# Slimmed Asymmetrical Contrastive Learning and Cross Distillation for Lightweight Model Training


[1]**Jian Meng,** [2]Li Yang, [3]Kyungmin Lee, [3]Jinwoo Shin, [4]Deliang Fan, [1]Jae-sun Seo

[1]Cornell Tech, USA, [2]University of North Carolina at Charlotte, USA

[3]KAIST, South Korea, [4]John Hopkins University, USA

Cornell University

NEURAL INFORMATION PROCESSING SYSTEMS

# Contrastive (Self-supervised) Learning

- Unsupervised representation learning leads to strong performance in various downstream tasks
  - Training ResNet-50 on ImageNet-1K with supervised and self-supervised learning (SSL):

| Method | CIFAR-10 | CIFAR-100 | Aircraft | Flowers | Birdsnap |
|---|---|---|---|---|---|
| Supervised (from scratch) | 94.8 | 78.2 | 83.8 | 92.0 | 76.0 |
| Supervised-Fine-tuned [1] | 97.5 | 86.1 | 86.0 | **97.6** | 75.8 |
| **BYOL-SSL-Fine-tuned [1]** | **97.8** | **86.4** | **88.1** | 97.0 | **76.3** |

- Learning powerful visual representation comes with cost…
  - The recent contrastive learning-based self-supervised learning requires wide and deep models.
  - Lightweight / sparse model (e.g., MobileNet) are **largely ignored** in contrastive learning.

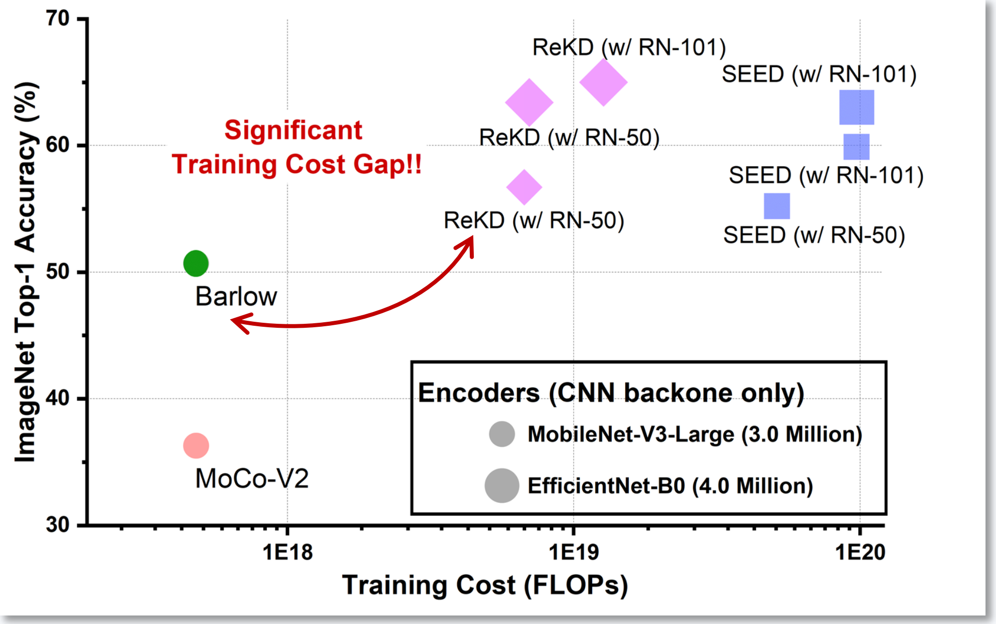
Powerful contrastive learning requires large-sized models, while the small-scale vision tasks are widely existing in the resource constrained edge devices.
**Strong vision learners ≠ Superior compatibility on edge**

Cornell University

[1] Grill, Jean-Bastien, et al. "Bootstrap your own latent-a new approach to self-supervised learning," NeurIPS, 2020

# Lightweight Contrastive Learning

- Insufficient learnability of model → Knowledge distillation (KD) with a **strong teacher**
  - SEED (*Fang, ICLR'21*): Pretrained teacher with CL (800 ep), distillation without labels (200 ep).
  - ReKD *(Zheng, AAAI'22)*: Pretrained teacher with CL (800 ep), distillation with relation knowledge (200 ep)
  - DisCo (*Gao, ECCV'22): Frozen Pre-trained teacher* + distilling the target student with both "teacher" and "mean student"




- Despite the distillation schemes, a strong teacher becomes an almost mandatory requirement.
  - *Extreme training cost* compared to vanilla contrastive learning.

