Effective Self-supervised Pre-training on Low-compute Networks without Distillation

Fuwen Tan
Samsung AI Center

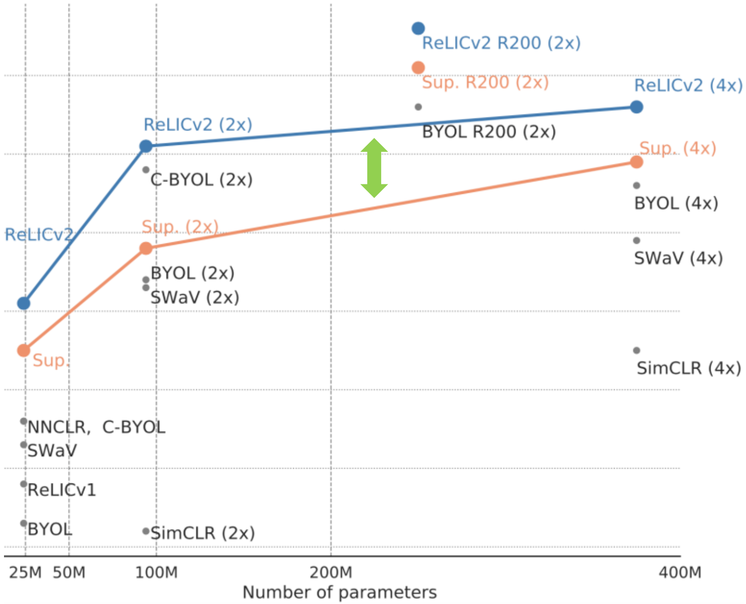
Fatemeh Saleh
Microsoft

Brais Martinez
Samsung AI Center

ICLR
International Conference On Learning Representations

Self-supervised learning (SSL) on low-compute networks

The power of SSL is unlocked for large visual backbones

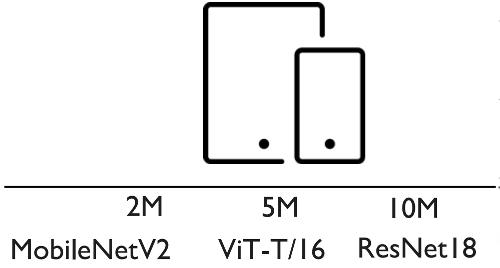


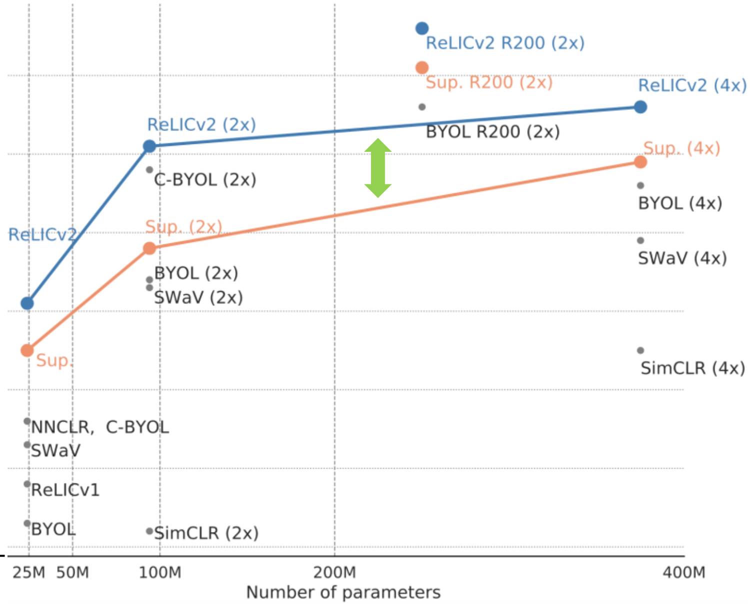
[Tomasev et al]

[Tomasev et al.] Pushing the limits of self-supervised ResNets: Can we outperform supervised learning without labels on ImageNet? *ICMLW 2022*

Self-supervised learning (SSL) on low-compute networks

SSL on low-compute visual backbones ?

