A Simple Framework for Contrastive Learning of Visual Representations

Ting Chen, Simon Kornblith, Mohammad Norouzi, Geoffrey Hinton

Google Research, Brain Team

Contribution

- Propose a technique for simple contrastive learning (SimCLR).
  - Framework is easily generalizable; does not require specific architecture.

- Demonstrate promising image classification performance with the use of SimCLR

Motivation

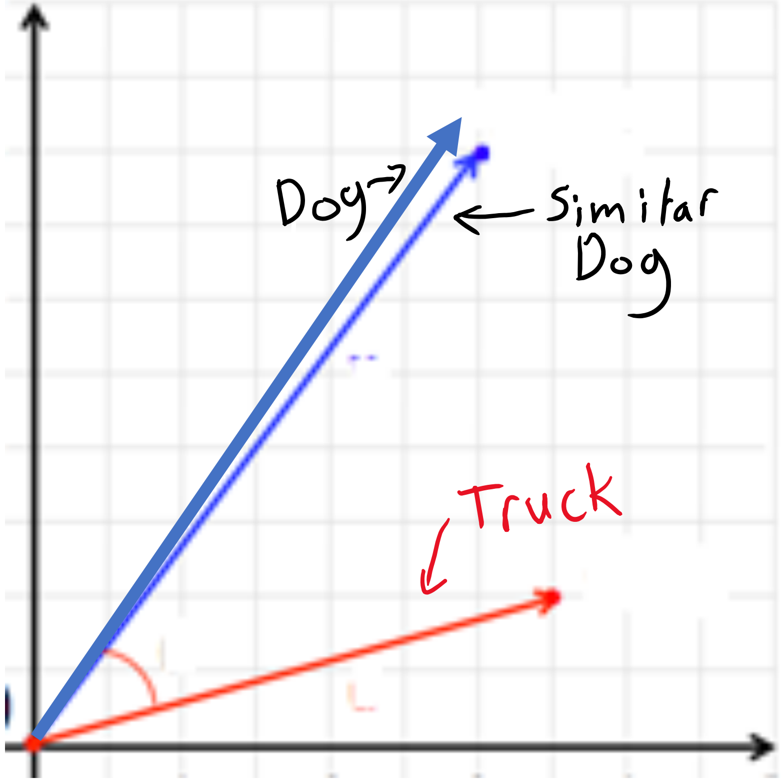
- Deep Neural Networks are commonly used for image-based tasks (e.g. classification).

- Generating adequate labeled data for training is often impractical, especially for highly specific tasks.

- Better performance from unsupervised training methods is desired.

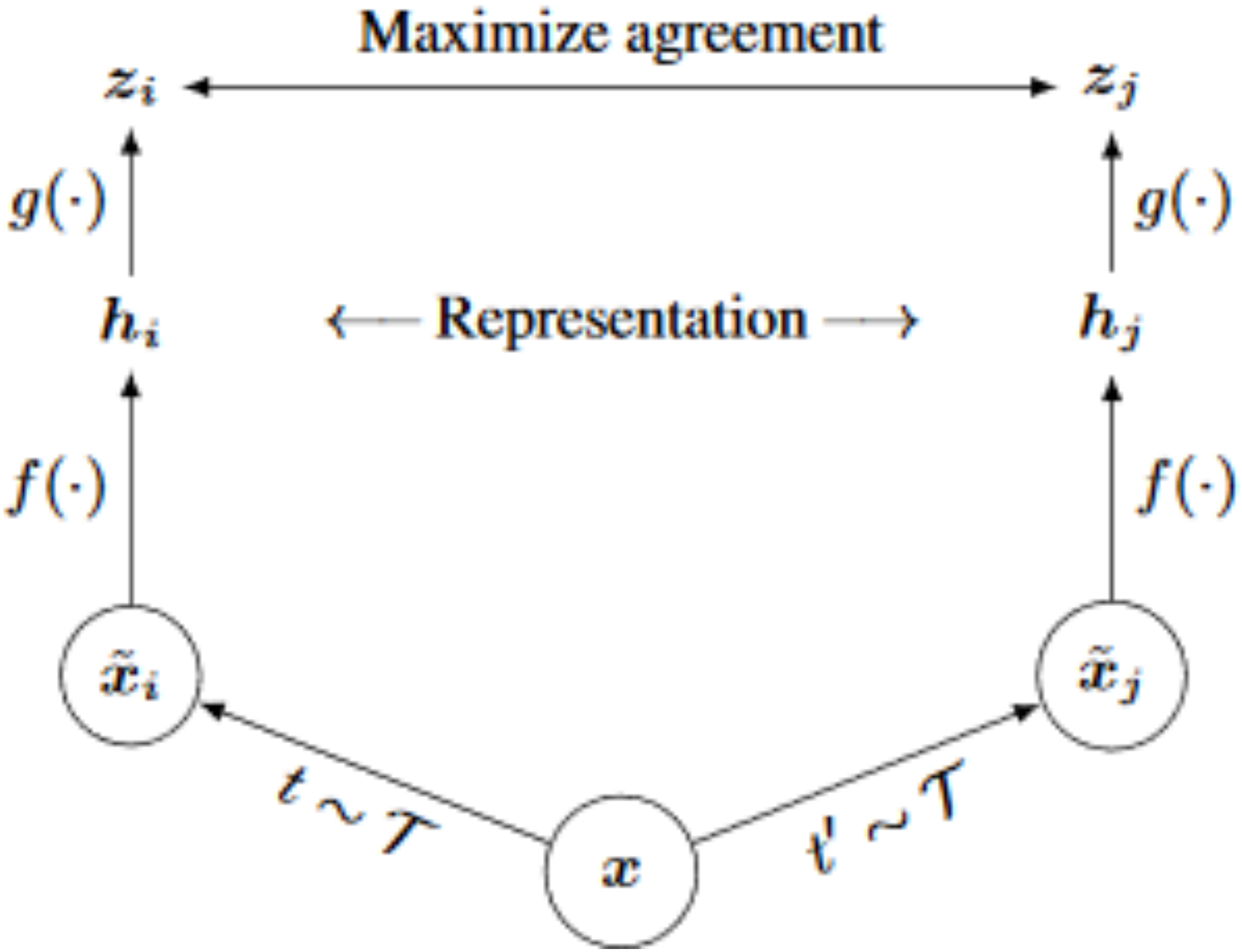
What is Contrastive Learning?

- Essentially a pretext task for teaching useful representations by *contrasting* different images.
- Network is incentivized to find feature representation where latent vector is *similar* when images are contextually similar.
  - Current contrastive learning methods sometimes require specific network architectures



