Date: 30th Nov. 2022

Introduction to SimCLR

(…and a little more)

Presenter: Yong-Min Shin
jordan3414@yonsei.ac.kr / yongminshin.simple.ink

Chen et al., A simple framework for contrastive learning of visual representations, ICML 2020 (7000+ citations)
Cole et al., When does contrastive visual representation learning work?, CVPR 2021

00 Three main topics

* Other awesome works couldn't fit into this presentation, refer to [4], [5] and more

1 Overview of self-supervised learning (SSL) [1]

- Idea of self-supervision
- Typical approach between NLP vs. Vision

2 SimCLR (A simple framework for contrastive learning of visual representations) [2]

- Overview and augmentation viewpoint
- Recipes for good representation learning

3 Towards understanding SSL [3]

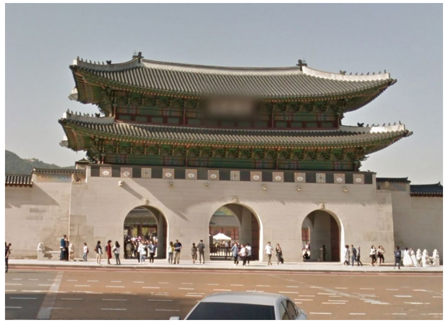
- Empirical study using SimCLR
- Analysis on 1) Dataset size 2) Dataset domain 3) Data quality 4) Task granularity

[1] LeCun, Lecture on YouTube at NYU (link: https://www.youtube.com/watch?v=tVwV14YkbYs&list=PL80I41oVxglKcAHllsU0txr3OuTTaWX2v&index=13) (2020)
[2] Chen et al., A simple framework for contrastive learning of visual representations, ICML 2020
[3] Cole et al., When does contrastive visual representation learning work?, CVPR 2021
[4] Tian et al., What makes for good views for contrastive learning?, NeurIPS 2020
[5] Wang & Liu, Understanding the behaviour of contrastive loss, CVPR 2021

01 Overview of SimCLR: Basic idea of self-supervision [1]

Self-supervised learning: Predict everything from everything else

1. ***Supervised learning**: **Learning with supervision** is extremely successful*
   - *Models adjust parameters by effective error signals*
   - *Assumption we have covered in this course: **Smoothness assumption** for semi-supervised learning*
2. ***Unsupervised learning**: **Labeling is very expensive**, unlabeled data is substantially larger*
   - *Assumption (belief, prior) of data structure is expressed in loss function*
   - *[5], [6]: Similar approach in graphs*
3. ***Self-supervised learning**: **Use the given data itself as supervision***
   - *Early ideas with Siamese nets & "metric learning": [7], [8]*
   - *First success in **natural language processing**: GPT [9], BERT [10]*
   - *Success translated to **image processing** domain: MoCo [11], SimCLR [1], BYOL [12], SimSiam [13] etc.*
   - *Biological motivation: Humans learn a large portion of the world by **observation** (even without supervision)*



...

Observe enough and we can understand
- *View angle*
- *Depth*
- *Brightness*
- *Shadow (+ direction of light)*

etc...

[1] LeCun, Lecture on YouTube at NYU (link: https://www.youtube.com/watch?v=tVwV14YkbYs&list=PL80I41oVxglKcAHllsU0txr3OuTTaWX2v&index=13) (2020)
[5] Perozzi et al., DeepWalk: Online learning of social representations, KDD 2014
[6] Hamilton et al., Inductive learning on large graphs, NeurIPS 2018
[7] Bromley, Guyon, LeCun, Sackinger and Shah, Signature verification using a "Siamese" time delay neural network, NeruIPS 1993
[8] Radford et al., Improving language understanding by generative pre-training, OpenAI blog (2018)
[10] Devlin et al., BERT: Pre-training of deep bidirectional transformers for language understanding, arXiv (2018)
[11] He et al., Momentum contrast for unsupervised visual representation learning, CVPR 2020
[12] Grill et al., Bootstrap your own latent: A new approach to self-supervised learning, NeurIPS 2020
[13] Chen et al., Exploring simple Siamese representation learning, CVPR 2021

01 Overview of SimCLR: Basic idea of self-supervision [1]

Self-supervised learning: *Predict everything from everything else*

1. Natural language processing

*"This is a **piece** of text extracted **from** a large set of **news** articles"*

Big gigantic Decoder

code

Big gigantic Encoder

*"This is a **[MASK]** of text extracted **[MASK]** a large set of **[MASK]** articles"*

2. Image processing: *Lean towards **augmentation-based** SSL*

vs.

Encoder → *h1*

Encoder → *h2*

Cost(h1, h2)

[1] LeCun, Lecture on YouTube at NYU (link: https://www.youtube.com/watch?v=tVwV14YkbYs&list=PL80I41oVxglKcAHllsU0txr3OuTTaWX2v&index=13) (2020)
Also, https://www.youtube.com/watch?v=ZaVP2SY23nc&list=PL80I41oVxglKcAHllsU0txr3OuTTaWX2v&index=14 (2020)

01 Overview of SimCLR [2]

Introduction: Unsupervised learning just as good as supervised learning

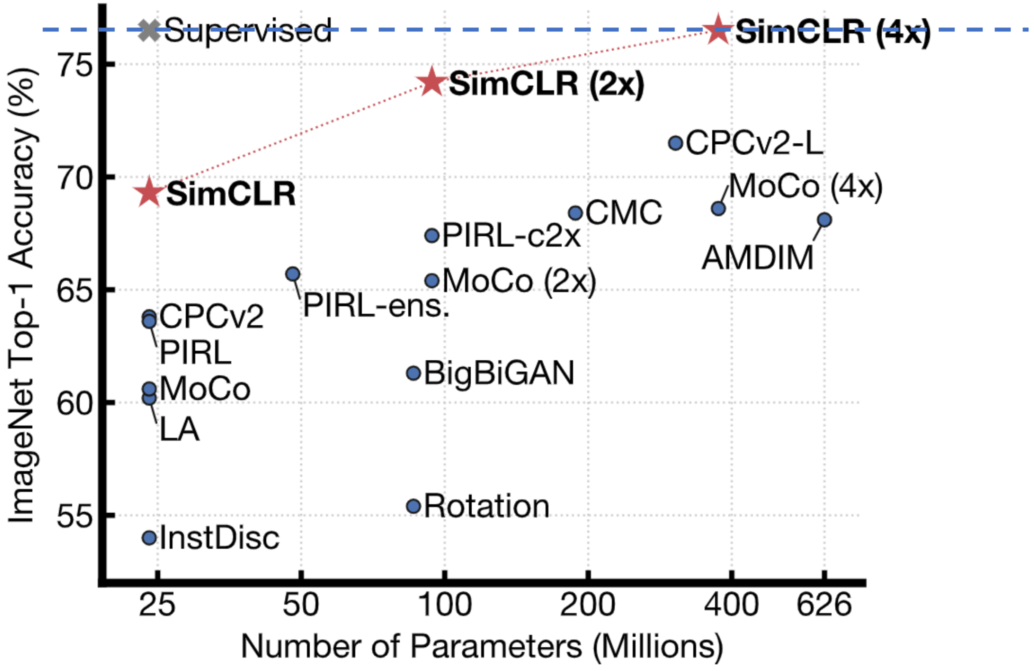


Figure 1. ImageNet Top-1 accuracy of linear classifiers trained on representations learned with different self-supervised methods (pretrained on ImageNet). Gray cross indicates supervised ResNet-50. Our method, SimCLR, is shown in bold.

