Stethoscope-Guided Supervised Contrastive Learning for Cross-domain Adaptation on Respiratory Sound Classification

June-Woo Kim[1,5], Sangmin Bae[2,5], Won-Yang Cho[3,5], Byungjo Lee[4,5], Ho-Young Jung[1,5]

[1] Department of AI, Kyungpook National University [2] KAIST AI

[3] SmartSound [4] National Cancer Center [5] MODULABS

kaen2891@knu.ac.kr, hoyjung@knu.ac.kr

TL;DR

- We introduce Stethoscope-Guided Supervised Contrastive Learning (SG-SCL), which aims to alleviate decreased performance arising from different stethoscope (recording device) types from the cross-domain perspective.

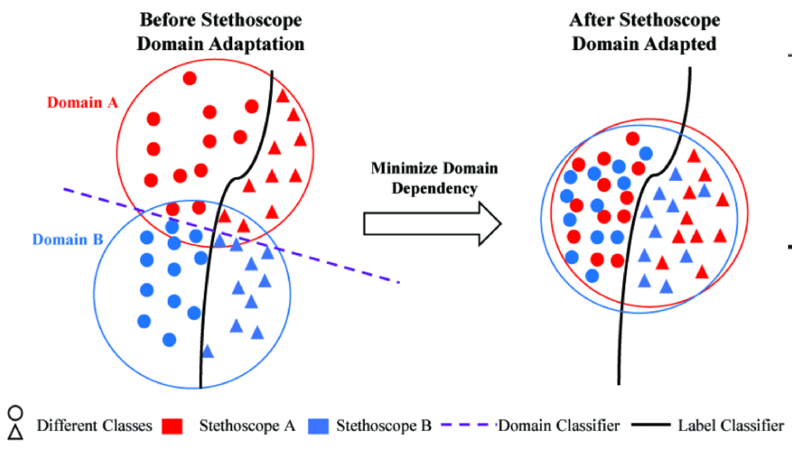


Fig. 1: Overview of our works

	label	train	test	sum
lung	Normal	2,063	1,579	3,642
	Crackle	1,215	649	1,864
	Wheeze	501	385	886
	Both	363	143	506
device	Meditron	997	459	1,459
	LittC2SE	594	0	594
	Litt3200	41	461	502
	AKGC417L	2,510	1,836	4,346