Idea behind SimCLR

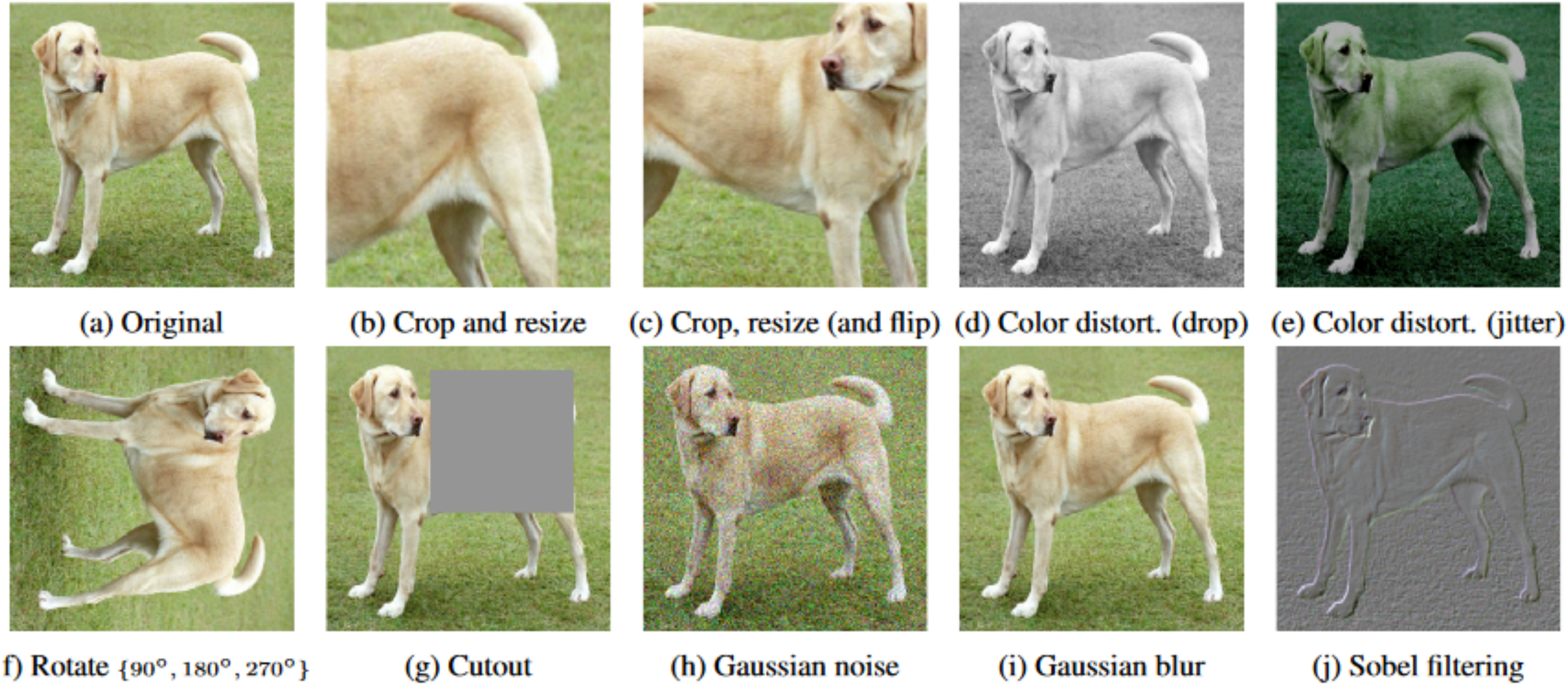
- Use data augmentation to create similar images on unlabeled data
- Pass paired images through encoder *f*
- Change Encoding into feature vector with projection head *g*
- Train network to maximize:
  - Agreement between 'positive examples'
  - Disagreement between 'negatives' (mismatches)



Experimental Summary

- Contrastive Learning was evaluated by training the SimCLR, then attaching it to a linear classifier (small amount of supervised learning) and testing for accuracy.
- Data augmentation used was ultimately random crop&resize with color distortions and Gaussian Blur.
- Base encoder architecture: ResNet-50
- Projection Head: 2-layer MLP
- Loss function: NT-Xent optimized with LARS.

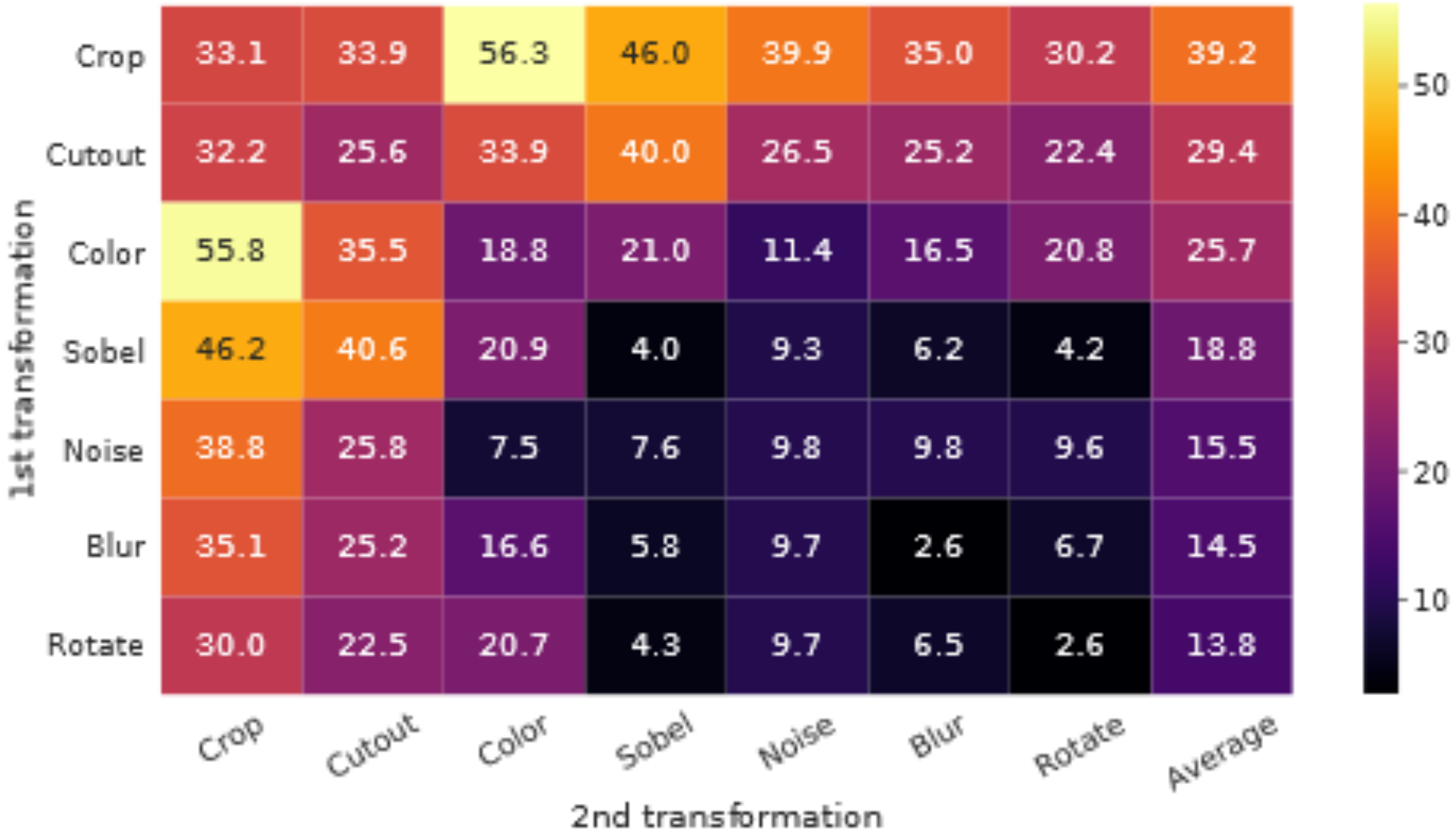
Tested Data Augmentations



(a) Original (b) Crop and resize (c) Crop, resize (and flip) (d) Color distort. (drop) (e) Color distort. (jitter)

f) Rotate {90°, 180°, 270°} (g) Cutout (h) Gaussian noise (i) Gaussian blur (j) Sobel filtering