Unsupervised learning ***reaches performance of supervised learning*** *for ImageNet*

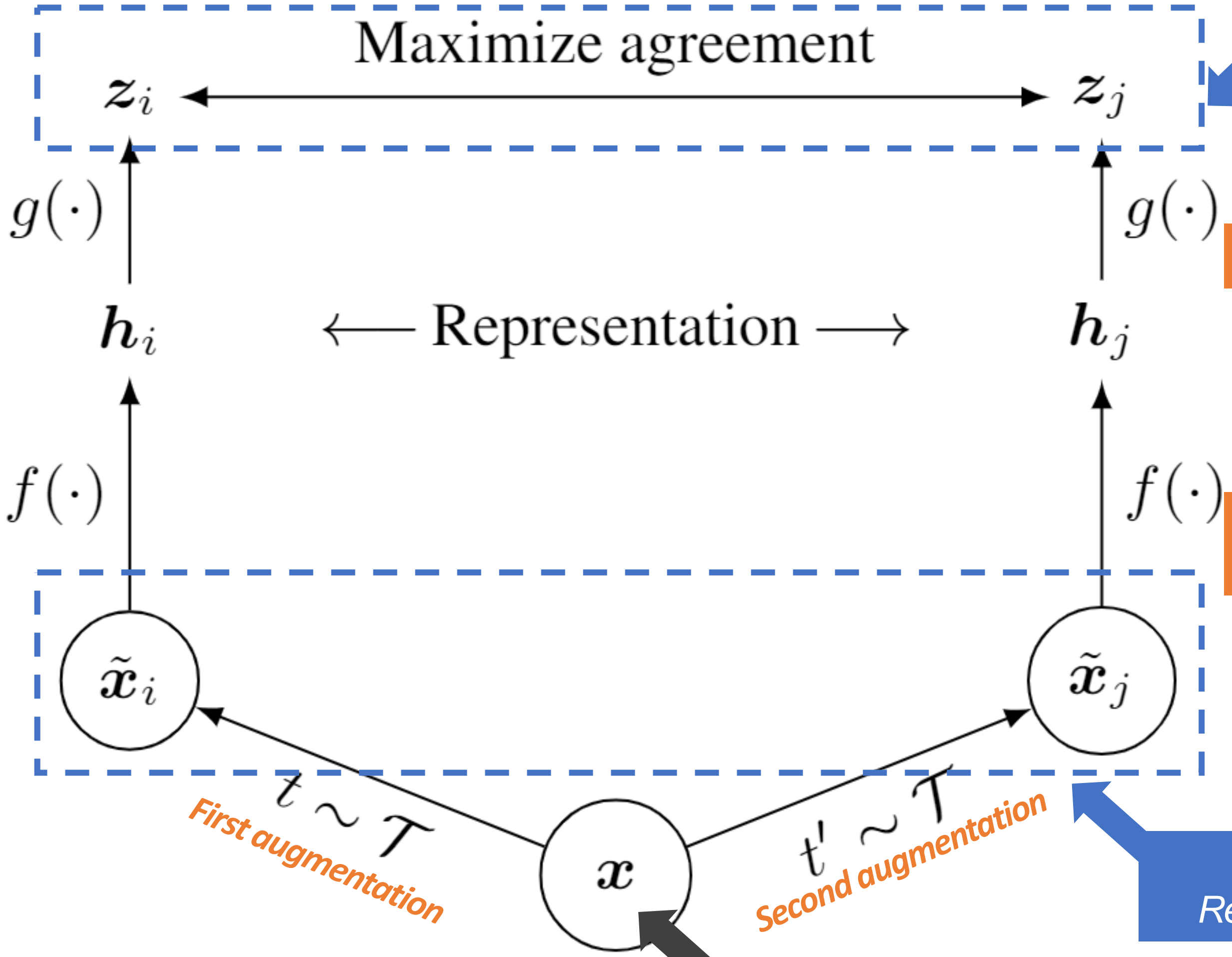
1. ***Reaching supervised learning performance***
   - *Representations from SimCLR + linear classifier* ***reaches similar performance from supervised learning***
   - *Since we user linear classifier, most benefit comes from SimCLR*

2. ***Crucial components***
   - *Composition of multiple data augmentation*
   - *Non-linear projection head*
   - *Contrastive cross entropy loss*
   - *Larger batch sizes and longer training*

[2] Chen et al., A simple framework for contrastive learning of visual representations, ICML 2020

01 Overview of SimCLR [2]

Overview of method



3. Determine whether they came from same image
[Contrastive loss function]
Cross entropy loss (+ other stuff)

[No explicit negative sampling]
For each data, consider the other 2(N - 1) samples as 'negative'

Projection head

Encoder
(ResNet backbone)

We use this encoder part at test time
(remove projection head)

2. Different data augmentations
Results in 2N (augmented) data points

1. Input data (Batch of N samples)

[2] Chen et al., A simple framework for contrastive learning of visual representations, ICML 2020

01 Overview of SimCLR [2]

A viewpoint on data augmentation [14]

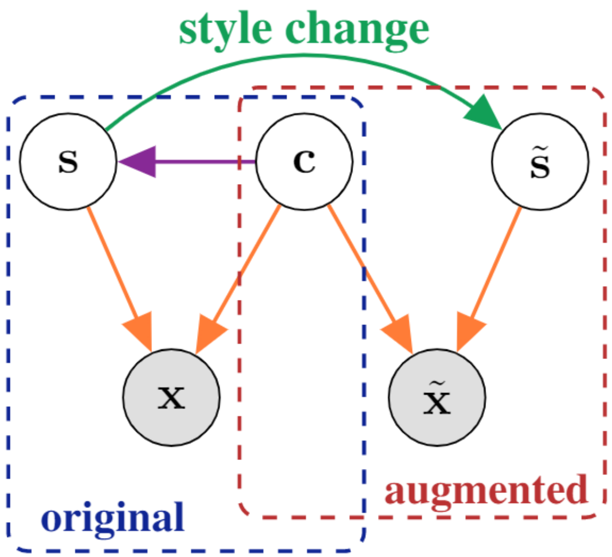


Figure 1: **Overview of our problem formulation.** We partition the latent variable $\mathbf{z}$ into content $\mathbf{c}$ and style $\mathbf{s}$, and allow for statistical and causal dependence of style on content. We assume that only style changes between the original view $\mathbf{x}$ and the augmented view $\tilde{\mathbf{x}}$, i.e., they are obtained by applying the same deterministic function $\mathbf{f}$ to $\mathbf{z} = (\mathbf{c}, \mathbf{s})$ and $\tilde{\mathbf{z}} = (\mathbf{c}, \tilde{\mathbf{s}})$.