Table 1: ICBHI dataset statistics

Motivation

1. Domain Adaptation (a.k.a. DANN [1])

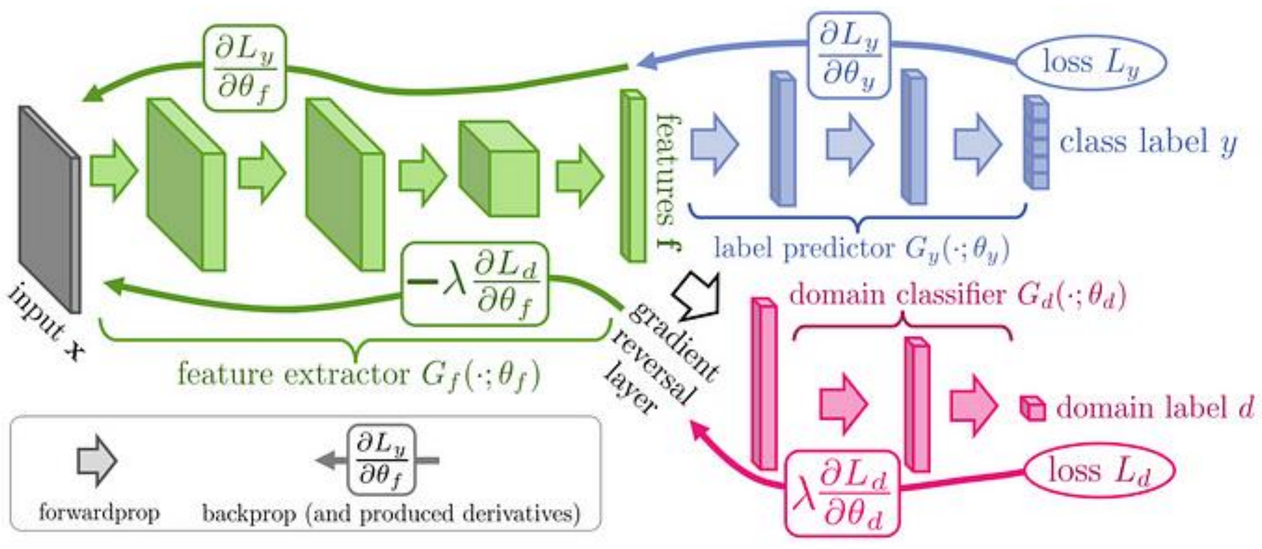


Fig. 2: Domain-adversarial training of neural networks architecture [1]

2. Supervised Contrastive Learning (SCL) [2]

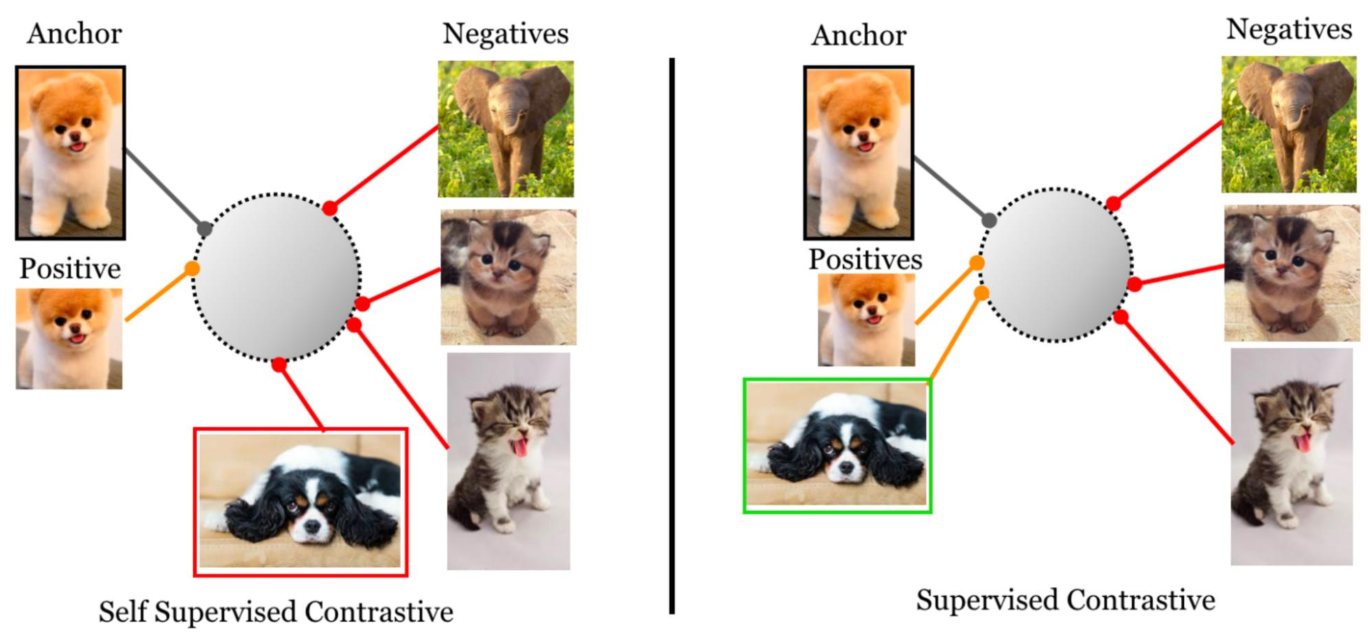


Fig. 3: Self-supervised contrastive learning vs. supervised contrastive learning [2]