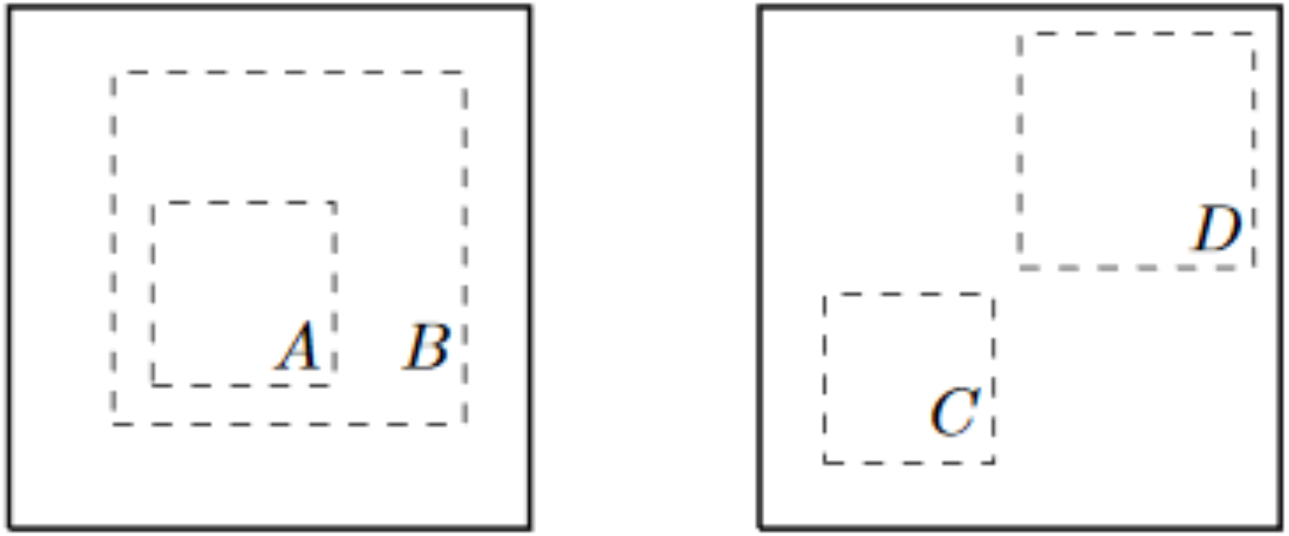
Performance Results of Different Transforms



1st transformation \ 2nd transformation	Crop	Cutout	Color	Sobel	Noise	Blur	Rotate	Average
Crop	33.1	33.9	56.3	46.0	39.9	35.0	30.2	39.2
Cutout	32.2	25.6	33.9	40.0	26.5	25.2	22.4	29.4
Color	55.8	35.5	18.8	21.0	11.4	16.5	20.8	25.7
Sobel	46.2	40.6	20.9	4.0	9.3	6.2	4.2	18.8
Noise	38.8	25.8	7.5	7.6	9.8	9.8	9.6	15.5
Blur	35.1	25.2	16.6	5.8	9.7	2.6	6.7	14.5
Rotate	30.0	22.5	20.7	4.3	9.7	6.5	2.6	13.8

Figure 5. Linear evaluation (ImageNet top-1 accuracy) under individual or composition of data augmentations, applied only to one branch. For all columns but the last, diagonal entries correspond to single transformation, and off-diagonals correspond to composition of two transformations (applied sequentially). The last column reflects the average over the row.

Conclusion: Crop+Color is Superior Augmentation

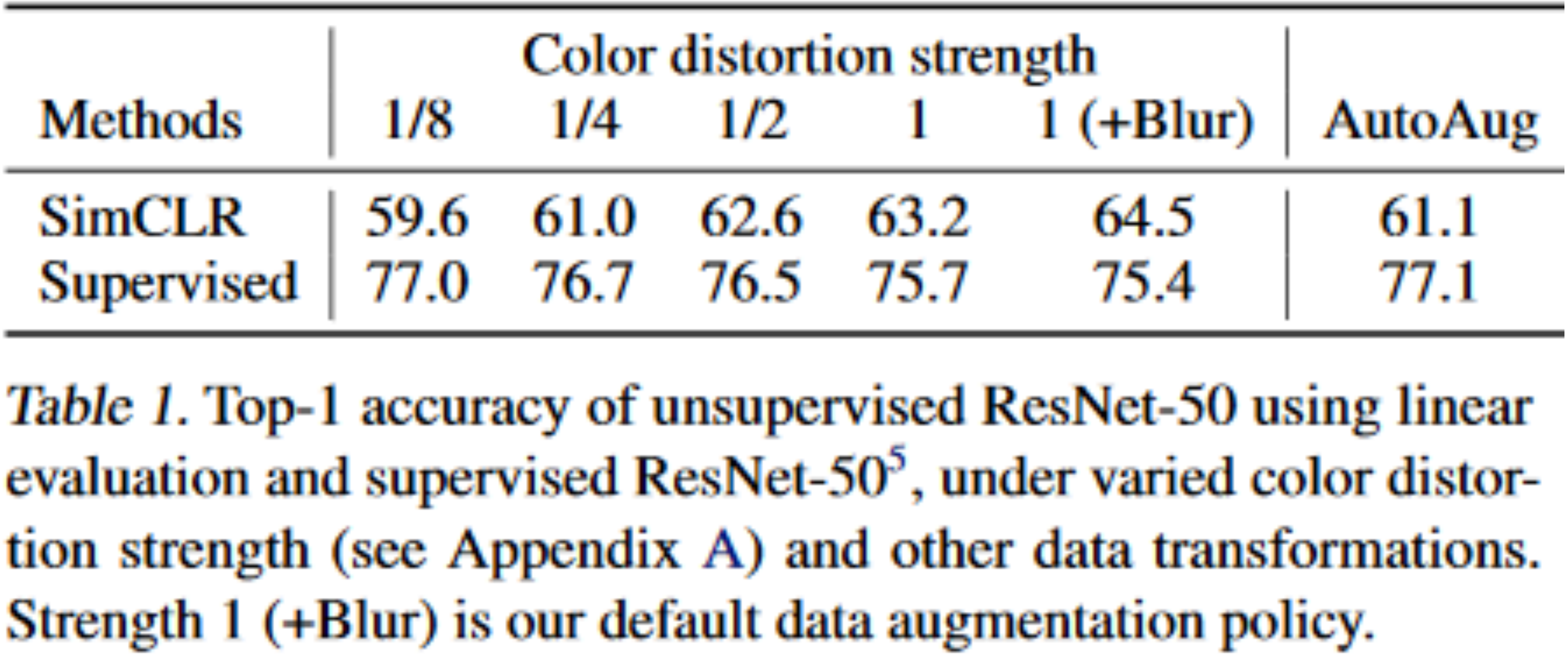


(a) Global and local views.

(b) Adjacent views.

Figure 3. Solid rectangles are images, dashed rectangles are random crops. By randomly cropping images, we sample contrastive prediction tasks that include global to local view (B→A) or adjacent view (D→C) prediction

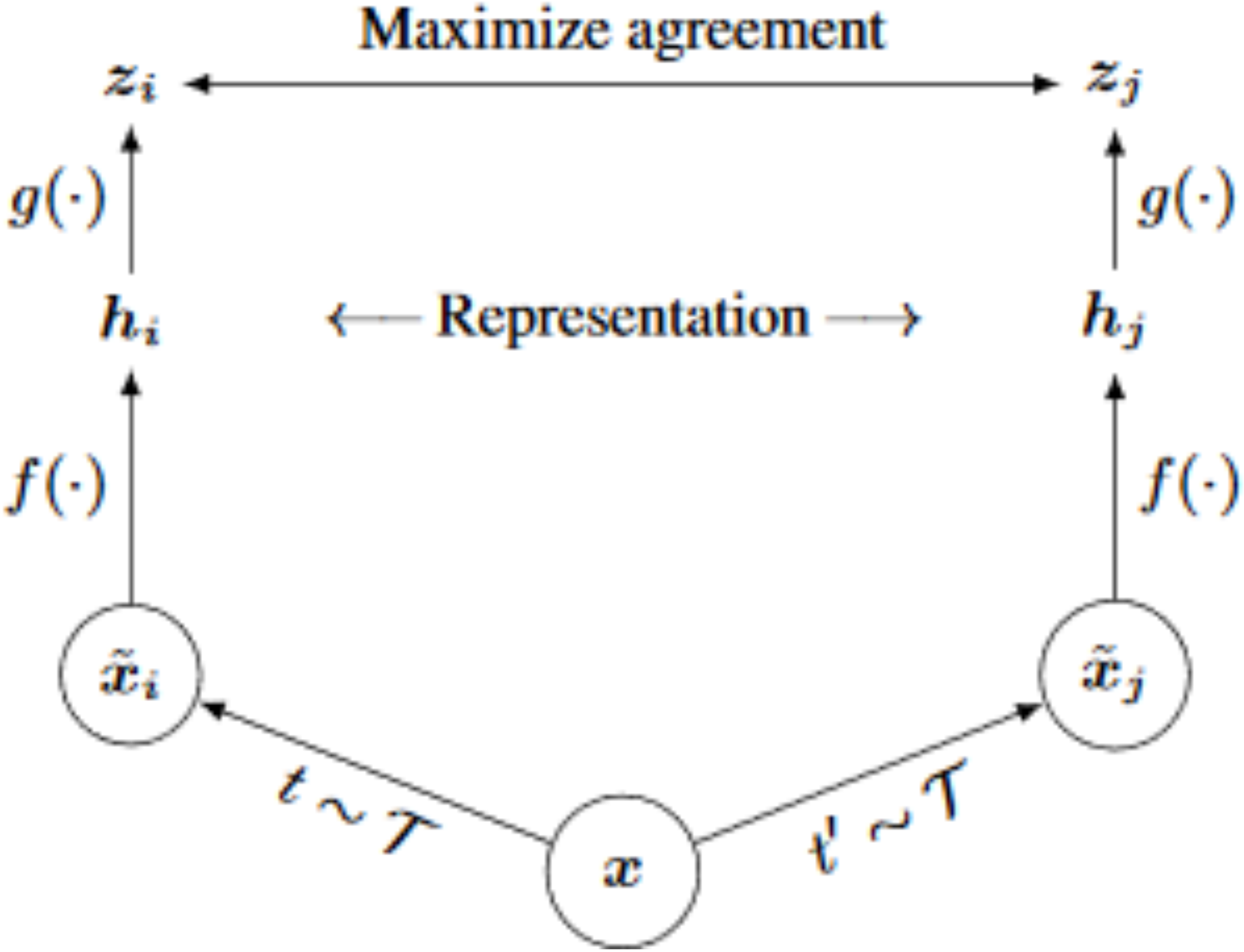
Contrastive Learning calls for *Stronger* Data Augmentation



