}{\delta})}{(1-\beta)m}}\right) + \mathcal{O}\left(\sqrt{\frac{d\log(\frac{\beta m}{d}) + \log(\frac{1}{\delta})}{\beta m}}\right)\right],$$

where $\gamma = \sqrt{\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{1-\beta}}$, *and* $\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathbb{U}_n,\mathbb{U}_t)$ *is the empirical* $\mathcal{H}$*-divergence estimated from noise and target samples in* $\mathcal{Z}$.

$\mathcal{Z}$: **domain-shared representation space**

[1] Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. A theory of learning from different domains. Machine learning, 79:151–175, 2010.
[2] Li, B., Wang, Y., Zhang, S., Li, D., Keutzer, K., Darrell, T., and Zhao, H. Learning invariant representations and risks for semi-supervised domain adaptation. In CVPR, pp. 1104–1113, 2021.

Methodology: Design of Noise Adaptation Framework (NAF)

Based on Theorem 4.1, we design NAF as follows:

$$\min_{g_t, g_n, f} \mathcal{L}_t + \alpha \mathcal{L}_n + \beta \mathcal{L}_{n,t}$$

$\mathcal{L}_t$: Empirical risk of labeled target samples, associated with $\hat{\epsilon}_t(\hat{f})$

$\mathcal{L}_n$: Empirical risk of noise, associated with $\hat{\epsilon}_n(\hat{f})$

$\mathcal{L}_{n,t}$: Distributional discrepancy between projected domains, associated with $\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathbb{U}_n, \mathbb{U}_t)$

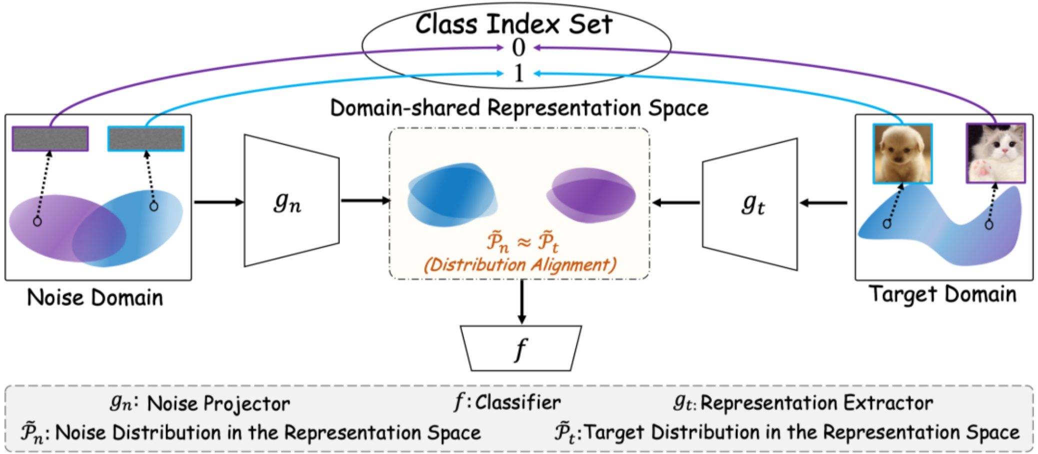
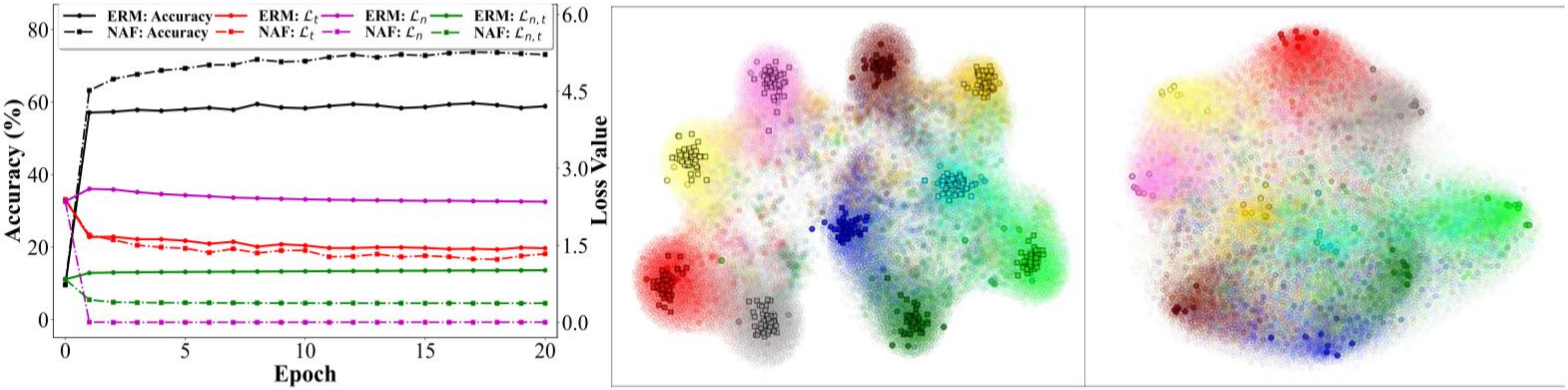