*Is there a contrastive learning algorithm that can train the high-performance lightweight model without using a mega-sized teacher?*

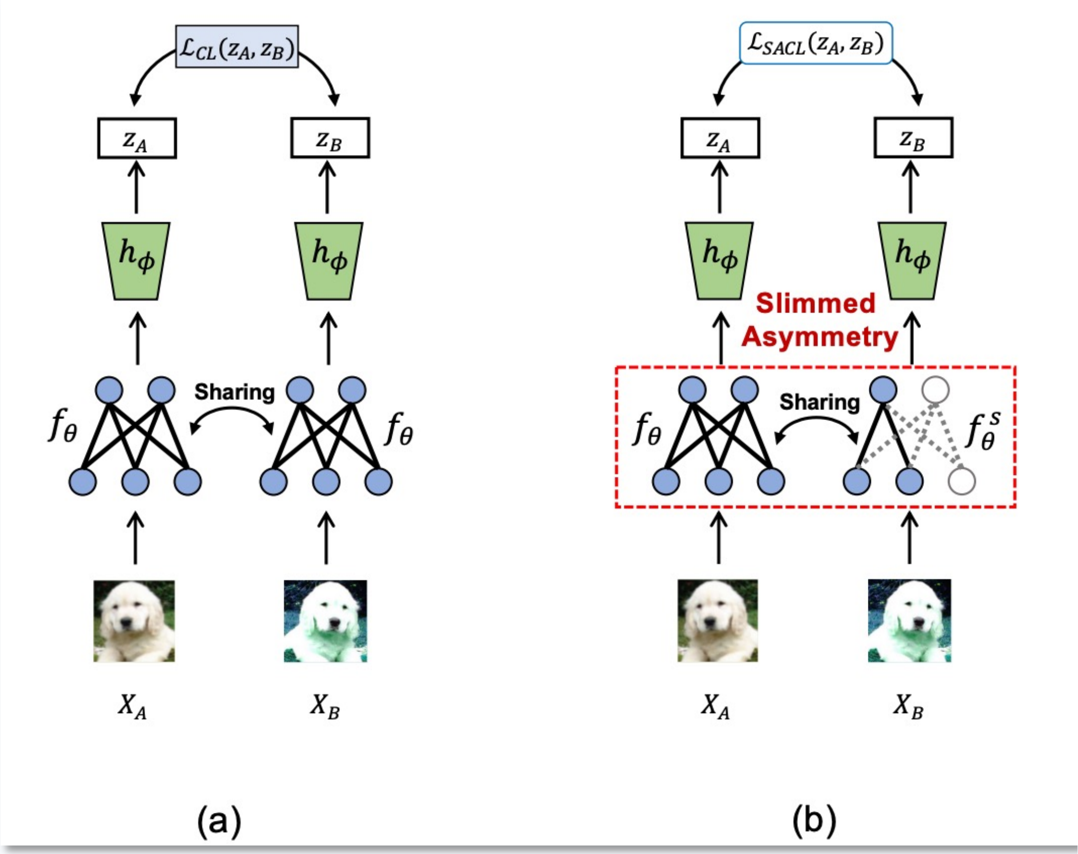
# Slimmed Asymmetric Contrastive Learning (SACL)

- *Question:* The necessity of employing the large ResNet teacher **haven't been fully justified.**
  - *Can we have a lighter teacher for faster training?*

- Slimmed Asymmetric Contrastive Learning (SACL)
  - Lightweight model can be considered as a subset model "sliced" from a wide, full-sized *host model.*
  - The *host model* $\theta$ is sliced by removing a *unified* amount of input **and** output channels to formulate $\theta_s$:

$$\theta_s \subset \theta \text{ and } \theta_s = \theta \cdot \mathcal{M}$$

  - Where $\mathcal{M}$ is the weight mask that disables **both** input and output channel with a given slice ratio (K×-1×)
  - Input $X_A$ and $X_B$ are separately encoded by the *host* and the *slimmed* encoder