	Color distortion strength					
Methods	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

Table 1. Top-1 accuracy of unsupervised ResNet-50 using linear evaluation and supervised ResNet-50[5], under varied color distortion strength (see Appendix A) and other data transformations. Strength 1 (+Blur) is our default data augmentation policy.

Encoder and Projection Head



Larger Encoder Networks = Better Performance

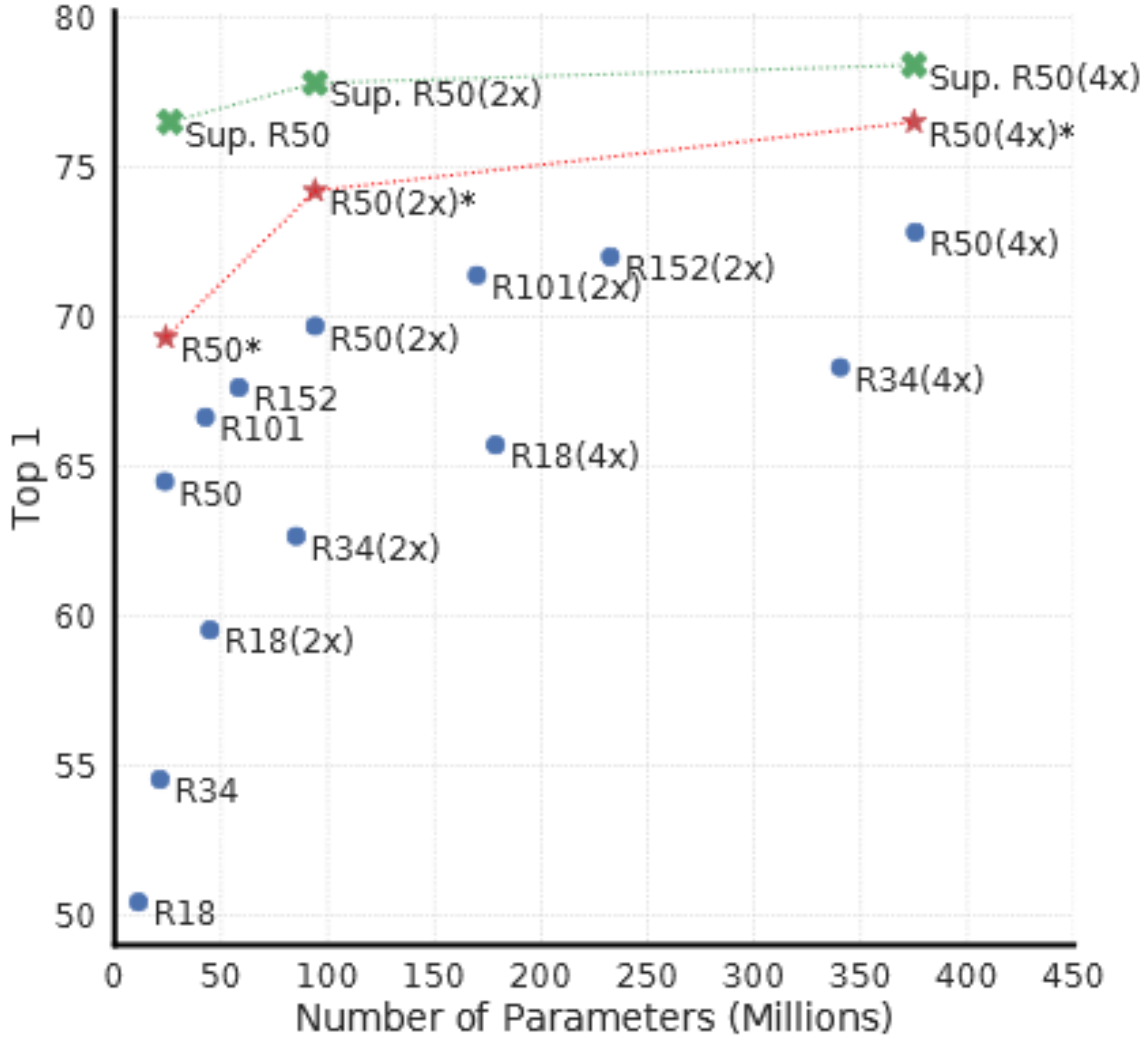
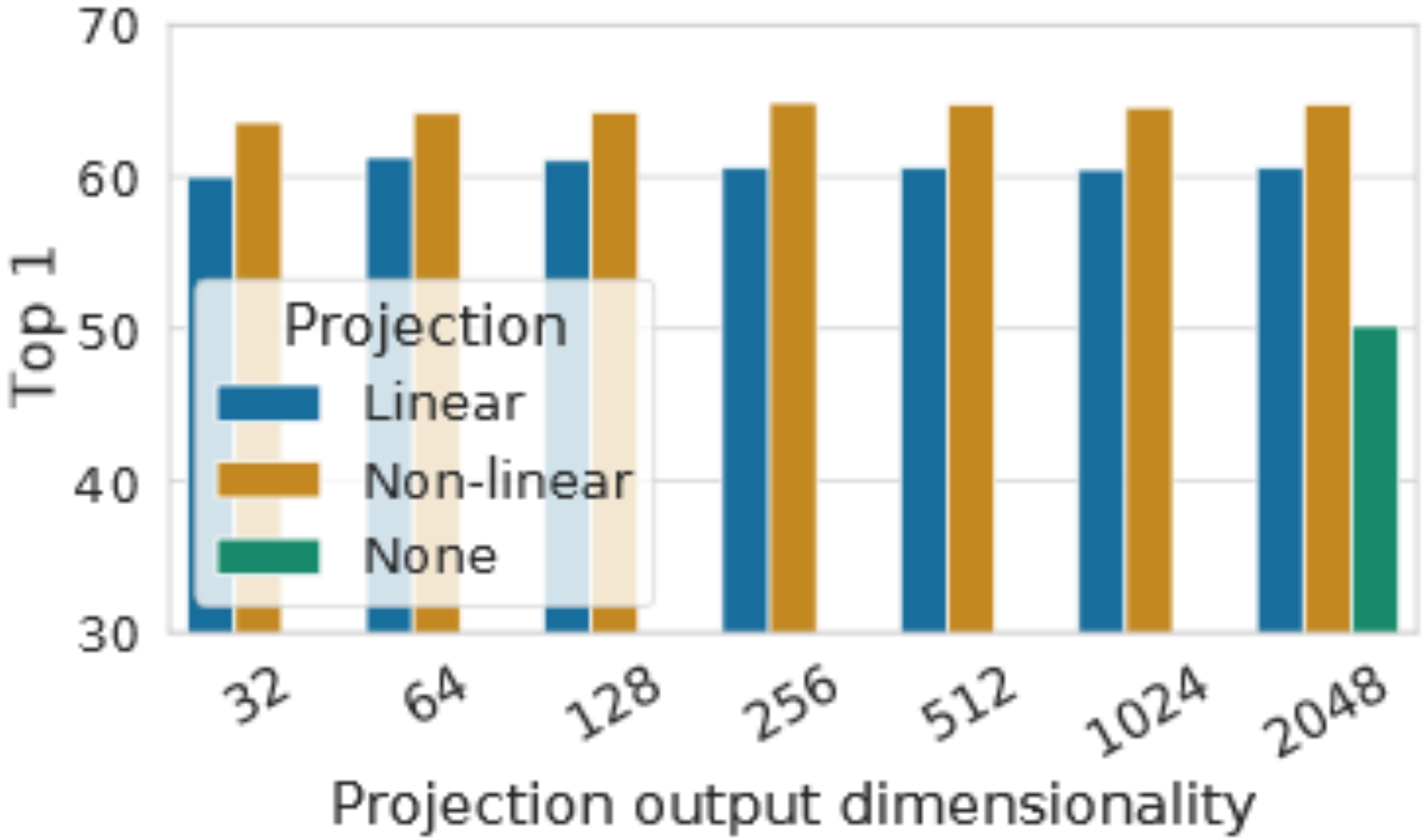