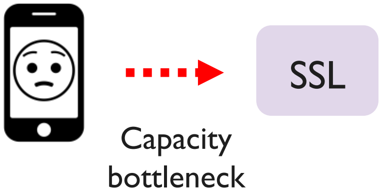




[Tomasev et al]

[Tomasev et al.] Pushing the limits of self-supervised ResNets: Can we outperform supervised learning without labels on ImageNet? *ICMLW 2022*

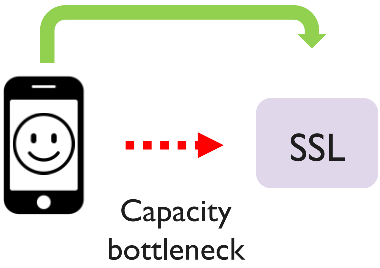
Previous research

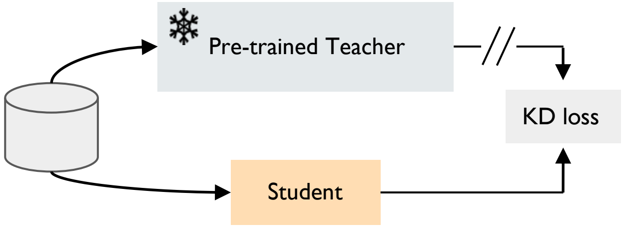


Previous research

KD w different losses

Knowledge Distillation (KD)

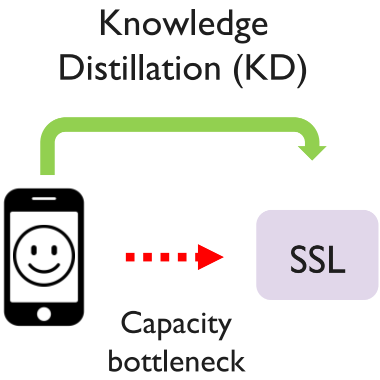




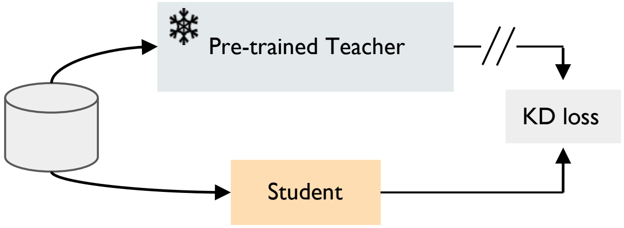
[Koohpayegani et al.]
[Fang et al.]
[Xu et al.]
[Gao et al.]

[Koohpayegani et al.] CompRess: Self-Supervised Learning by Compressing Representations. *NeurIPS 2020.*
[Fang et al.] SEED: Self-supervised Distillation for Visual Representation. *ICLR 2021.*
[Xu et al.] Bag of Instances Aggregation Boosts Self-supervised Distillation. *ICLR 2022.*
[Gao et al.] DisCo: Remedy Self-supervised Learning on Lightweight Models with Distilled Contrastive Learning. *ECCV 2022.*

Previous research



KD w different losses



[Koohpayegani et al.]
[Fang et al.]
[Xu et al.]
[Gao et al.]