Method

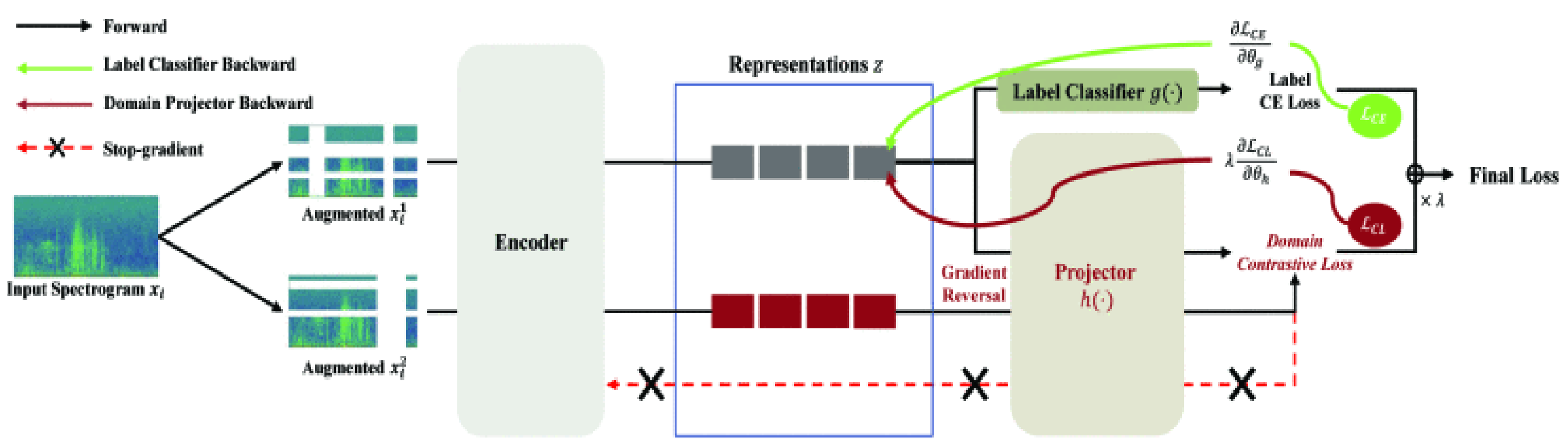


Fig. 4: Overall illustration of proposed SG-SCL for cross-domain adaptation

1. DAT (Domain Adaptation Training)

- $\mathcal{L}_{\text{DAT}} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{DA}}$

where $\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{n} y_i \log(\hat{y}_i)$ and $\mathcal{L}_{\text{DA}} = -\sum_{i=1}^{n} d_i \log(\hat{d}_i)$ are CE loss with lung sound label y and stethoscope domain label d, and the predicted probabilities $\hat{y}$ and $\hat{d}$ are obtained by label and domain classifiers, λ is a domain regularization parameter with reversal gradients, respectively.

2. SG-SCL (Stethoscope-Guided SCL)

- $\mathcal{L}_{\text{CL}} = \sum_{i \in I} -\log\left\{\frac{1}{|P(i)|}\sum_{p \in P(i)} \frac{e(h(z_i)\cdot \text{sgd}(h(z_p))/\tau)}{\sum_{a \in A(i)} e(h(z_i)\cdot \text{sgd}(h(z_a))/\tau}\right\}$

where index $\boldsymbol{i}$ is the anchor index from $A(i) \equiv I \setminus \{i\}$, $\boldsymbol{P(i)} \equiv \{p \in A(i): d_p = d_i\}$ represents the collection of all positive samples within the multi-viewed batch that corresponds to the $\boldsymbol{i}$–th sample, $\boldsymbol{z}$ is the encoder output, $\boldsymbol{e}$ and $\textbf{sgd}(\cdot)$ denote the exponential function and stop-gradient operation, $\boldsymbol{h}$ is projector, both $\boldsymbol{z}$ and $\boldsymbol{h}$ have the same dimension, and the final loss is $\mathcal{L}_{CE} + \lambda\mathcal{L}_{CL}$.

[1] Ganin, Yaroslav, et al. "Domain-adversarial training of neural networks." Journal of machine learning research 2016.
[2] Khosla, Prannay, et al. "Supervised contrastive learning." NeurIPS 2020.