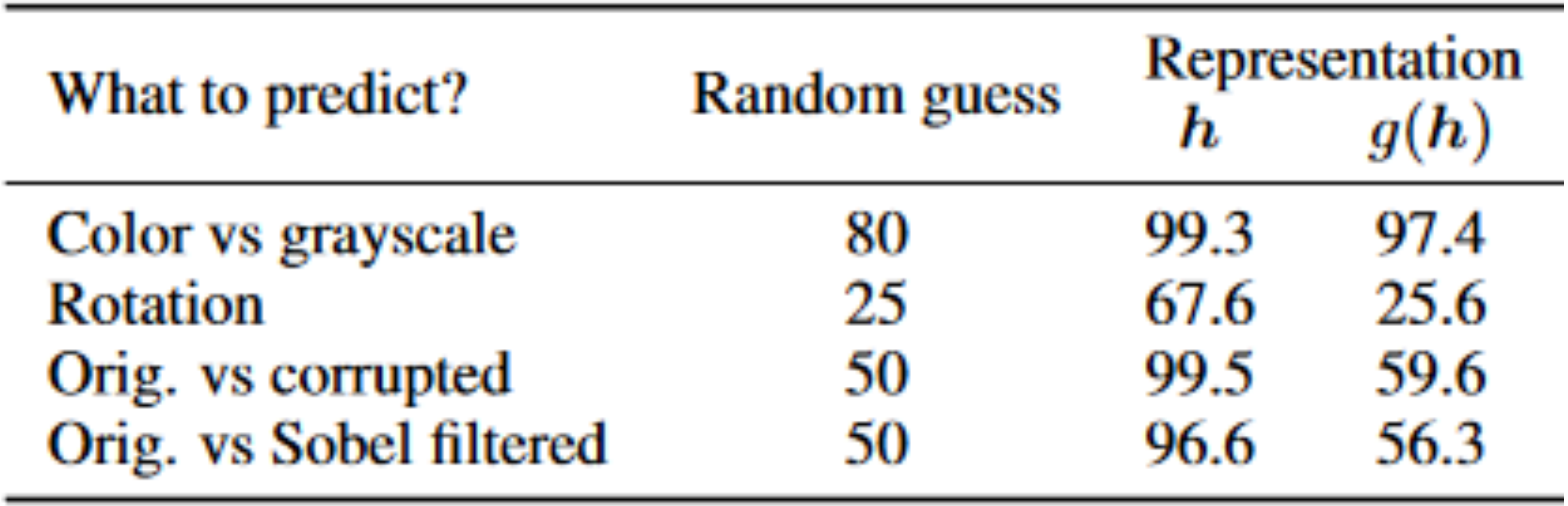
Figure 7. Linear evaluation of models with varied depth and width. Models in blue dots are ours trained for 100 epochs, models in red stars are ours trained for 1000 epochs, and models in green crosses are supervised ResNets trained for 90 epochs (He et al., 2016)

Non-linear Projection Head gave best results



- **Figure 8**. Linear evaluation of representations with different projection heads g(·)and various dimensions of z=g(h). The representation h (before projection) is 2048-dimensional here.

Projection Head testing of other Transforms



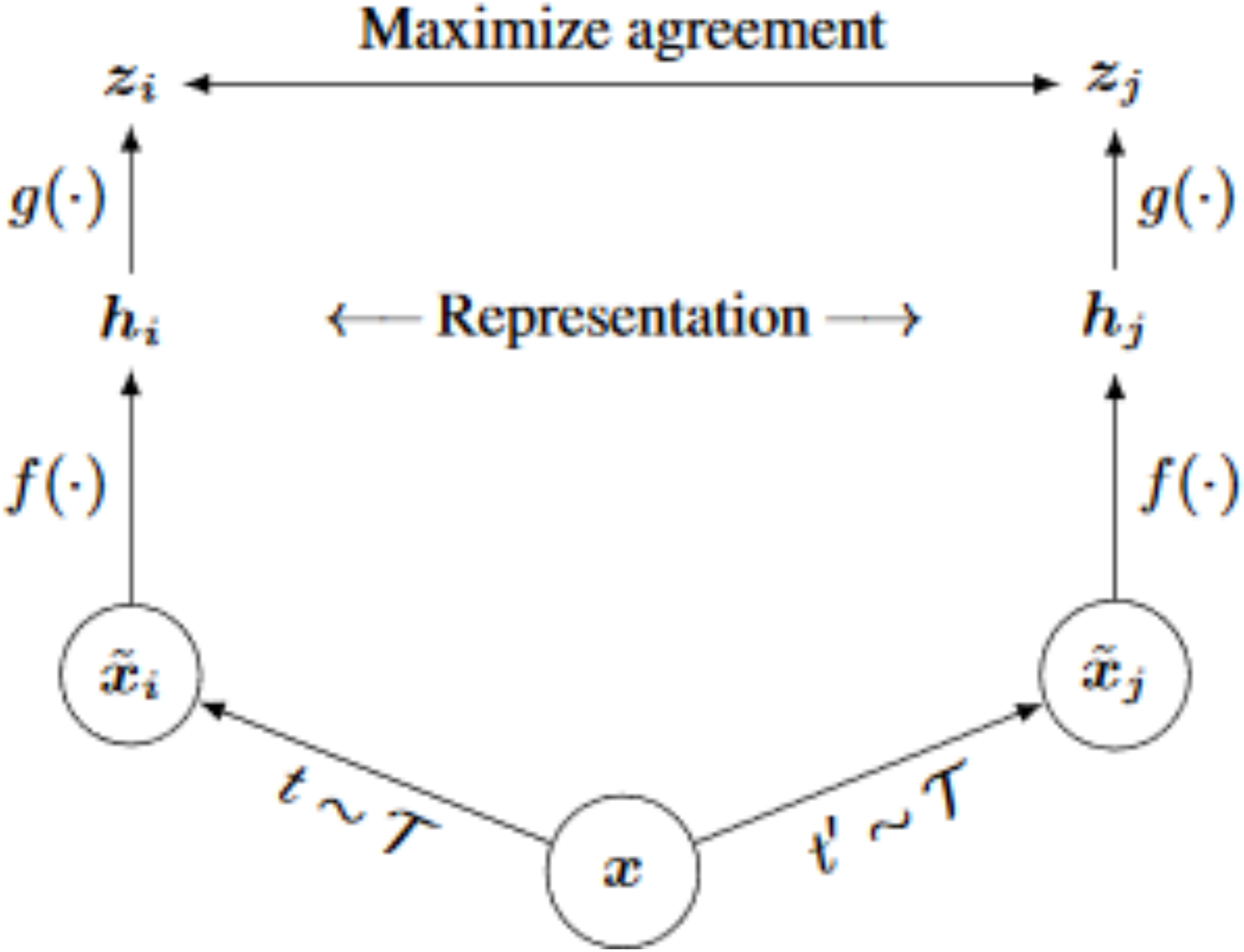
What to predict?	Random guess	Representation h	$g(h)$
Color vs grayscale	80	99.3	97.4
Rotation	25	67.6	25.6
Orig. vs corrupted	50	99.5	59.6
Orig. vs Sobel filtered	50	96.6	56.3