1. *Assumption: **Style** and **content (semantic characteristics)** are related*
2. *Data that we measure is **created by a deterministic process from style & content***
3. *Then, **augmentation only changes the style** of the data and leaves the content unchanged*

[2] Chen et al., A simple framework for contrastive learning of visual representations, ICML 2020
[14] Kügelgen et al., Self-supervised learning with data augmentations provably isolates content from style, NeurIPS 2021

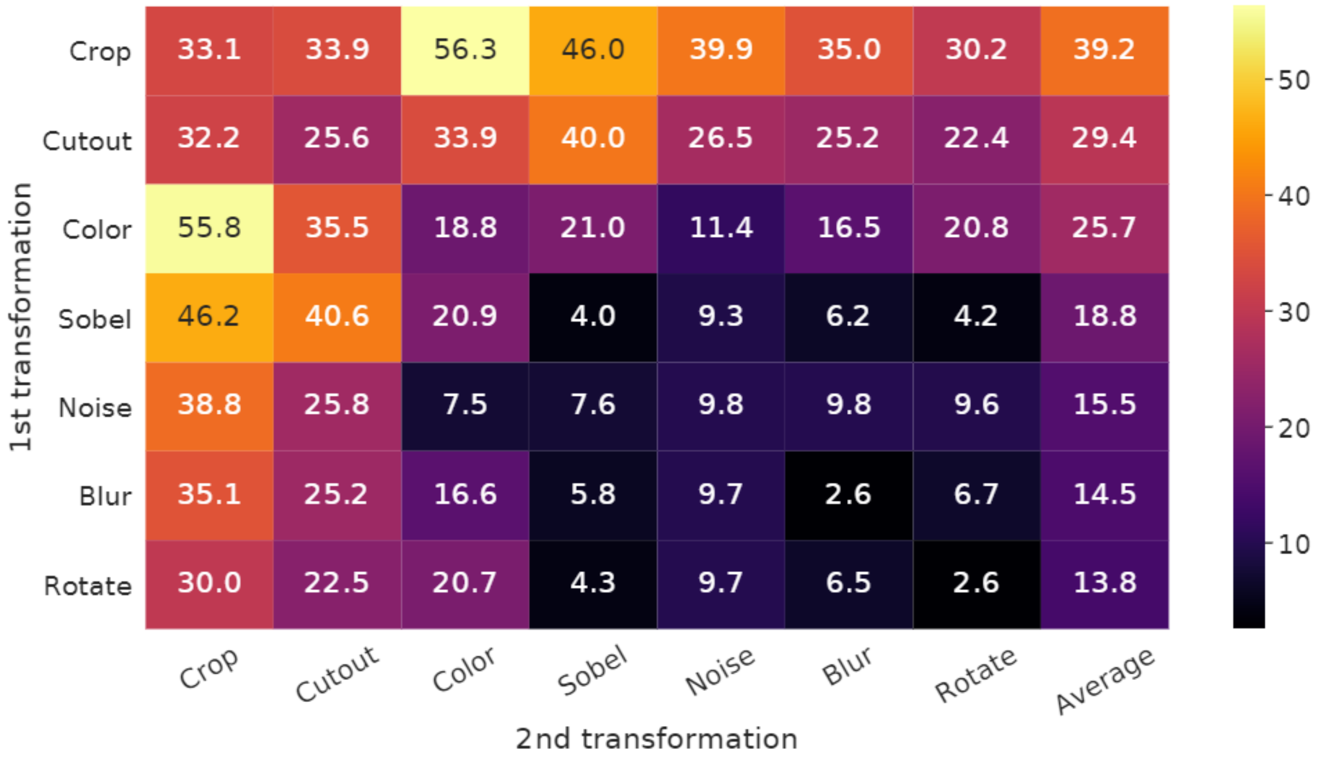
01 Overview of SimCLR: Recipes for good representations [2]

1. Composition of data augmentation is crucial for learning good representations

[Settings of augmentation ablation study]

1. *Only apply one (diagonal in Figure 5) or two (off-diagonal in Figure 5) augmentation to one of the branches*
2. *The remaining branch is always the identity*

**This is not the original setting and thus hurts the performance*



Random cropping + random color distortion stands out

[2] Chen et al., A simple framework for contrastive learning of visual representations, ICML 2020

01 Overview of SimCLR: Recipes for good representations [2]

2. CL needs stronger data augmentations than supervised learning

Stronger color distortion

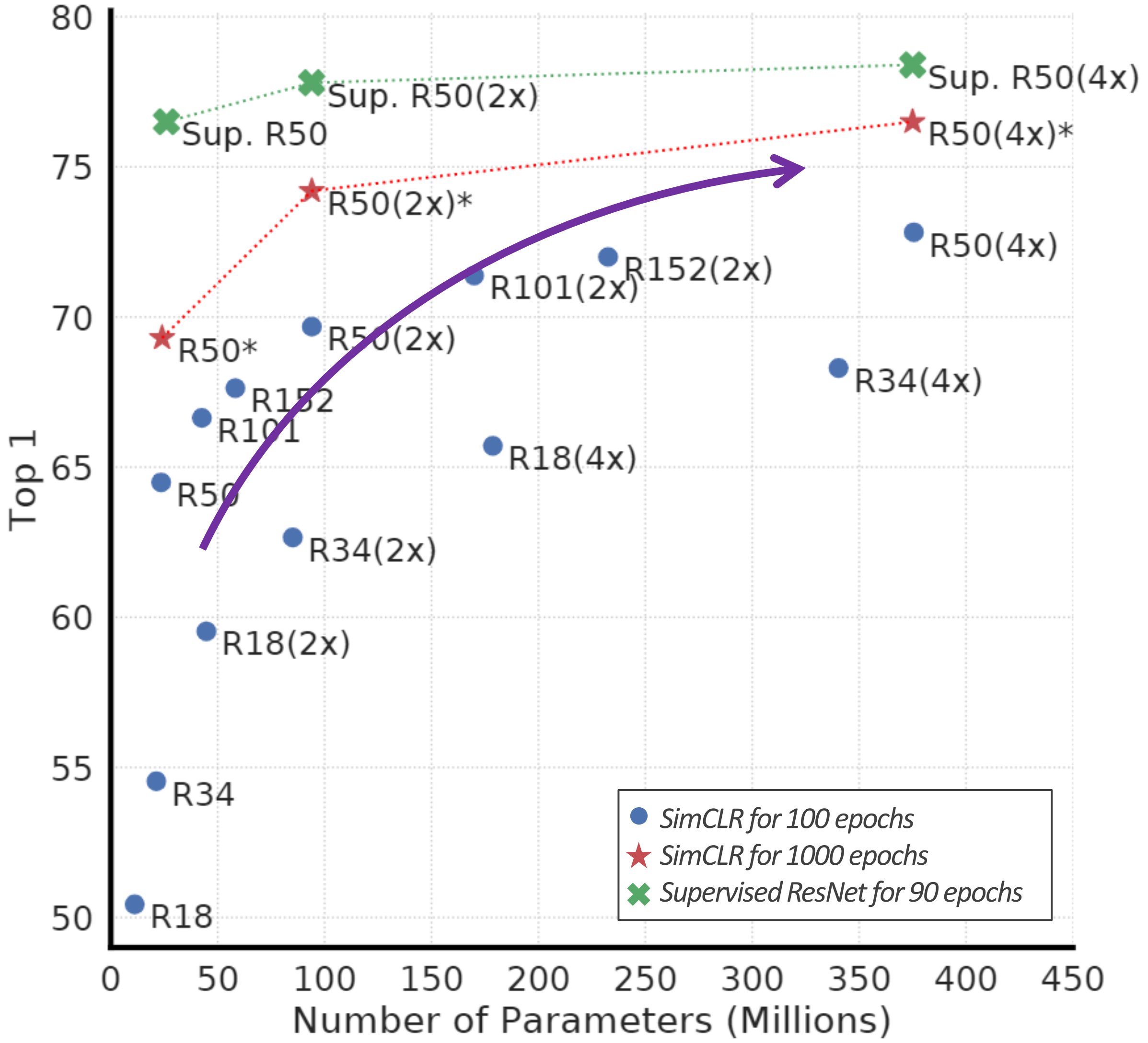
Methods	Color distortion strength 1/8	1/4	1/2	1	1 (+Blur)	AutoAug
SimCLR	59.6	61.0	62.6	63.2	64.5	61.1
Supervised	77.0	76.7	76.5	75.7	75.4	77.1