Quantitative Results

anchor	target	S_p (%)	S_e (%)	**Score** (%)
z_i	z_p	**89.84**$_{\pm 3.92}$	13.61$_{\pm 5.67}$	51.73$_{\pm 1.53}$
z_i	$h(z_p)$	76.30$_{\pm 1.55}$	**44.60**$_{\pm 2.20}$	60.45$_{\pm 0.44}$
$h(z_i)$	z_p	81.87$_{\pm 3.20}$	39.83$_{\pm 1.05}$	60.85$_{\pm 1.60}$
$h(z_i)$	$h(z_p)$	77.25$_{\pm 3.43}$	36.35$_{\pm 17.97}$	60.78$_{\pm 0.85}$
$h(z_i)$	$\text{sgd}(z_p)$	76.31$_{\pm 6.35}$	43.79$_{\pm 4.38}$	60.05$_{\pm 1.19}$
$h(z_i)$	$\text{sgd}(h(z_p))$	79.87$_{\pm 8.89}$	43.55$_{\pm 5.93}$	**61.71**$_{\pm 1.61}$

Table. 2: SG-SCL performance based on two factors: anchor representation z_i and that of target z_p

- We got the best result when the stop-gradient was applied to the target representations $\boldsymbol{z_p}$, which are the second augmented samples from same source in the multi-viewed batch.
- We found that allowing gradient flows through both anchor z_i and target representations z_p simultaneously did not show an improvement (4th rows in Table 2).

architecture	method	pretrain	S_p (%)	S_e (%)	**Score** (%)
EfficientNet	CE	IN	73.48$_{\pm 5.93}$	**39.24**$_{\pm 2.43}$	57.46$_{\pm 1.05}$
	DAT		**89.99**$_{\pm 7.15}$	11.78$_{\pm 7.05}$	50.89$_{\pm 0.69}$
	SG-SCL		81.58$_{\pm 3.38}$	33.85$_{\pm 3.75}$	**57.72**$_{\pm 1.32}$
ResNet18	CE	IN	74.72$_{\pm 3.43}$	33.95$_{\pm 3.88}$	54.33$_{\pm 0.91}$
	DAT		**91.26**$_{\pm 6.32}$	12.03$_{\pm 5.86}$	51.51$_{\pm 0.34}$
	SG-SCL		75.58$_{\pm 6.36}$	**34.63**$_{\pm 5.98}$	**55.10**$_{\pm 1.18}$
CNN6	CE	IN	80.13$_{\pm 2.64}$	35.91$_{\pm 3.52}$	58.10$_{\pm 0.59}$
	DAT		**88.57**$_{\pm 5.66}$	13.73$_{\pm 6.47}$	51.15$_{\pm 0.45}$
	SG-SCL		78.16$_{\pm 3.49}$	**38.05**$_{\pm 4.41}$	**58.11**$_{\pm 0.64}$
AST	CE	IN + AS	77.14$_{\pm 3.35}$	41.97$_{\pm 2.21}$	59.55$_{\pm 0.88}$
	DAT		77.11$_{\pm 7.20}$	41.99$_{\pm 5.00}$	59.81$_{\pm 1.25}$
	SG-SCL		**79.87**$_{\pm 8.89}$	**43.55**$_{\pm 5.93}$	**61.71**$_{\pm 1.61}$

Table 3: Respiratory sound classification performance according to different architectures using CE, DAT, and SG-SCL

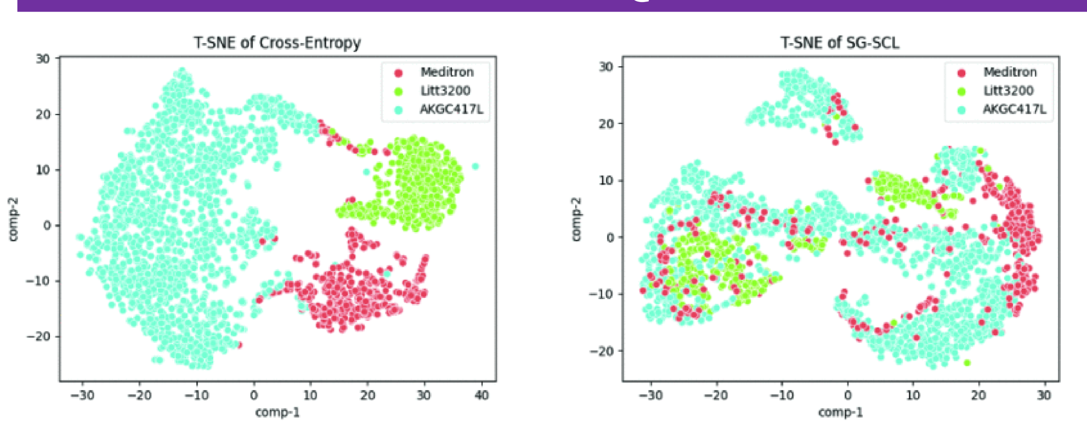
- We trained CE, DAT, and SG-SCL methods on different architectures with the ICBHI dataset under the same conditions without additional learning techniques.
- As a result, the proposed SG-SCL method achieved the best Score in all architectures.