Table 3. Accuracy of training additional MLPs on different representations to predict the transformation applied. Other than crop and color augmentation, we additionally and independently add rotation (one of {0◦,90◦,180◦,270◦}), Gaussian noise, and Sobel filtering transformation during the pretraining for the last three rows. Both h and g(h)are of the same dimensionality, i.e. 2048.

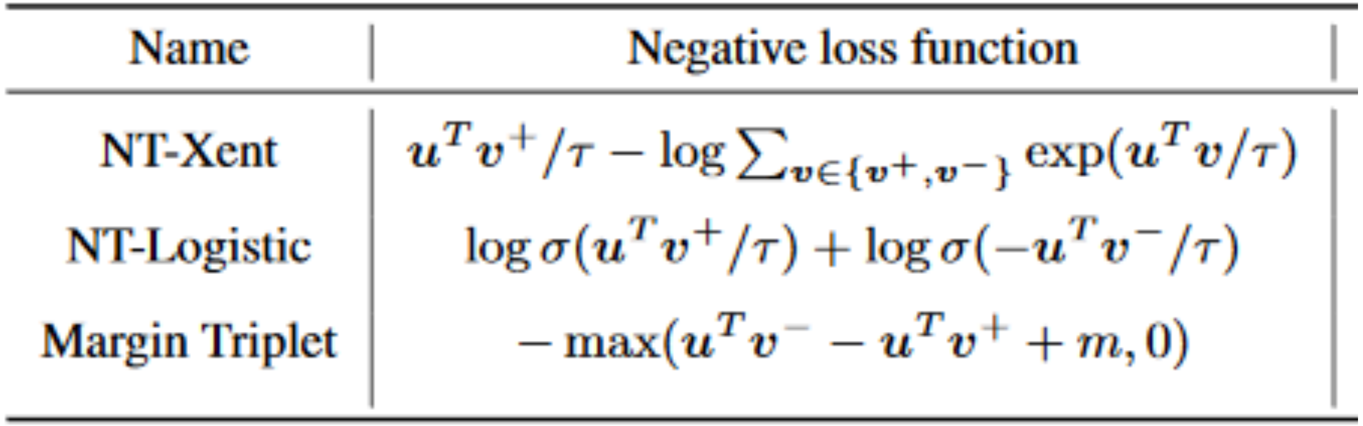
Projection head not used in final classification

- Use of projection head is only used for training encoder via contrastive loss
- Layer before projection head *g* is a better feature representation for the image.
- “z=g(h) is trained to be invariant to data transformation. Thus, g can remove information that may be useful for the downstream task, such as the color or orientation of objects”

Contrastive Loss

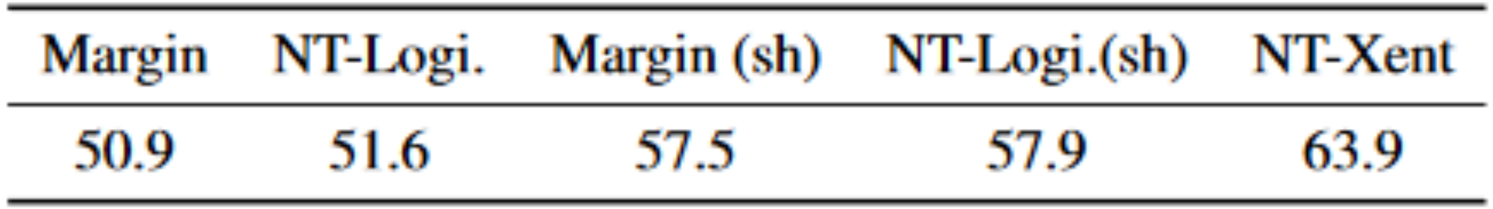


NT-Xent Shows best performance of loss functions



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

Table 2. Negative loss functions. All input vectors, i.e. u, v+, v−, are`l2 normalized. NT-Xent is an abbreviation for “Normalized Temperature-scaled Cross Entropy”. Different loss functions impose different weightings of positive and negative examples



Margin	NT-Logi.	Margin (sh)	NT-Logi.(sh)	NT-Xent
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.

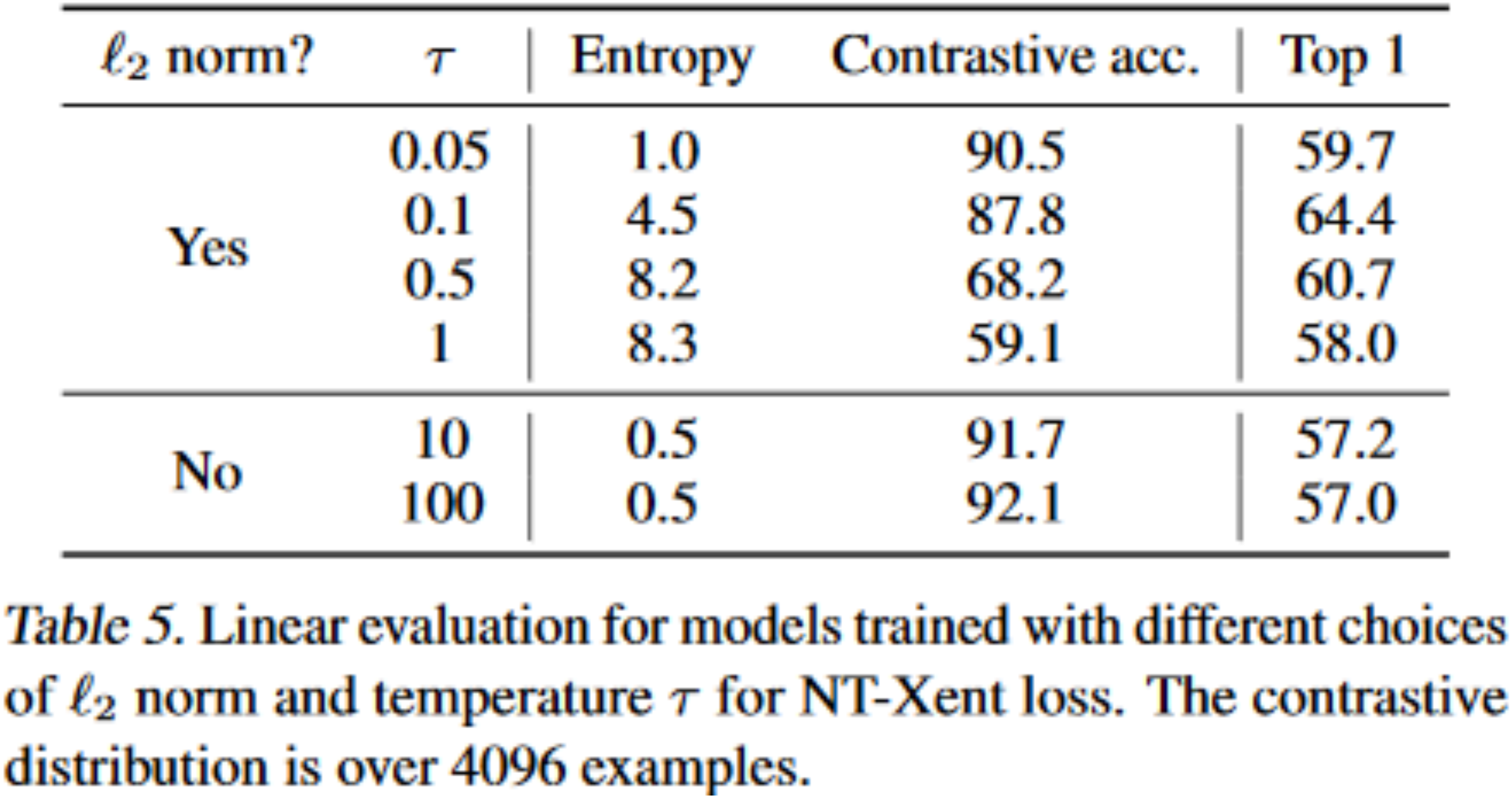
NT-Xent

$$\ell_{i,j} = -\log \frac{\exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}\,, \tag{1}$$