Figure 3. Under the SSNA setting, a randomly generated noise domain and a target domain share the same class index set. In NAF, noise and target samples are projected into a domain-shared representation space via a noise projector $g_n(\cdot)$ and a representation extractor $g_t(\cdot)$, respectively. By classifying noise according to the class indices in this representation space using a classifier $f(\cdot)$, the noise domain can induce a discriminative structure, which may facilitate alignment with the target domain and improve target representation separability.

Methodology: Does NAF Tighten the Bound?

NAF: minimizing $\mathcal{L}_t$, $\mathcal{L}_n$, and $\mathcal{L}_{n,t}$ **vs.** **ERM**: minimizing $\mathcal{L}_t$ only



(a) NAF vs. ERM *(b)* NAF *(c)* ERM

Figure 4. (a) Training loss and accuracy curves for NAF and ERM on CIFAR-10 with ResNet-18. $\mathcal{L}_t$ denotes the empirical risk of labeled target samples, $\mathcal{L}_n$ is the empirical risk of noise, and $\mathcal{L}_{n,t}$ measures the distributional discrepancy between domains. (b) Representations learned by NAF on CIFAR-10 with ResNet-18, where ■’ indicates noise representation; •’ and ‘○’ represent labeled and unlabeled target representations, respectively. (c) Representations learned by ERM on CIFAR-10 with ResNet-18, with the same symbol scheme as in (b). Colors correspond to different classes.

Discriminative structure of noise domain is essential!!!

Experiments: Performance Gain from NAF

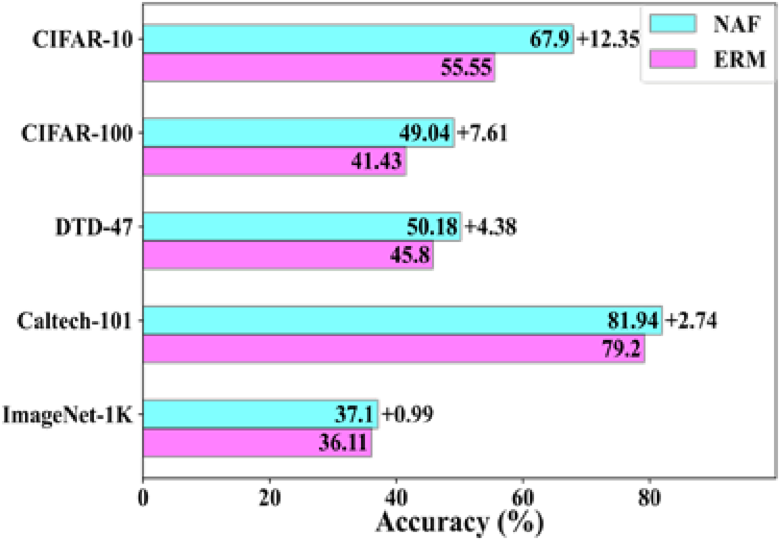


Figure 2. Accuracy (%) of NAF and ERM on five benchmark datasets, *i.e.*, CIFAR-10, CIFAR-100, DTD-47, Caltech-101, and ImageNet-1K, using ResNet-18. NAF outperforms ERM across all the datasets, demonstrating the effectiveness of NAF in transferring knowledge from the noise domain to the target domain.