Pros

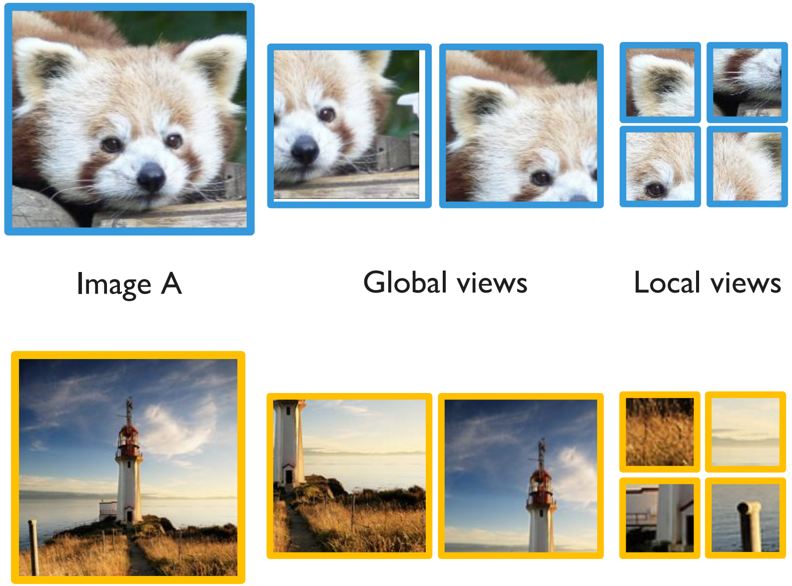
- Re-use strong teachers
- Easier to optimize

Cons

- New pre-training data
  - Teacher pre-training
- New on-device tasks
  - Continual Learning
  - Federated Learning

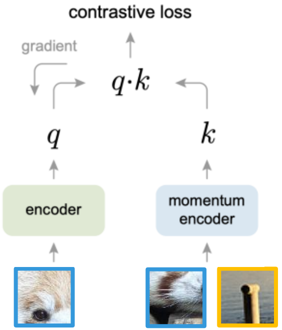
[Koohpayegani et al.] CompRess: Self-Supervised Learning by Compressing Representations. *NeurIPS 2020.*
[Fang et al.] SEED: Self-supervised Distillation for Visual Representation. *ICLR 2021.*
[Xu et al.] Bag of Instances Aggregation Boosts Self-supervised Distillation. *ICLR 2022.*
[Gao et al.] DisCo: Remedy Self-supervised Learning on Lightweight Models with Distilled Contrastive Learning. *ECCV 2022.*

State-of-the-art SSL

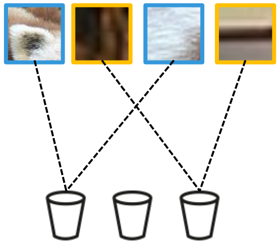


Self-supervision:

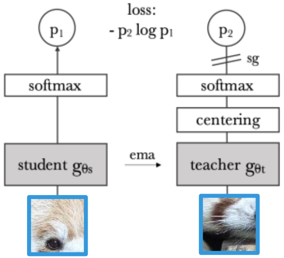
aligning different views of the same image



Contrastive learning
[Chen et al.]



Clustering
[Caron et al.]



Feature matching
[Caron et al.]

[Chen et al.] Improved Baselines with Momentum Contrastive Learning.
[Caron et al.] Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. *NeurIPS 2020*.
[Caron et al.] Emerging Properties in Self-Supervised Vision Transformers. *ICCV 2021*.

State-of-the-art SSL

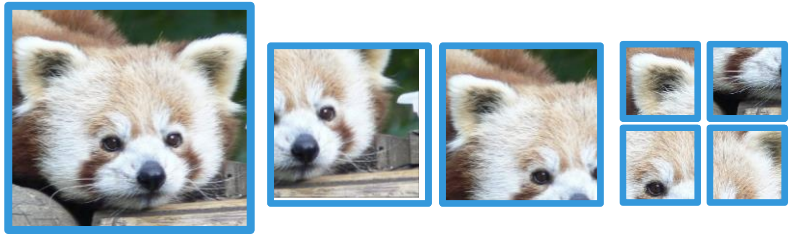


Image A | Global views | Local views

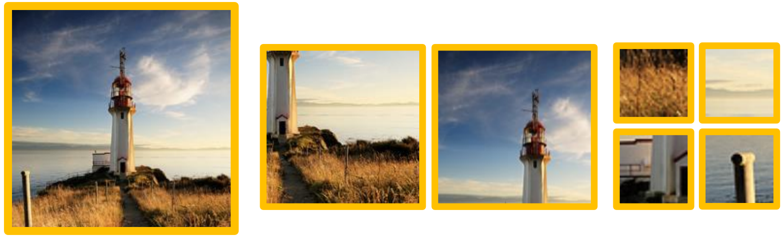
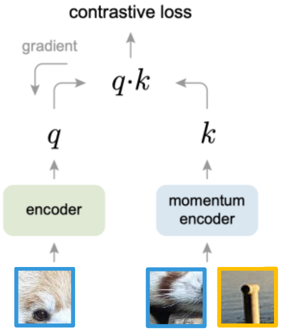