(a) Normal contrastive learning

(b) Proposed slimmed asymmetry contrastive learning

# Cross Distillation (XD)

- Asymmetry is not the “one-and-done” solution for lightweight CL due to the sparsity-induced distortion.
  - *How to further enhance the training performance?*

- Given the asymmetrical contrastive encoders $f_\theta$ and $f_\theta^s$,
  - We first encode $X^A$ and $X^B$ based on SACL, leading to the embeddings $z^A$ and $z^B$

$$X^A \rightarrow f_\theta \rightarrow z^A$$

$$X^B \rightarrow f_\theta^s \rightarrow z^B$$

  - Subsequently, we **freeze** both $f_\theta$ and $f_\theta^s$, while **reversing** the order of inputs for encoding

$$X^B \rightarrow [f_\theta] \rightarrow [\hat{z}^B]$$

$$X^A \rightarrow [f_\theta^s] \rightarrow [\hat{z}^A]$$

Where $[\cdot]$ represents the frozen encoder.

Cornell University.

# Cross Distillation (Continued)

- Now we have a pair of latent code (e.g., $z^A$ and $[\hat{z}^A]$) for each input (e.g., $X^A$) that contains the latent information distorted by sparsity **only**.
    - To minimize the discrepancy, we compute the cross-distillation loss $\mathcal{L}_{\text{CD}}$ as:

$$\mathcal{L}_{\text{CD}} = \frac{\mathcal{L}^A_{\text{CD}}(z^A, [\hat{z}^A]) + \mathcal{L}^B_{\text{CD}}(z^B, [\hat{z}^B])}{2}$$

- We define the total training loss as the weighted combination between contrastive loss $\mathcal{L}_{\text{SACL}}$ and cross-distillation loss $\mathcal{L}_{\text{CD}}$

$$\mathcal{L} = \alpha \mathcal{L}_{\text{SACL}} + (1 - \alpha) \mathcal{L}_{\text{CD}}$$

Loss contains augmentation. & asymmetry

Loss that minimizes asymmetry only

Cornell University

# Cross Distillation (Continued)

- Why cross-distillation?
  - Cross distillation enables the optimization across the **feature dimensions** inside latent space

- When the encoders are completely dense (no SACL):
  - $C_{ii}^{AA} \rightarrow 1.0$, *inner-correlation* loss → 0.0
  - Minimizing the cross-distillation (XD) loss ⇔ Decorrelate the features across different dimensions

| Method | Linear Eval. Acc (%) | Training Epochs | Pretrained Teacher |
|---|---|---|---|
| ReKD | 56.70 | 200 | ResNet-50 |
| SEED | 55.20 | 200 | ResNet-50 |
| **XD (Ours)** | **57.16** | **100** | **N/A** |