1. *Stronger color augmentation improves unsupervised learning*
2. *Supervised methods have the opposite trend*

[2] Chen et al., A simple framework for contrastive learning of visual representations, ICML 2020

01 Overview of SimCLR: Recipes for good representations [2]

3. Unsupervised CL benefits more from bigger models

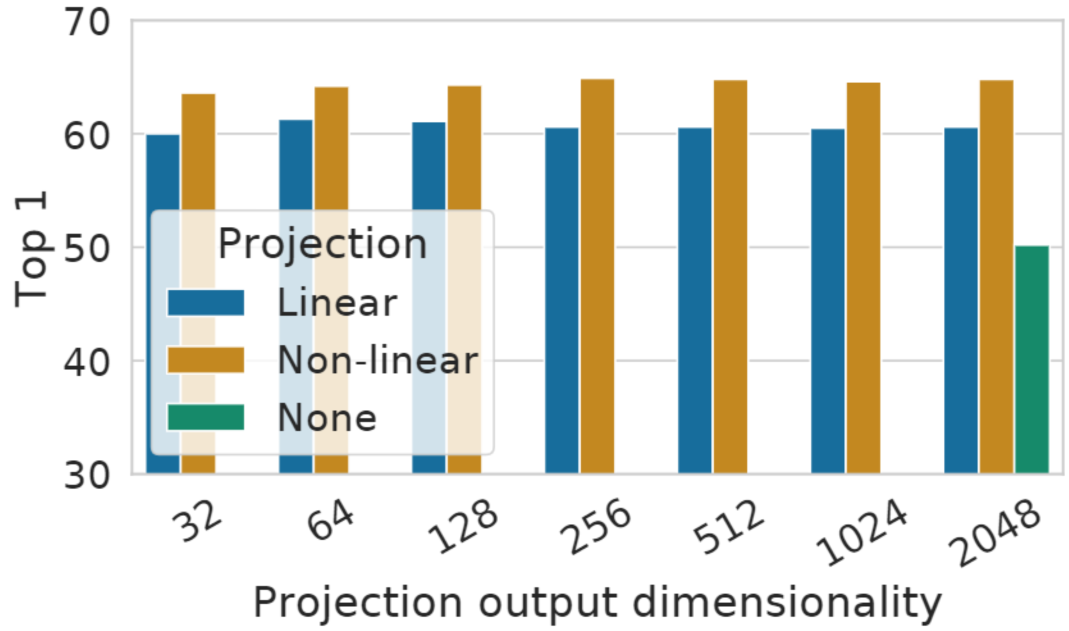


Gap between supervised and unsupervised models gets less when the model size increases

[2] Chen et al., A simple framework for contrastive learning of visual representations, ICML 2020

01 Overview of SimCLR: Recipes for good representations [2]

4. Non-linear projection head improves the representation quality of the layer before it



What to predict?	Random guess	Representation $\boldsymbol{h}$	Representation $g(\boldsymbol{h})$
Color vs grayscale	80	99.3	97.4
Rotation	25	67.6	25.6
Orig. vs corrupted	50	99.5	59.6
Orig. vs Sobel filtered	50	96.6	56.3

Loss of information

*Plot: **Non-linear projections** > **linear projections** > **None***

- *Hypothesis: Contrastive loss can lose some information critical for some downstream tasks*
- *Another experiment: Compare amount of information before & after non-linear projection*
- *Table: **A lot of information is lost after non-linear projection***

[2] Chen et al., A simple framework for contrastive learning of visual representations, ICML 2020

01 Overview of SimCLR: Recipes for good representations [2]

5. Normalized cross entropy loss with adjustable temperature works better then alternatives

Margin	NT-Logi.	Margin (sh)	NT-Logi.(sh)	NT-Xent (SimCLR)
50.9	51.6	57.5	57.9	63.9

Table 4. Linear evaluation (top-1) for models trained with different loss functions. "sh" means using semi-hard negative mining.