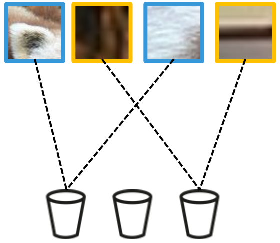
Image B | Global views | Local views

Self-supervision:

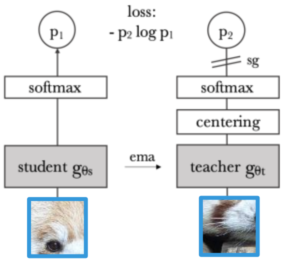
aligning different views of the same image



Contrastive learning
[Chen et al.]



Clustering
[Caron et al.]

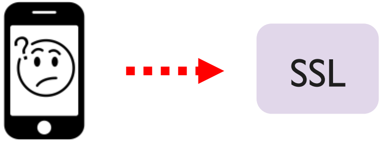


Feature matching
[Caron et al.]

A form of "*regularization*"

[Chen et al.] Improved Baselines with Momentum Contrastive Learning.
[Caron et al.] Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. *NeurIPS 2020*.
[Caron et al.] Emerging Properties in Self-Supervised Vision Transformers. *ICCV 2021*.

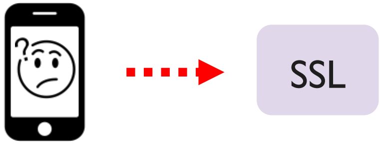
Diagnosing SSL for lightweight networks



Previous consensus:
Capacity bottleneck

Our finding:
Model complexity vs regularization strength

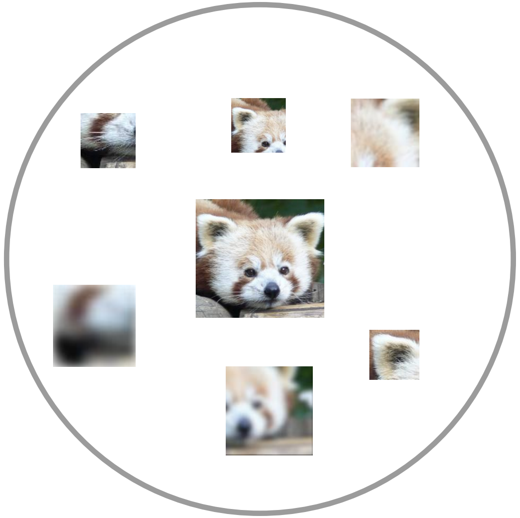
Diagnosing SSL for lightweight networks



Aligning views
of diverse scales & contexts

Previous consensus:
Capacity bottleneck

Our finding:
Model complexity vs regularization strength