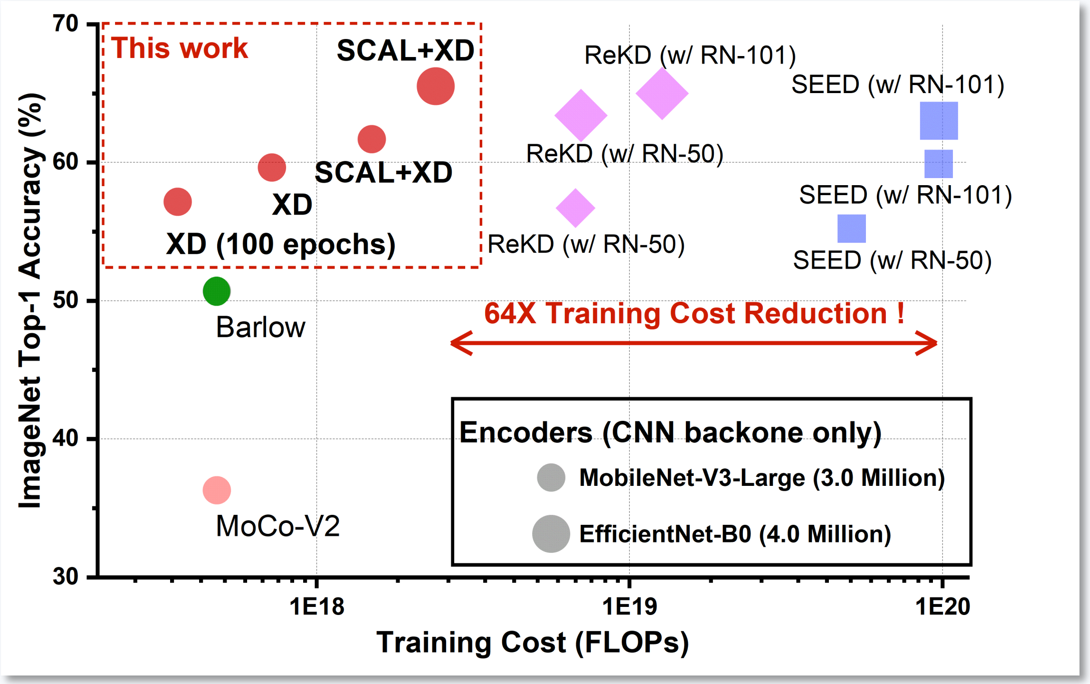
MobileNet-V3 ImageNet-1K Linear evaluation accuracy comparison between XD (proposed) and SOTA methods

- Minimizing $\mathcal{L}_{\mathrm{CD}}$ avoids the aliasing feature across different dimensions
  - Decorrelation at the embedding level ultimately has a decorrelation effect at the representation level
  - Outperform previous SOTA method without heavy distillation and pre-trained teacher

Cornell University.

# Slimmed Asymmetrical CL (SACL) + Cross Distillation (XD)

- New SoTA Performance on lightweight contrastive learning
  - **64×** training cost reduction compared to SOTA lightweight contrastive learning method.
  - **Train from scratch** with lightweight encoder (e.g., EfficientNet, MobileNet).




Substantial training cost reduction of the proposed method

| Method | Encoder | Linear Eval. (%) | Epochs | Pre-train | Teacher | Training FLOPs (e+17) |
|---|---|---|---|---|---|---|
| ‡**SACL-XD (Ours)** | Eff-B0 (1.5×-1×) | **65.32 (+2.12)** | 200 | ✗ | - | **24 (2.9× ↓)** |
| §**SACL-XD (Ours)** | Mob-V3 (1.5×-1×) | **61.69 (+1.79)** | 200 | ✗ | - | **15 (64.7× ↓)** |
| **SACL-XD (Ours)** | Mob-V1 (1.5×-1×) | **59.34** | 200 | ✗ | - | **19** |
| **XD only (Ours)** | Mob-V3 (1×) | **59.42** | 200 | ✗ | - | **7.2** |
| **XD only (Ours)** | Mob-V3 (1×) | **57.16** | **100** | ✗ | - | **3.6** |
| **XD only (Ours)** | Mob-V1 (1×) | **55.84** | **100** | ✗ | - | **9.0** |
| §SSL-Small [24] | Mob-V3 (1×) | 48.70 | 200 | ✗ | - | 67 |
| §SSL-Small [24] | Eff-B0 (1×) | 55.90 | 200 | ✗ | - | 67 |
| ReKD [32] | Mob-V3 (1×) | 56.70 | 200 | ✗ | ResNet-50 | 67 |
| ReKD [32] | Mob-V3 (1×) | 59.60 | 200 | ✗ | ResNet-101 | 125 |
| ReKD [32] | Eff-B0 (1×) | 63.40 | 200 | ✗ | ResNet-50 | 70 |
| OSS [9] | Eff-B0 (1×) | 64.10 | 800+200 | ✗ | ResNet-50 | 67 |
| *SEED [14] | Mob-V3 (1×) | 55.20 | 800+200 | ✓ | ResNet-50 | 512 |
| *SEED [14] | Mob-V3 (1×) | 59.90 | 800+200 | ✓ | ResNet-101 | 971 |
| *SEED [14] | Eff-B0 (1×) | 61.30 | 800+200 | ✓ | ResNet-50 | 516 |
| †MoCo-V2 [7] | Mob-V3 (1×) | 36.30 | 200 | ✗ | - | 4.8 |
| †MoCo-V2 [7] | Eff-B0 (1×) | 42.20 | 200 | ✗ | - | 8.5 |