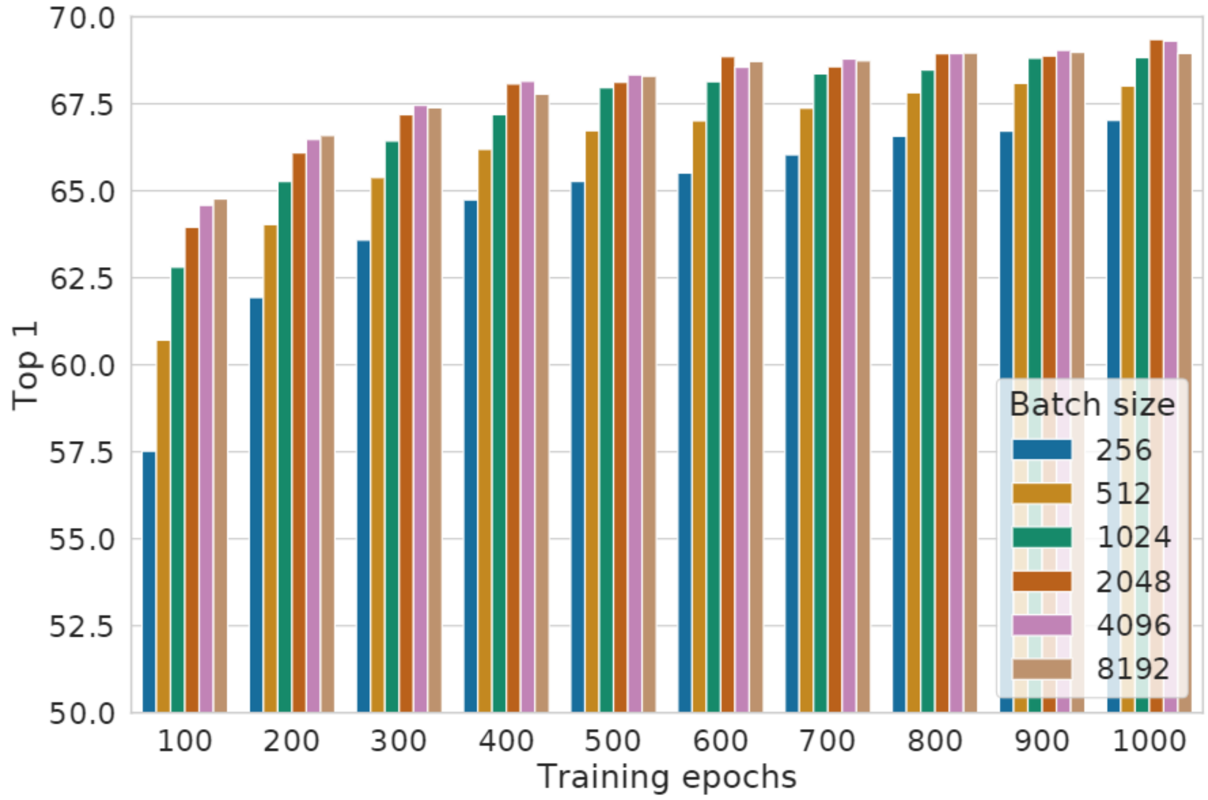
NT-Xent performs best over alternatives

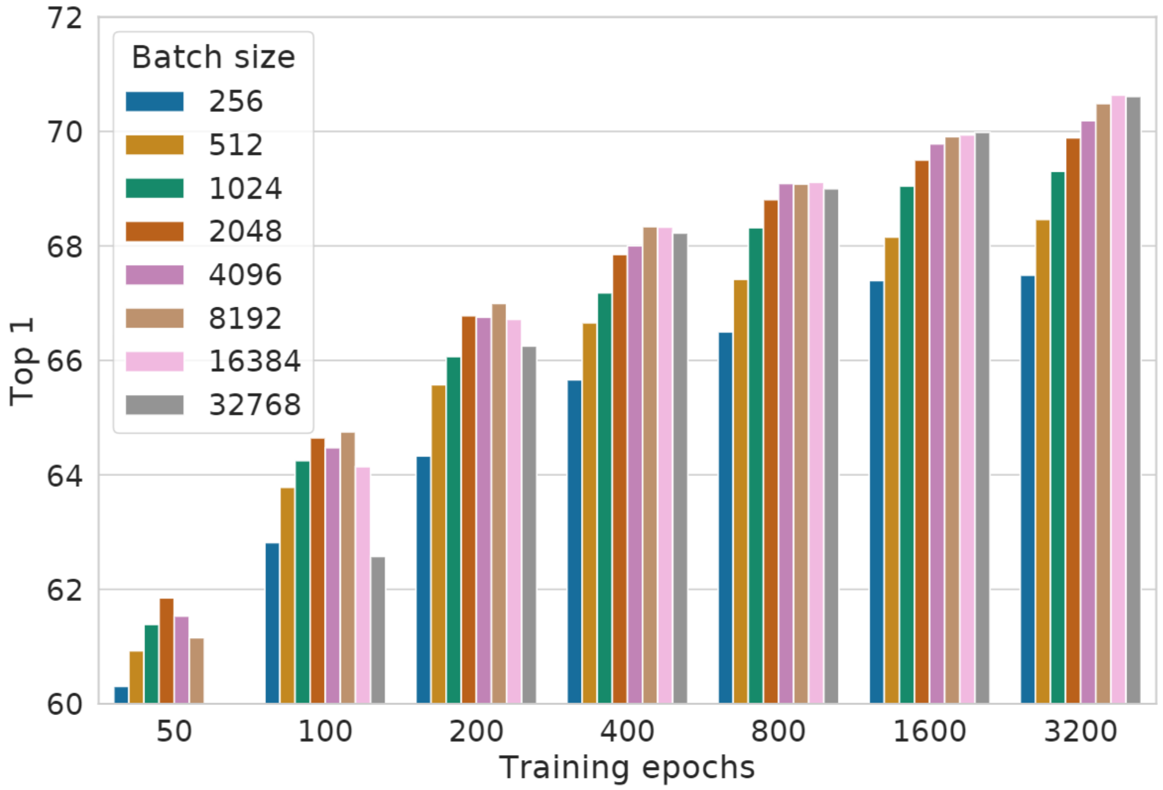
Name	Negative loss function
NT-Xent	$\boldsymbol{u}^T\boldsymbol{v}^+/\tau - \log\sum_{\boldsymbol{v}\in\{\boldsymbol{v}^+,\boldsymbol{v}^-\}}\exp(\boldsymbol{u}^T\boldsymbol{v}/\tau)$
NT-Logistic	$\log\sigma(\boldsymbol{u}^T\boldsymbol{v}^+/\tau) + \log\sigma(-\boldsymbol{u}^T\boldsymbol{v}^-/\tau)$
Margin Triplet	$-\max(\boldsymbol{u}^T\boldsymbol{v}^- - \boldsymbol{u}^T\boldsymbol{v}^+ + m, 0)$

[2] Chen et al., A simple framework for contrastive learning of visual representations, ICML 2020

01 Overview of SimCLR: Recipes for good representations [2]

6. CL benefits more from *larger batch sizes* and *longer training*

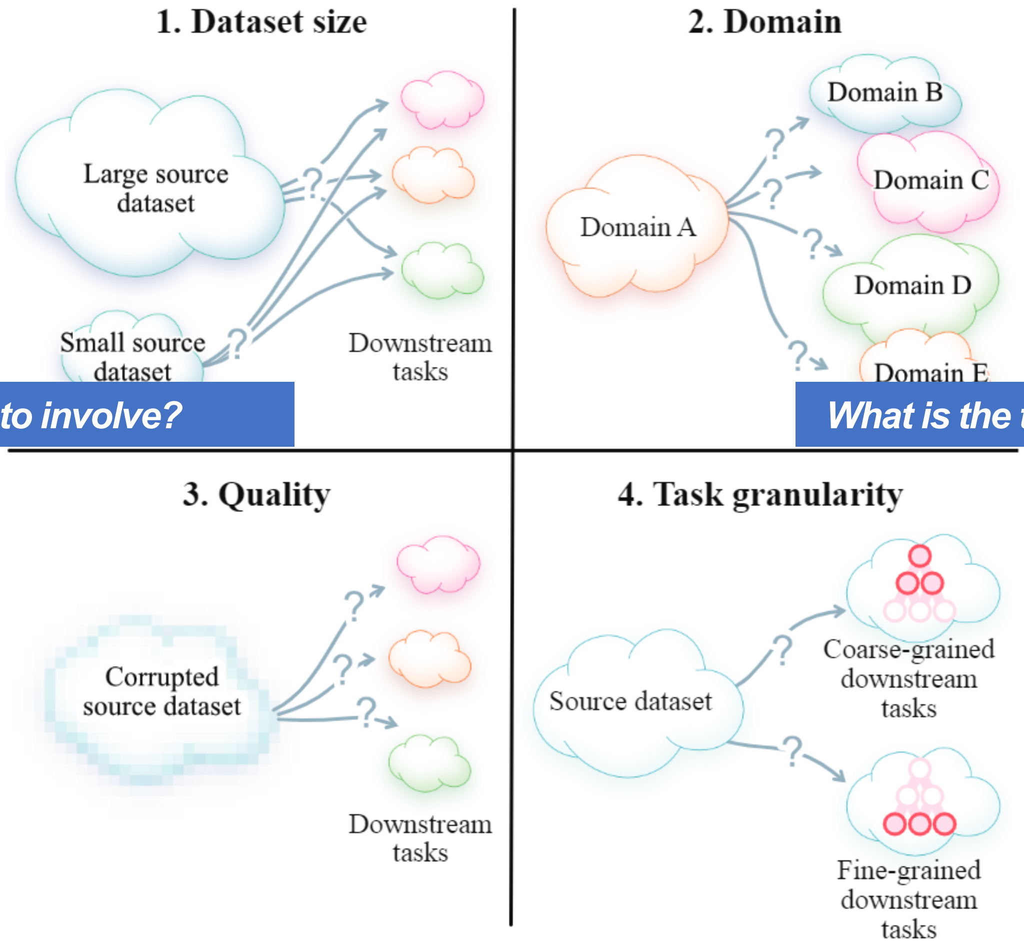




[2] Chen et al., A simple framework for contrastive learning of visual representations, ICML 2020

03 When does it work?: Focus on empirical analysis for visual representations

An empirical analysis of SSL using SimCLR



How much data do we need to involve?