τ is temperature constant.
Where sim() is the normalized dot product between two vectors

$$\mathrm{sim}(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}^\top \boldsymbol{v} / \|\boldsymbol{u}\| \|\boldsymbol{v}\|$$

Experimentation with Temperature and Norm



ℓ_2 norm?	τ	Entropy	Contrastive acc.	Top 1
Yes	0.05	1.0	90.5	59.7
	0.1	4.5	87.8	64.4
	0.5	8.2	68.2	60.7
	1	8.3	59.1	58.0
No	10	0.5	91.7	57.2
	100	0.5	92.1	57.0

Table 5. Linear evaluation for models trained with different choices of ℓ_2 norm and temperature τ for NT-Xent loss. The contrastive distribution is over 4096 examples.

Final Testing on Classification Task

- ImageNet (among other databases) used
- Contrastive learning applied. Learned encoder is then attached to linear classifier (also ResNet-50)

Contrastive Learning Benefits From Larger batch sizes and longer training

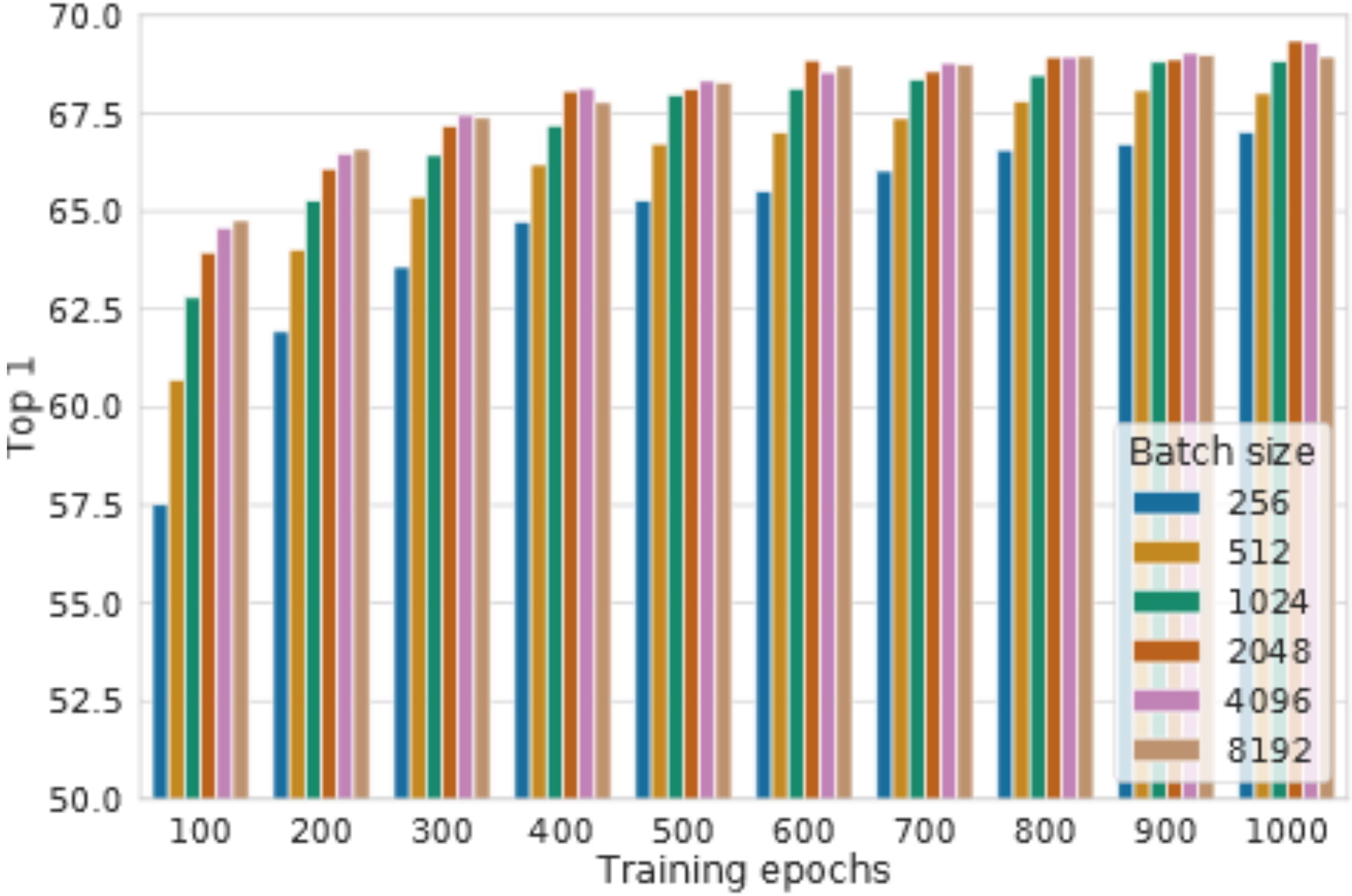
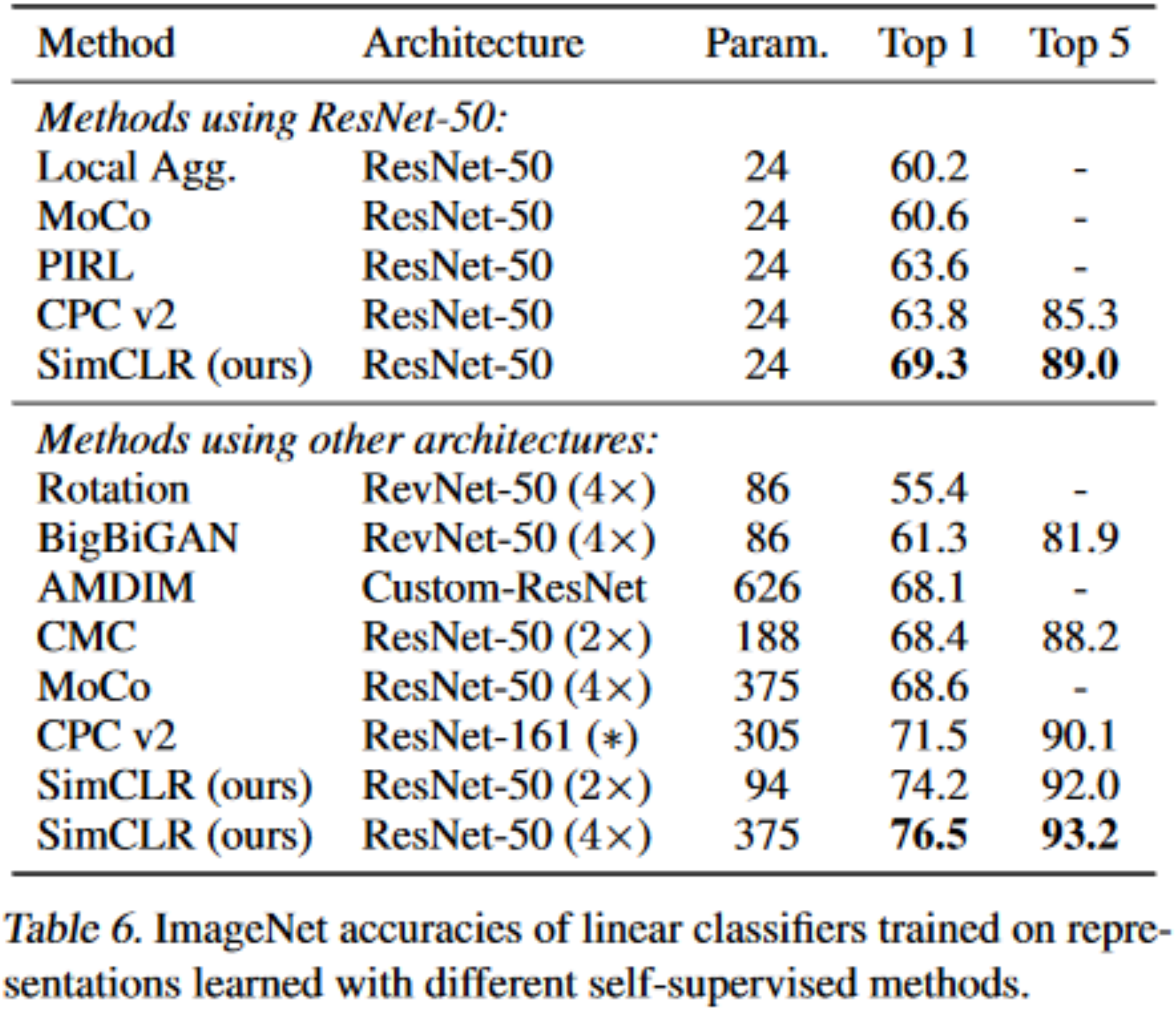