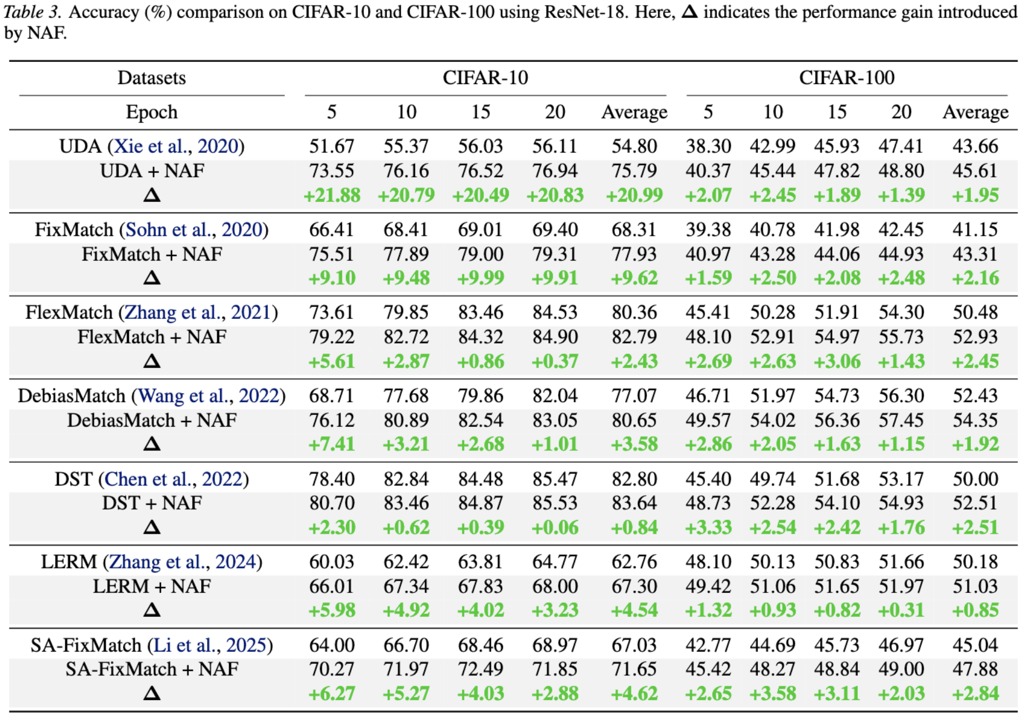
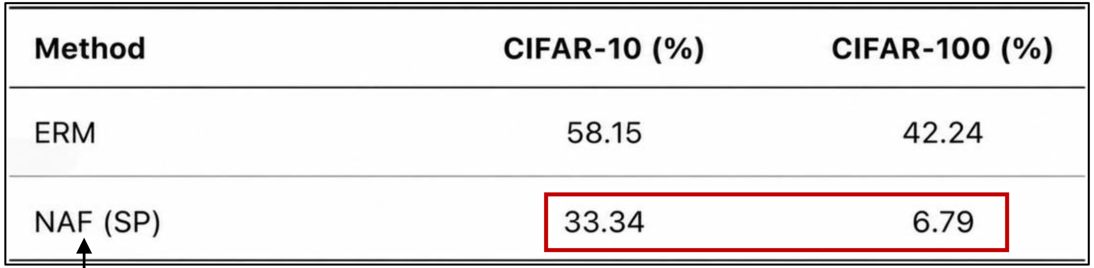


Table 3. Accuracy (%) comparison on CIFAR-10 and CIFAR-100 using ResNet-18. Here, **Δ** indicates the performance gain introduced by NAF.

Datasets	CIFAR-10					CIFAR-100				
Epoch	5	10	15	20	Average	5	10	15	20	Average
UDA (Xie et al., 2020)	51.67	55.37	56.03	56.11	54.80	38.30	42.99	45.93	47.41	43.66
UDA + NAF	73.55	76.16	76.52	76.94	75.79	40.37	45.44	47.82	48.80	45.61
Δ	+21.88	+20.79	+20.49	+20.83	+20.99	+2.07	+2.45	+1.89	+1.39	+1.95
FixMatch (Sohn et al., 2020)	66.41	68.41	69.01	69.40	68.31	39.38	40.78	41.98	42.45	41.15
FixMatch + NAF	75.51	77.89	79.00	79.31	77.93	40.97	43.28	44.06	44.93	43.31
Δ	+9.10	+9.48	+9.99	+9.91	+9.62	+1.59	+2.50	+2.08	+2.48	+2.16
FlexMatch (Zhang et al., 2021)	73.61	79.85	83.46	84.53	80.36	45.41	50.28	51.91	54.30	50.48
FlexMatch + NAF	79.22	82.72	84.32	84.90	82.79	48.10	52.91	54.97	55.73	52.93
Δ	+5.61	+2.87	+0.86	+0.37	+2.43	+2.69	+2.63	+3.06	+1.43	+2.45
DebiasMatch (Wang et al., 2022)	68.71	77.68	79.86	82.04	77.07	46.71	51.97	54.73	56.30	52.43
DebiasMatch + NAF	76.12	80.89	82.54	83.05	80.65	49.57	54.02	56.36	57.45	54.35
Δ	+7.41	+3.21	+2.68	+1.01	+3.58	+2.86	+2.05	+1.63	+1.15	+1.92
DST (Chen et al., 2022)	78.40	82.84	84.48	85.47	82.80	45.40	49.74	51.68	53.17	50.00
DST + NAF	80.70	83.46	84.87	85.53	83.64	48.73	52.28	54.10	54.93	52.51
Δ	+2.30	+0.62	+0.39	+0.06	+0.84	+3.33	+2.54	+2.42	+1.76	+2.51
LERM (Zhang et al., 2024)	60.03	62.42	63.81	64.77	62.76	48.10	50.13	50.83	51.66	50.18
LERM + NAF	66.01	67.34	67.83	68.00	67.30	49.42	51.06	51.65	51.97	51.03
Δ	+5.98	+4.92	+4.02	+3.23	+4.54	+1.32	+0.93	+0.82	+0.31	+0.85
SA-FixMatch (Li et al., 2025)	64.00	66.70	68.46	68.97	67.03	42.77	44.69	45.73	46.97	45.04
SA-FixMatch + NAF	70.27	71.97	72.49	71.85	71.65	45.42	48.27	48.84	49.00	47.88
Δ	+6.27	+5.27	+4.03	+2.88	+4.62	+2.65	+3.58	+3.11	+2.03	+2.84