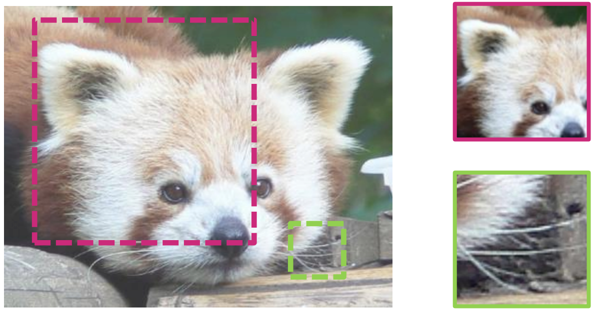


Large networks

Small networks

Regularization strength: from the view matching perspective

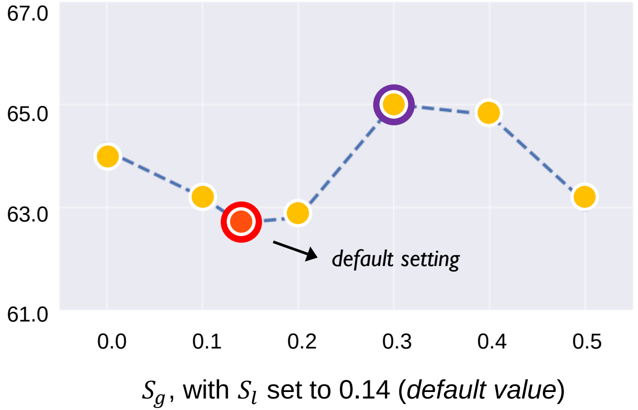
Aligning views of different
"crop scales"

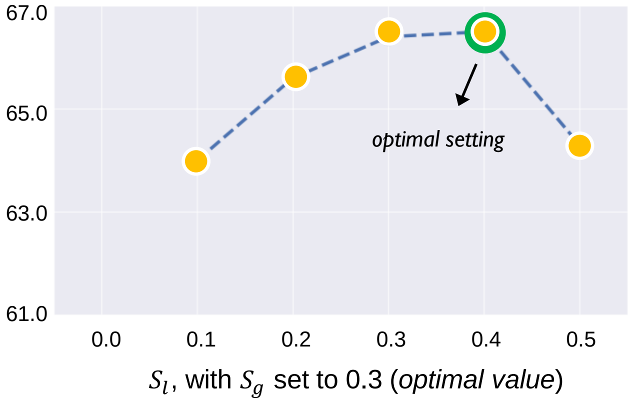


Global view B:
$S_g \sim 1.0$
the area of A

Local view B:
$0.05 \sim S_l$
the area of A

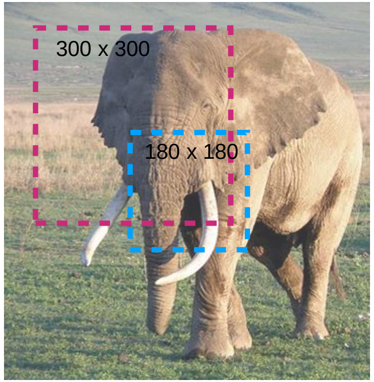
Image A





Regularization strength: from the view matching perspective

Aligning views of different
"pixel scales"



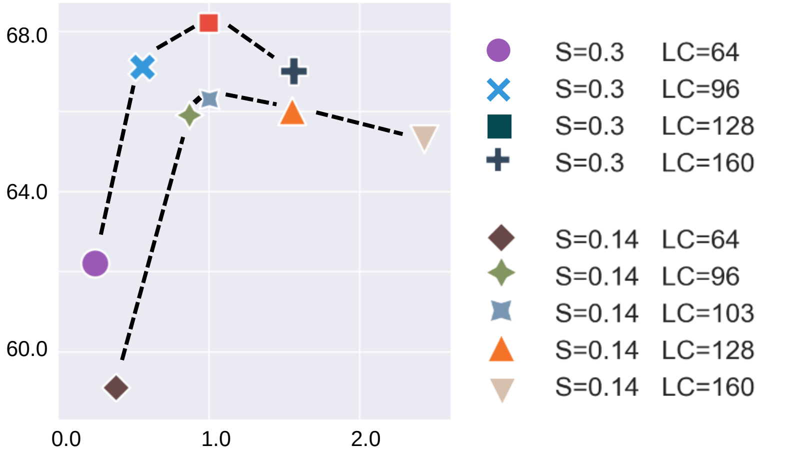
$$Pixel\ scale = \frac{target\ res.}{source\ res.}$$

224 x 224

$$\frac{224 \times 224}{300 \times 300} \approx 0.56$$

96 x 96

$$\frac{96 \times 96}{180 \times 180} \approx 0.28$$



$mean(PS_l/PS_g)$
PS_l, PS_g: pixel scale of global & local views

$S = S_g = S_l$
LC: target res. of the local views

The best performance achieves when the views have similar pixel scales (i.e. $mean(PS_l/PS_g) \approx 1.0$).