What is the transferability between different data?

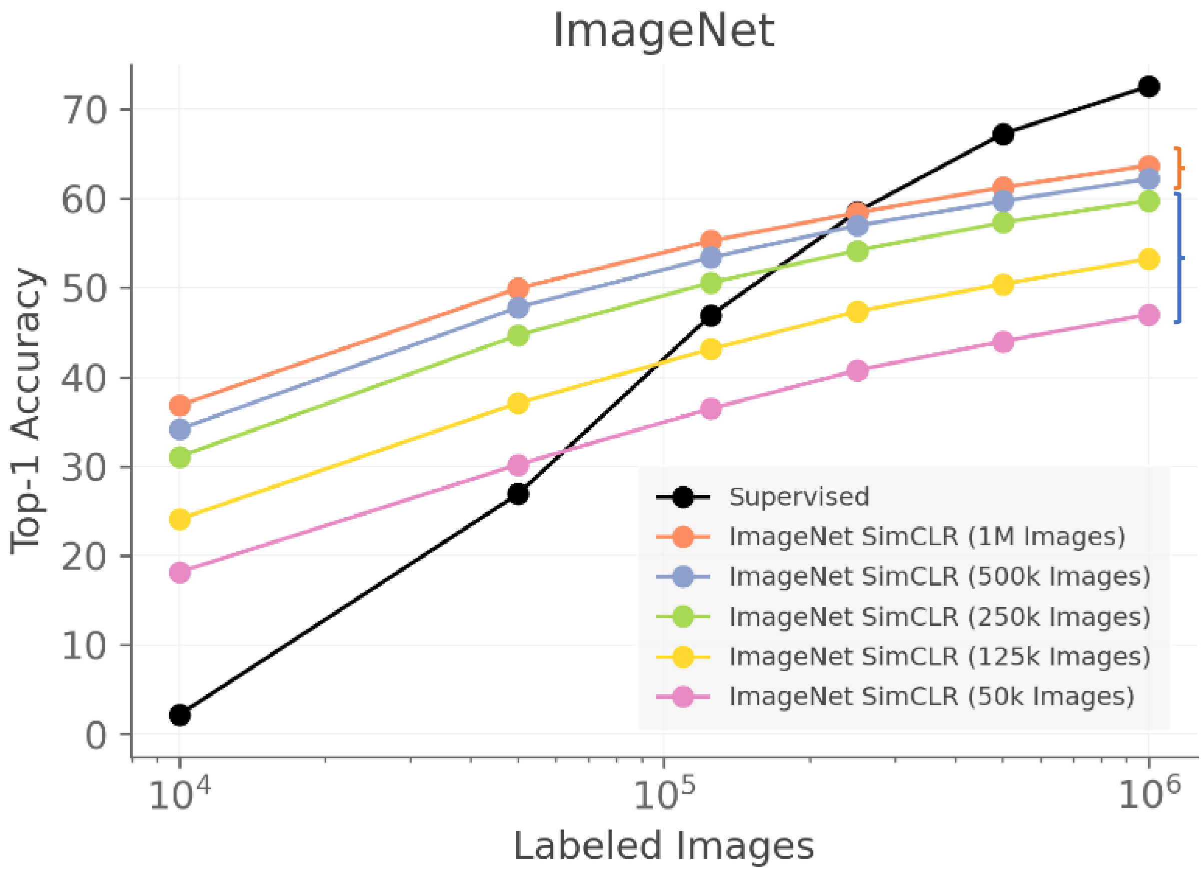
How much is SSL robust to image corruption?

Can SSL help for more difficult tasks?

[3] Cole et al., When does contrastive visual representation learning work?, CVPR 2021

03 When does it work?: Focus on empirical analysis for visual representations

1. Dataset size: There is *little benefit beyond 500k*



(a) Linear Evaluation

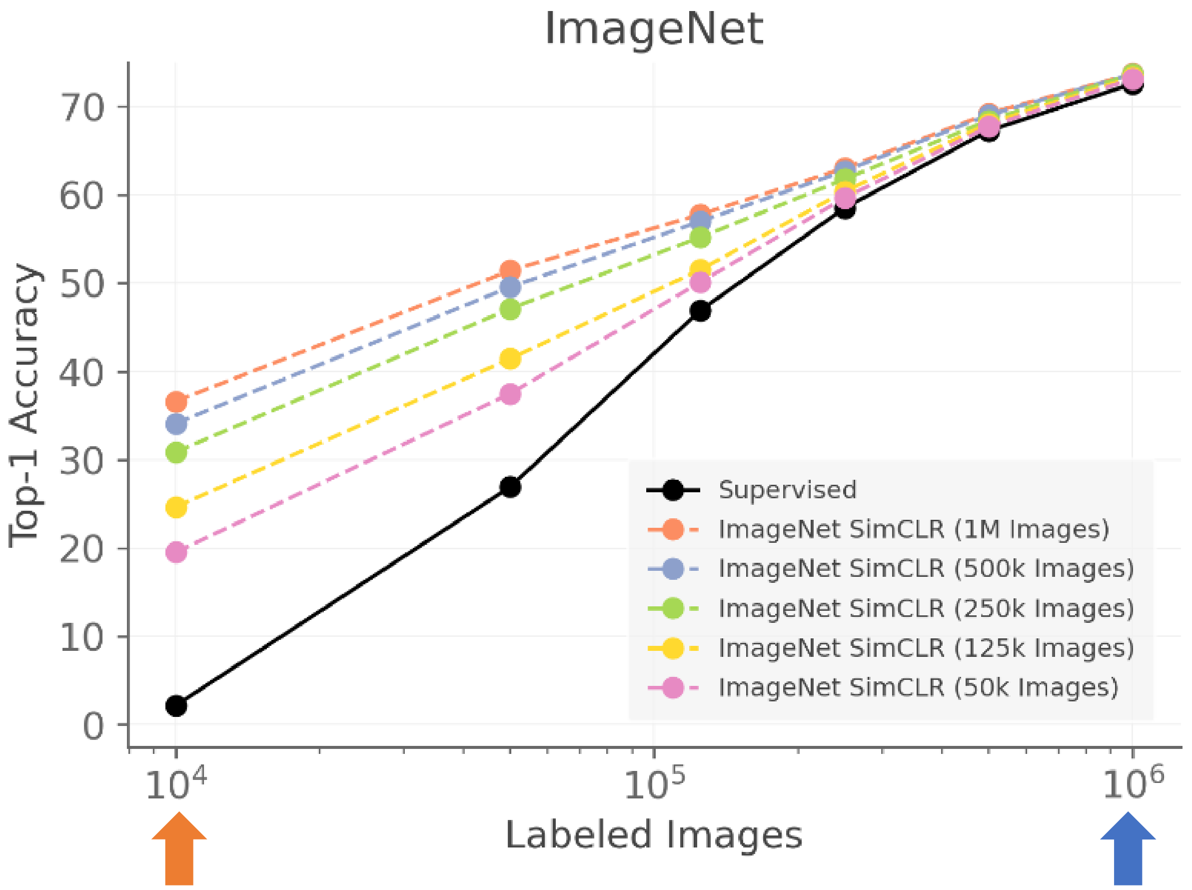
Diminishing returns after 500k

Significant performance increase from more images

[3] Cole et al., When does contrastive visual representation learning work?, CVPR 2021