Experiments: Role of Discriminative Structure of Noise Domain



Method	CIFAR-10 (%)	CIFAR-100 (%)
ERM	58.15	42.24
NAF (SP)	33.34	6.79

All noise collapses to a single point

Experiments: Noise Domain vs. Real Source Domain

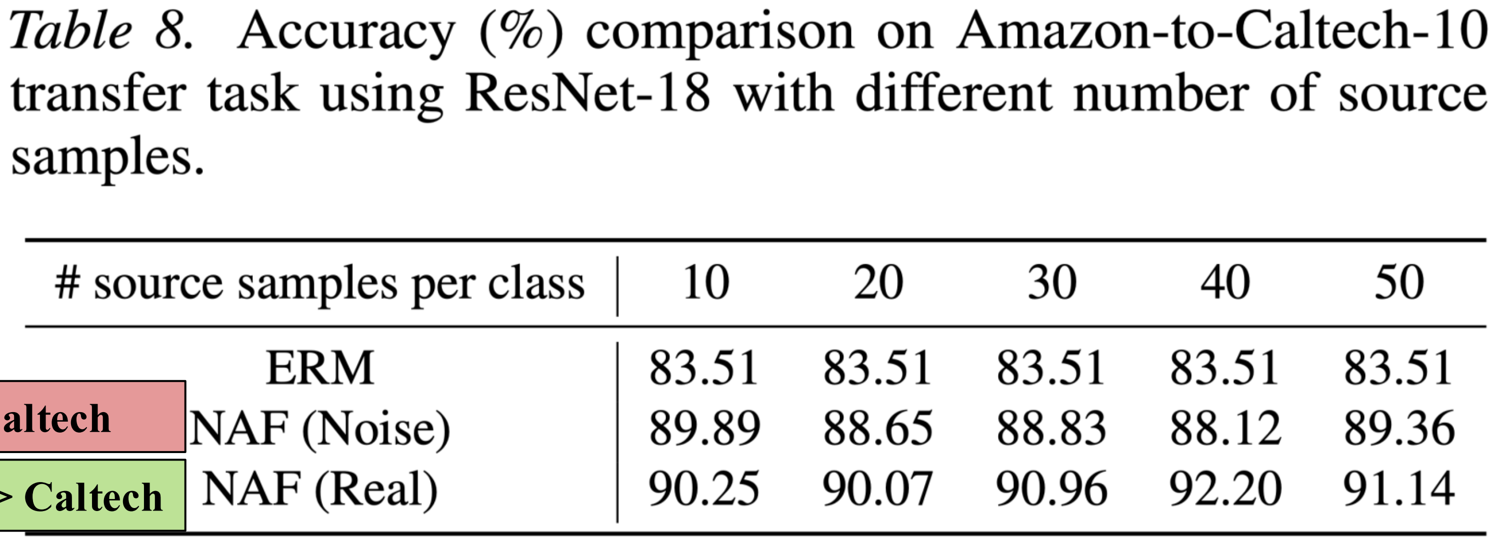


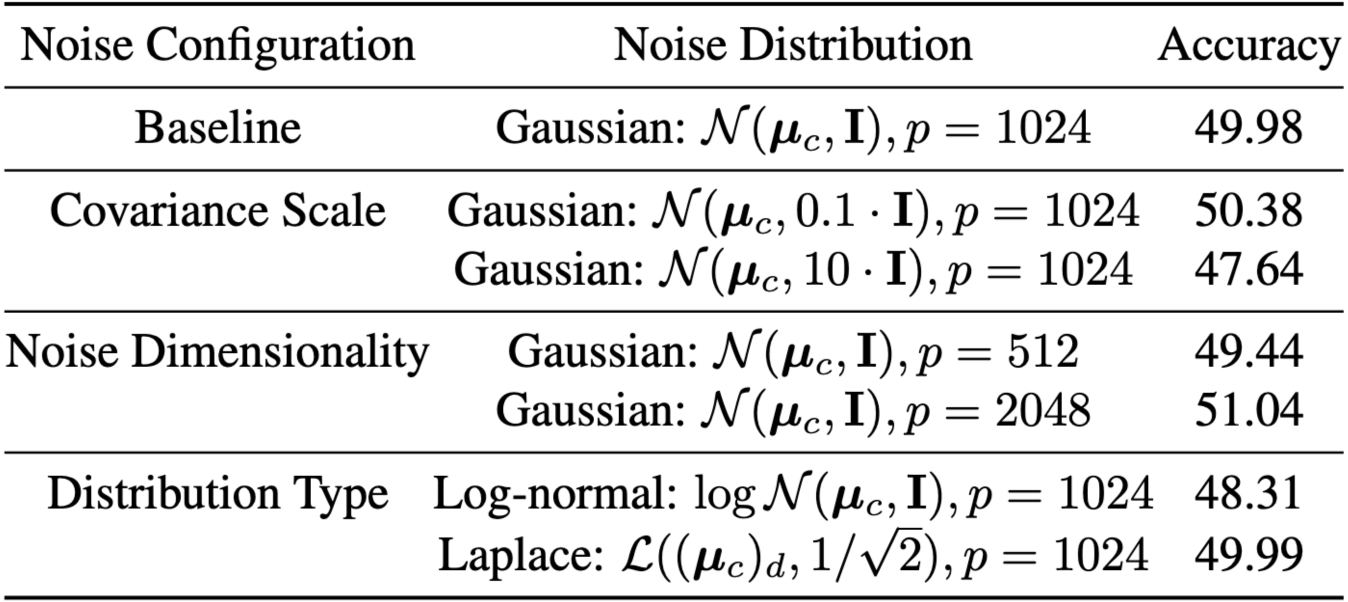
Table 8. Accuracy (%) comparison on Amazon-to-Caltech-10 transfer task using ResNet-18 with different number of source samples.

# source samples per class	10	20	30	40	50
ERM	83.51	83.51	83.51	83.51	83.51
NAF (Noise)	89.89	88.65	88.83	88.12	89.36
NAF (Real)	90.25	90.07	90.96	92.20	91.14

Noise --> Caltech

Amazon --> Caltech

Experiments: Role of Noise Distribution



Noise Configuration	Noise Distribution	Accuracy
Baseline	Gaussian: $\mathcal{N}(\boldsymbol{\mu}_c, \mathbf{I}), p = 1024$	49.98
Covariance Scale	Gaussian: $\mathcal{N}(\boldsymbol{\mu}_c, 0.1 \cdot \mathbf{I}), p = 1024$	50.38
	Gaussian: $\mathcal{N}(\boldsymbol{\mu}_c, 10 \cdot \mathbf{I}), p = 1024$	47.64
Noise Dimensionality	Gaussian: $\mathcal{N}(\boldsymbol{\mu}_c, \mathbf{I}), p = 512$	49.44
	Gaussian: $\mathcal{N}(\boldsymbol{\mu}_c, \mathbf{I}), p = 2048$	51.04
Distribution Type	Log-normal: $\log \mathcal{N}(\boldsymbol{\mu}_c, \mathbf{I}), p = 1024$	48.31
	Laplace: $\mathcal{L}((\boldsymbol{\mu}_c)_d, 1/\sqrt{2}), p = 1024$	49.99

Conclusion

1. The **discriminative structure** of source samples in the **representation space** is a key factor in transfer learning.

2. A deeper understanding of **representation space** may be essential for future progress in AI.

Thank you all for your time and participation!

Paper: https://arxiv.org/abs/2606.00558

Code: https://github.com/AIResearch-Group/SSNA

Contact: yaoyuan.hitsz@gmail.com