Regularization strength: from the view matching perspective

Re-balance the global and local losses

L_g: *loss between two global views*

L_l: *loss between a pair of global-local views*

P_{gg}: # of *global-global view pairs*

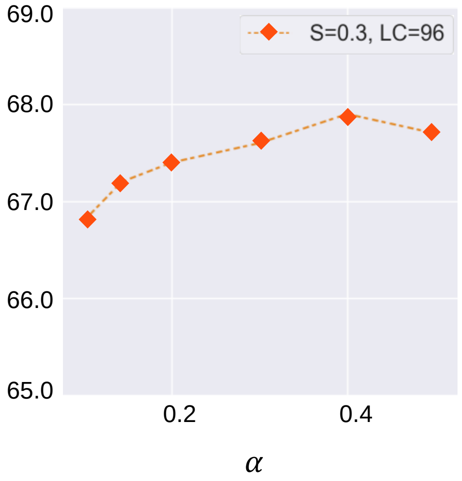
P_{gl}: # of *global-local view pairs*

Default formulation

$$L = \frac{L_g + L_l}{P_{gg} + P_{gl}}$$

Our formulation

$$L = \alpha \cdot \frac{L_g}{P_{gg}} + (1 - \alpha) \cdot \frac{L_l}{P_{gl}}$$



$S = S_g = S_l$

LC: target res. of the local views

Improve representative SSL approaches

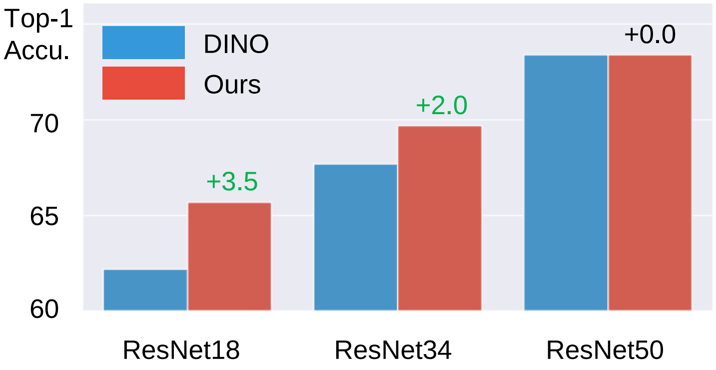
Representative SSL w. MobileNetV2 [Sandler et al.] as the backbone		Linear evaluation on ImageNet-1K	
		Top-1 (%)	Top-5 (%)
MoCo-v2 [Chen et al.] w. local views	Baseline	60.6	83.3
	Ours	61.6 (+1.0)	84.2 (+0.9)
SwAV [Caron et al.]	Baseline	65.2	85.6
	Ours	67.3 (+2.1)	87.2 (+1.6)
DINO [Caron et al.]	Baseline	66.2	86.4
	Ours	68.3 (+2.1)	87.8 (+1.4)
Supervised		71.9	90.3

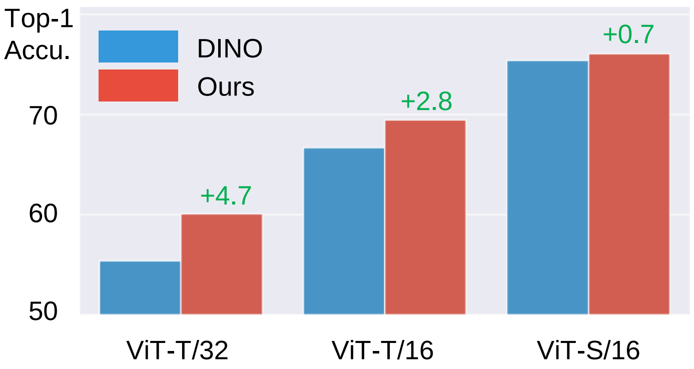
[Sandler et al.] MobileNetV2: Inverted Residuals and Linear Bottlenecks
[Chen et al.] Improved Baselines with Momentum Contrastive Learning.
[Caron et al.] Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. *NeurIPS 2020*.
[Caron et al.] Emerging Properties in Self-Supervised Vision Transformers. *ICCV 2021*.