Universal SOTA performance for both XD and SACL + XD

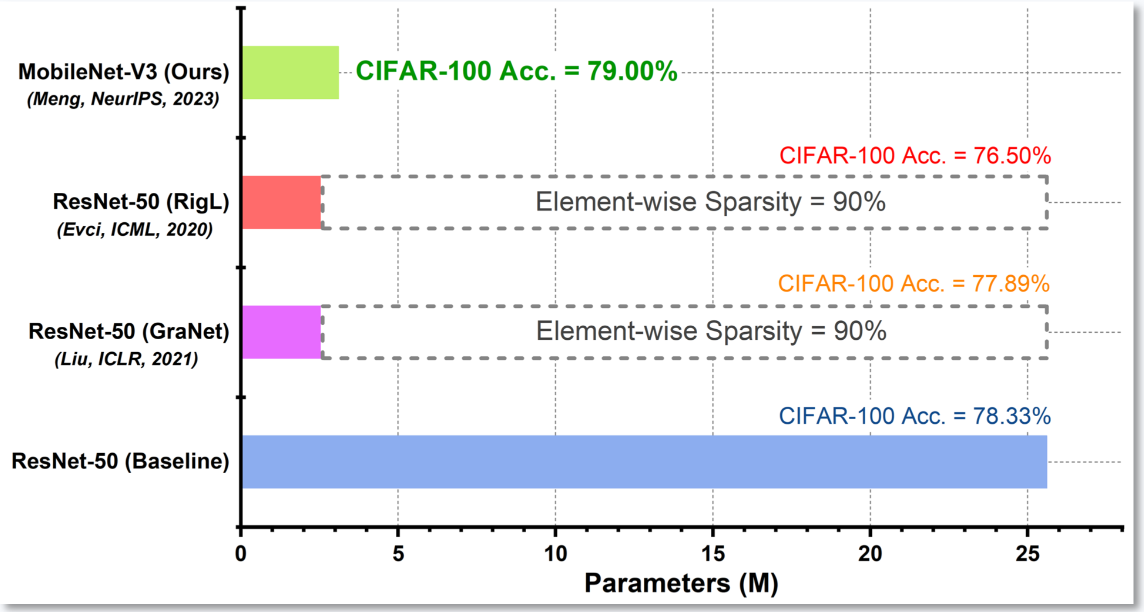
# Slimmed Asymmetrical CL (SACL) + Cross Distillation (XD)

- New SoTA Performance on downstream vision tasks

| Method | Encoder | CIFAR-10 | CIFAR-100 | Aircraft | Flowers | Cars |
|---|---|---|---|---|---|---|
| Supervised (from scratch) | Mob-V3 | 92.97 | 73.69 | 65.37 | 79.89 | 68.18 |
| Supervised-FT | Mob-V3 | 94.53 | 78.56 | 68.29 | 89.94 | 82.43 |
| **XD (Ours, 100 ep)** | Mob-V3 | **94.80** | **79.00** | **71.39** | **90.05** | **82.77** |
| **SACL + XD (Ours)** | Mob-V1 (1.5× – 1×) | **94.92** | **79.64** | **72.21** | **90.48** | **83.14** |

Downstream performance of the lightweight model pre-trained on ImageNet-1K + minimum fine-tuning

- From the efficient inference point of view…
  - High-performance unsupervised pre-training of SACL+XD empower the lightweight model with strong visual representation
  - Arguably, the superior downstream performance of lightweight model outperforms supervised pruning
  - Dedicated sparse accelerator is **NOT** required




Cornell University

# Conclusion

- We propose a novel contrastive learning algorithm which trains the powerful lightweight encoder without introducing strong teacher

- We have investigated the lightweight contrastive learning from the perspectives of latent space and aliasing reduction.

- With the proposed cross-distillation and slimmed asymmetric CL, our method empower the lightweight model with highly efficient contrastive learning, leading to the strong accuracy-efficiency tradeoff.

# Thank you!