03 When does it work?: Focus on empirical analysis for visual representations

1. Dataset size: SSL provides a good model initialization



Big gains from SSL pre-training in scarce supervision settings

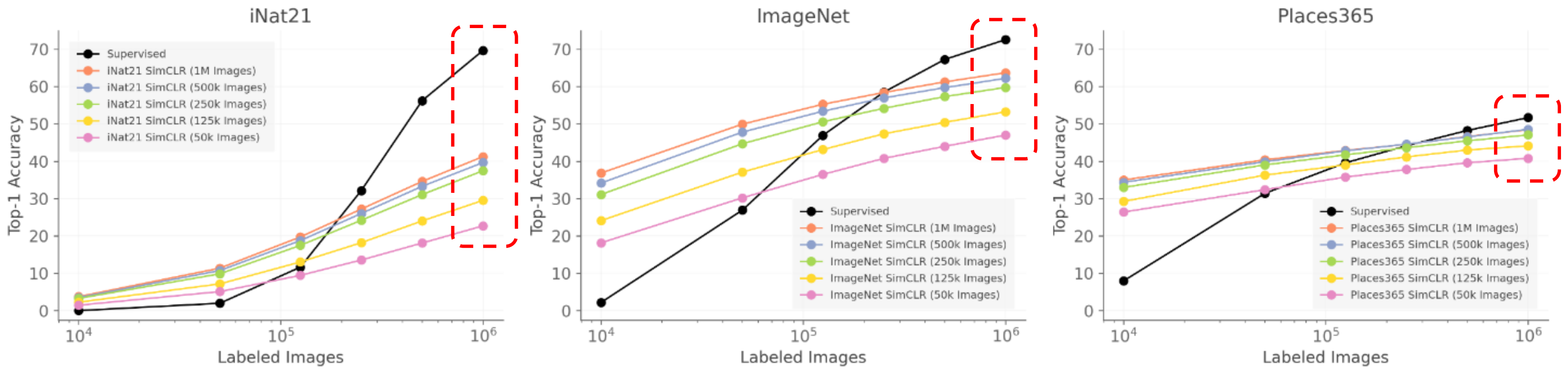
(b) Fine-Tuning

Gains from pre-training decreases as supervision becomes abundant

[3] Cole et al., When does contrastive visual representation learning work?, CVPR 2021

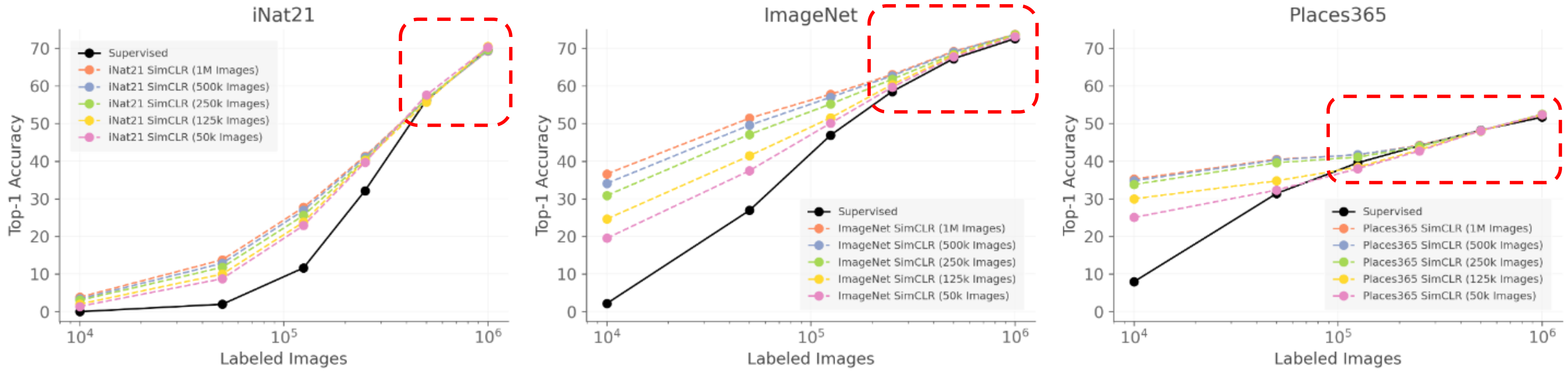
03 When does it work?: Focus on empirical analysis for visual representations

1. Dataset size: SSL needs a lot of labeled images to match supervised performance



(a) Linear Evaluation

[Linear evaluation]
Starts to match the performance ***near ~1M labeled images***
+ iNat21 is a challenging dataset



(b) Fine-Tuning

[Fine-tuning]
Starts to match the performance ***near 100~500k labeled images***

[3] Cole et al., When does contrastive visual representation learning work?, CVPR 2021

03 When does it work?: Focus on empirical analysis for visual representations

2. Domain: Pre-training from the *same* domain is *always better*

*Linear evaluation

Pretraining	iNat21	ImageNet	Places365	GLC20
iNat21 (1M) SimCLR	**0.493**	0.519	0.416	0.707
ImageNet (1M) SimCLR	0.373	**0.644**	0.486	0.716
Places365 (1M) SimCLR	0.292	0.491	**0.501**	0.693
GLC20 (1M) SimCLR	0.187	0.372	0.329	**0.769**