Figure 9. Linear evaluation models (ResNet-50) trained with different batch size and epochs. Each bar is a single run from scratch.

Proposed reason: Longer contrastive training or larger batches means more negative (mismatch) examples

Comparison to State-of-the-Art



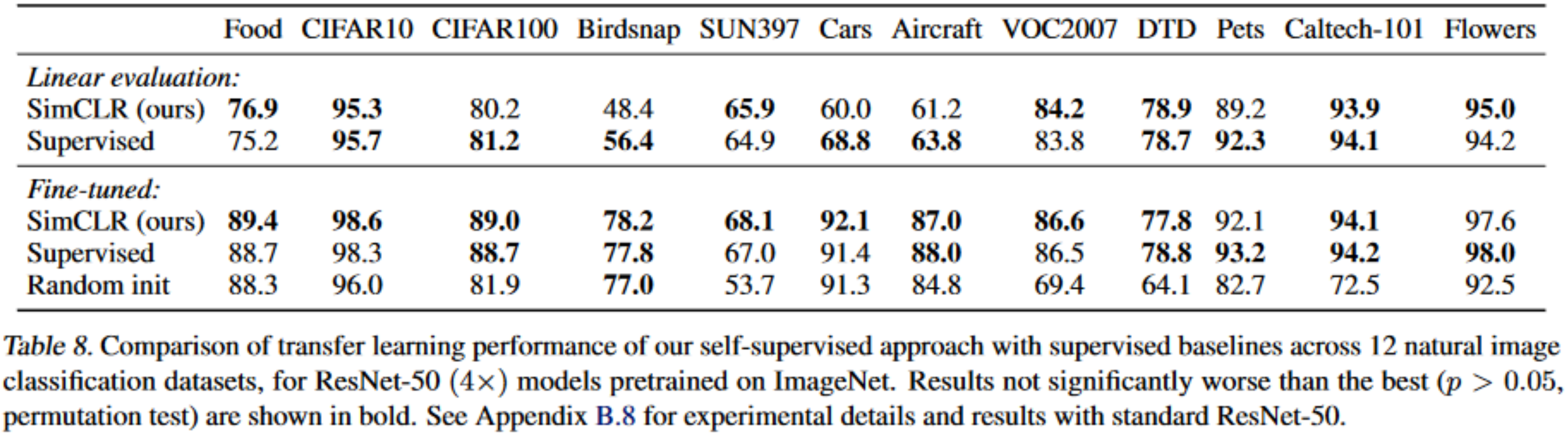
Method	Architecture	Param.	Top 1	Top 5
Methods using ResNet-50:				
Local Agg.	ResNet-50	24	60.2	-
MoCo	ResNet-50	24	60.6	-
PIRL	ResNet-50	24	63.6	-
CPC v2	ResNet-50	24	63.8	85.3
SimCLR (ours)	ResNet-50	24	**69.3**	**89.0**
Methods using other architectures:				
Rotation	RevNet-50 (4×)	86	55.4	-
BigBiGAN	RevNet-50 (4×)	86	61.3	81.9
AMDIM	Custom-ResNet	626	68.1	-
CMC	ResNet-50 (2×)	188	68.4	88.2
MoCo	ResNet-50 (4×)	375	68.6	-
CPC v2	ResNet-161 (*)	305	71.5	90.1
SimCLR (ours)	ResNet-50 (2×)	94	74.2	92.0
SimCLR (ours)	ResNet-50 (4×)	375	**76.5**	**93.2**

Table 6. ImageNet accuracies of linear classifiers trained on representations learned with different self-supervised methods.

		Label fraction	
Method	Architecture	1%	10%
		Top 5	
Supervised baseline	ResNet-50	48.4	80.4
Methods using other label-propagation:			
Pseudo-label	ResNet-50	51.6	82.4
VAT+Entropy Min.	ResNet-50	47.0	83.4
UDA (w. RandAug)	ResNet-50	-	88.5
FixMatch (w. RandAug)	ResNet-50	-	89.1
S4L (Rot+VAT+En. M.)	ResNet-50 (4×)	-	91.2
Methods using representation learning only:			
InstDisc	ResNet-50	39.2	77.4
BigBiGAN	RevNet-50 (4×)	55.2	78.8
PIRL	ResNet-50	57.2	83.8
CPC v2	ResNet-161($*$)	77.9	91.2
SimCLR (ours)	ResNet-50	75.5	87.8
SimCLR (ours)	ResNet-50 (2×)	83.0	91.2
SimCLR (ours)	ResNet-50 (4×)	**85.8**	**92.6**

Table 7. ImageNet accuracy of models trained with few labels.

Evaluation on other Classification Tasks



	Food	CIFAR10	CIFAR100	Birdsnap	SUN397	Cars	Aircraft	VOC2007	DTD	Pets	Caltech-101	Flowers
Linear evaluation:												
SimCLR (ours)	**76.9**	**95.3**	80.2	48.4	**65.9**	60.0	61.2	**84.2**	**78.9**	89.2	**93.9**	**95.0**
Supervised	75.2	**95.7**	**81.2**	**56.4**	64.9	**68.8**	**63.8**	83.8	**78.7**	**92.3**	**94.1**	94.2
Fine-tuned:												
SimCLR (ours)	**89.4**	**98.6**	**89.0**	**78.2**	**68.1**	**92.1**	**87.0**	**86.6**	**77.8**	92.1	**94.1**	97.6
Supervised	88.7	98.3	**88.7**	**77.8**	67.0	91.4	**88.0**	86.5	**78.8**	**93.2**	**94.2**	**98.0**
Random init	88.3	96.0	81.9	**77.0**	53.7	91.3	84.8	69.4	64.1	82.7	72.5	92.5

Table 8. Comparison of transfer learning performance of our self-supervised approach with supervised baselines across 12 natural image classification datasets, for ResNet-50 $(4\times)$ models pretrained on ImageNet. Results not significantly worse than the best ($p > 0.05$, permutation test) are shown in bold. See Appendix B.8 for experimental details and results with standard ResNet-50.

Conclusion

- Contrastive Learning with data augmentation is a viable method for teaching feature representations.
- Data augmentation strength, training batch size, etc. are critical parameters for success with contrastive learning
- “We note that the superiority of our framework relative to previous work is not explained by any single design choice, but by their composition.”