	method	architecture	pretrain	venue	S_p (%)	S_e (%)	**Score** (%)
4-class eval.	CNN-MoE [19]	C-DNN	-	*JBHI'21*	72.40	21.50	47.00
	RespireNet [3] (CBA+BRC+FT)	ResNet34	IN	*EMBC'21*	72.30	40.10	56.20
	Ren *et al.* [4]	CNN8-Pt	-	*ICASSP'22*	72.96	27.78	50.37
	Chang *et al.* [20]	CNN8-dilated	-	*INTERSPEECH'22*	69.92	35.85	52.89
	Wang *et al.* [5] (Splice)	ResNeSt	IN	*ICASSP'22*	70.40	40.20	55.30
	Nguyen *et al.* [6] (CoTuning)	ResNet50	IN	*TBME'22*	79.34	37.24	58.29
	Moummad *et al.* [16] (SCL)	CNN6	AS	*arXiv'22*	75.95	39.15	57.55
	Bae *et al.* [7] (Fine-tuning)	AST	IN + AS	*INTERSPEECH'23*	77.14	41.97	59.55
	Bae *et al.* [7] (Patch-Mix CL)	AST	IN + AS	*INTERSPEECH'23*	**81.66**	43.07	**62.37***
	DAT [ours]	AST	IN + AS	*ICASSP'24*	77.11$_{\pm 7.2}$	42.50$_{\pm 5.39}$	59.81$_{\pm 1.25}$
	SG-SCL [ours]	AST	IN + AS	*ICASSP'24*	79.87$_{\pm 8.89}$	**43.55**$_{\pm 5.93}$	61.71$_{\pm 1.61}$
2-class eval.	CNN-MoE [19]	C-DNN	-	*JBHI'21*	72.40	37.50	54.10
	Nguyen *et al.* [6] (CoTuning)	ResNet50	IN	*TBME'22*	79.34	50.14	64.74
	Bae *et al.* [7] (Fine-tuning)	AST	IN + AS	*INTERSPEECH'23*	77.14	56.40	66.77
	Bae *et al.* [7] (Patch-Mix CL)	AST	IN + AS	*INTERSPEECH'23*	**81.66**	55.77	68.71*
	DAT [ours]	AST	IN + AS	*ICASSP'24*	77.11$_{\pm 7.2}$	56.98$_{\pm 7.42}$	67.04$_{\pm 1.29}$
	SG-SCL [ours]	AST	IN + AS	*ICASSP'24*	79.87$_{\pm 8.89}$	**57.97**$_{\pm 8.96}$	**68.93**$_{\pm 1.47}$

Table 4: Comprehensive comparison of the ICBHI dataset for the respiratory sound classification task (60-40% official split)

- In the 4-class evaluation, the proposed SG-SCL achieved a 61.71% Score.
- Our SG-SCL achieved a state-of-the-art Score with a 68.93% in the 2-class evaluation.
- Moreover, the proposed SG-SCL obtained the highest Sensitivity ($\boldsymbol{S_e}$) in both 4-class and 2-class evaluations, suggesting that our method is the most accurate model for actually classifying abnormal respiratory sounds.

Qualitative Results



(a) AST fine-tuning (b) Proposed SG-SCL

Fig. 5: T-SNE results on ICBHI test set for stethoscope labels

Paper Code

- The AST fine-tuning results in Fig. 5 (a) show that the representations are clustered according to the stethoscopes, while our SG-SCL results in Fig. 5 (b) are well mixed regardless of the recording device type.