ImageNet is the best when transferring between datasets

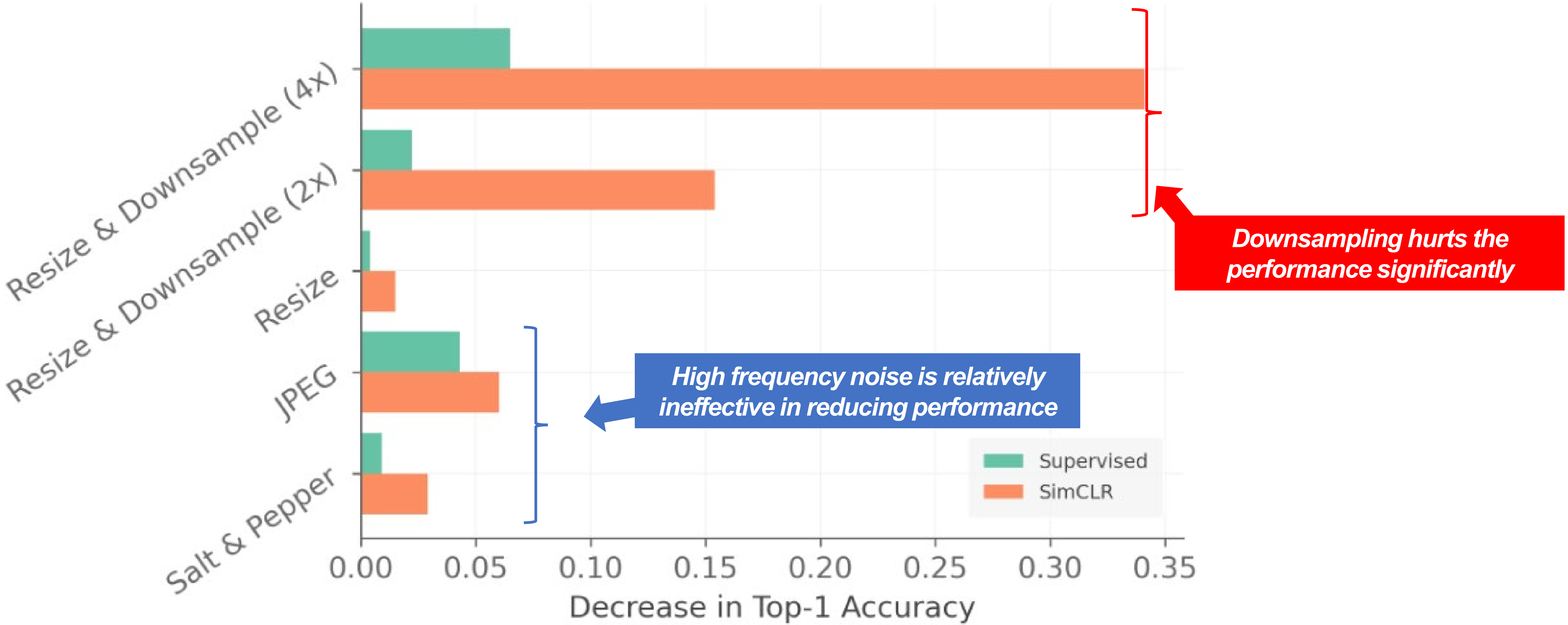
Pre-training with the same domain is dominantly better

*Also, adding & combining different datasets usually **does not benefit** the performance*

[3] Cole et al., When does contrastive visual representation learning work?, CVPR 2021

03 When does it work?: Focus on empirical analysis for visual representations

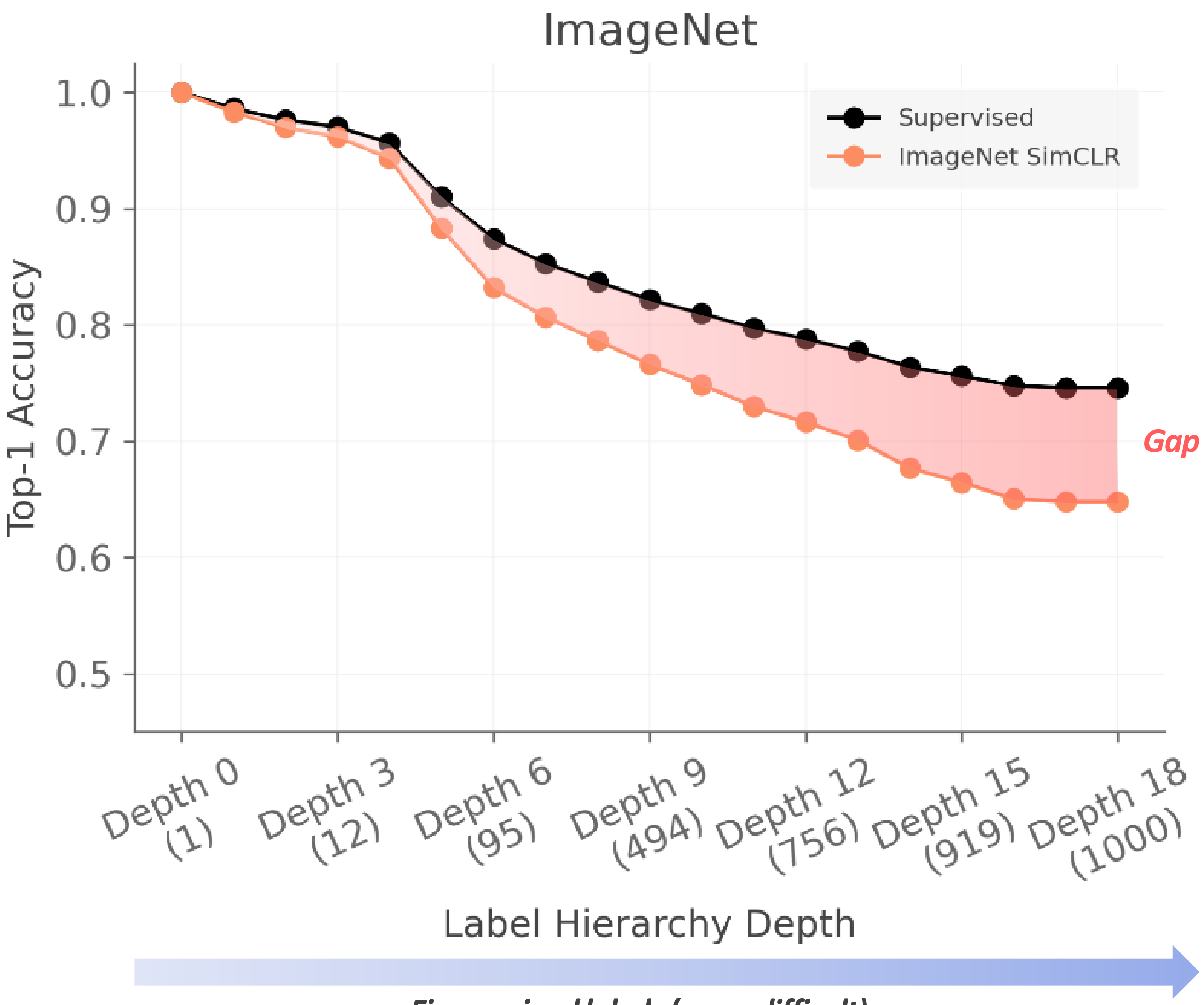
3. Quality: SimCLR is critical in image resolution, and robust in noise



[3] Cole et al., When does contrastive visual representation learning work?, CVPR 2021

03 When does it work?: Focus on empirical analysis for visual representations

4. Task granularity: SimCLR is *critical* in *image resolution*, and *robust* in *noise*



Gap between supervised vs. SimCLR increases

Fine-grained labels (more difficult)

[3] Cole et al., When does contrastive visual representation learning work?, CVPR 2021

04 Summary

SimCLR: One of the most impactful works in vision (2020)

1. *How to perform good? [2]*
   - *Diverse & strong augmentations*
   - *Large models, large batches, longer training*
   - *Non-linear projection*
   - *NX-Tent loss function*
2. *Broader analysis [3]*
   - *Dataset size has diminishing returns*
   - *SSL provides good initialization*
   - *Still need lot of labeled data*
   - *Keep the dataset domain consistent*
   - *Use high resolution images*
   - *May not be powerful in datasets with subtler class differences*

[2] Chen et al., A simple framework for contrastive learning of visual representations, ICML 2020
[3] Cole et al., When does contrastive visual representation learning work?, CVPR 2021