Improve representative visual backbones

Method	Linear evaluation on ImageNet-1K						
	Top-1 (%)						
	MobileNetV2 #Par.2.2M, GFLOPS 0.31	ResNet18 #Par.11.2M, GFLOPS 1.8	ResNet34 #Par.21.3M, GFLOPS 3.7	ResNet50 #Par.23.5M, GFLOPS 4.1	ViT-T/32 #Par. 5.5M, GFLOPS 0.31	ViT-T/16 #Par. 5.5M, GFLOPS 1.26	ViT-S/16 #Par.21.7M, GFLOPS 4.6
Supervised	71.9	69.8	73.3	76.1	-	72.2	79.9
DINO baseline	66.2	62.2	67.7	73.4	55.4	66.7	75.4
DINO + Ours	68.3 (+2.1)	65.7 (+3.5)	69.7 (+2.0)	73.4 (+0.0)	60.1 (+4.7)	69.5 (+2.8)	76.1 (+0.7)

The improvement is larger for smaller networks





Improve **downstream applications**

Backbone	Method	Mask R-CNN FPN 1x on COCO		Semi-supervised Learning on ImageNet-1K	
		Object Det.	Instance Seg.	1% label	10% label
MobileNetV2	Supervised	33.1	29.8	-	-
	DINO baseline	30.9	28.1	47.9	61.3
	DINO + Ours	32.1 (+1.2)	29.1 (+1.0)	50.6 (+2.7)	63.5 (+2.2)
ResNet18	Supervised	34.5	31.6	-	-
	DINO baseline	32.7	30.6	44.5	59.2
	DINO + Ours	34.1 (+1.4)	31.8 (+1.2)	49.8 (+5.3)	63.0 (+3.8)
ResNet34	Supervised	38.7	35.0	-	-
	DINO baseline	37.6	34.6	52.4	65.4
	DINO + Ours	38.6 (+1.0)	35.5 (+0.9)	55.2 (+2.8)	67.2 (+1.8)

Compare to SOTA, all with Knowledge Distillation

Method	Linear evaluation on ImageNet-1K					
	MobileNetV2		ResNet18		ResNet34	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
Supervised	71.9	90.3	69.8	89.1	73.3	91.4
CompRess [Koohpayegani et al.]	65.8	-	62.6	-	-	-
SimReg [Navaneet et al.]	**69.1**	-	65.1	-	-	-
SEED [Fang et al.]	-	-	63.0	84.9	65.7	86.8
DisCo [Xu et al.]	-	-	65.2	86.8	67.6	88.6
BINGO [Gao et al.]	-	-	65.5	87.0	68.9	89.0
Ours	68.8	87.8	**66.8**	**87.3**	**70.8**	**90.0**

[Koohpayegani et al.] CompRess: Self-Supervised Learning by Compressing Representations. *NeurIPS 2020.*
[Navaneet et al.] SimReg: A Simple Regression Based Framework for Self-supervised Knowledge Distillation. BMVC 2021.
[Fang et al.] SEED: Self-supervised Distillation for Visual Representation. *ICLR 2021.*
[Xu et al.] Bag of Instances Aggregation Boosts Self-supervised Distillation. *ICLR 2022.*
[Gao et al.] DisCo: Remedy Self-supervised Learning on Lightweight Models with Distilled Contrastive Learning. *ECCV 2022.*

Thank you!

https://github.com/saic-fi/SSLight

ICLR
International Conference On
Learning Representations