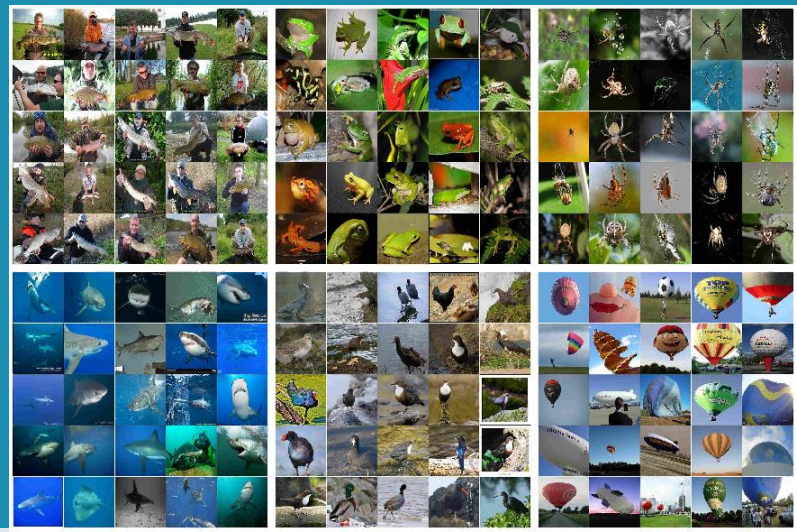
KU LEUVEN PSI ETH zürich CVL

ECCV'20 ONLINE 23-28 AUGUST 2020

SCAN: Learning to Classify Images without Labels



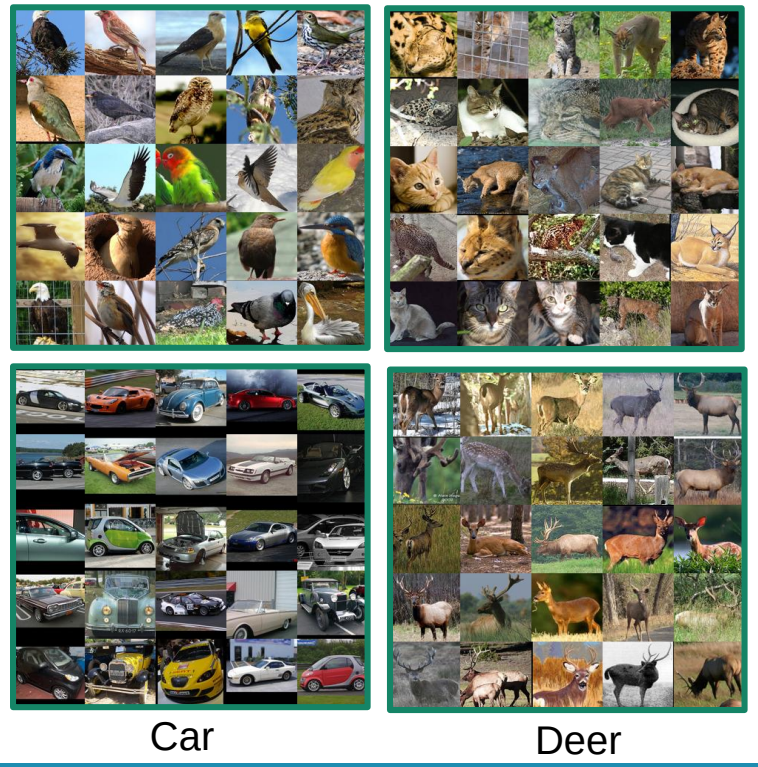
Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans and Luc Van Gool

Unsupervised Image Classification

Task: Group a set unlabeled images into semantically meaningful clusters.

Unlabeled Data

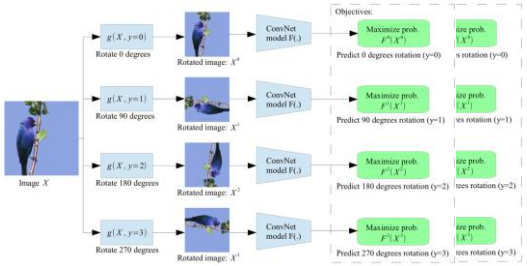
Cluster

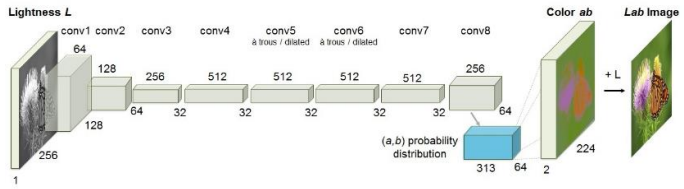


Prior work – Two dominant paradigms

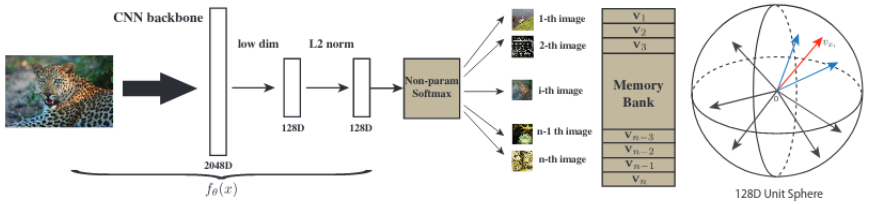
I. Representation Learning

Idea: Use a self-supervised learning pretext task
+ off-line clustering (K-means)





Ex 1: Predict Transformations



Ex 2: Instance Discrimination

Problem: K-means leads to cluster degeneracy.

II. End-To-End Learning

Idea: - Leverage architecture of CNNs as a prior.
(e.g. DAC, DeepCluster, DEC, etc.)

or - Maximize mutual information between an image and its augmentations
(e.g. IMSAT, IIC)

Problems:

- Cluster learning depends on initialization, and is likely to latch onto low-level features.

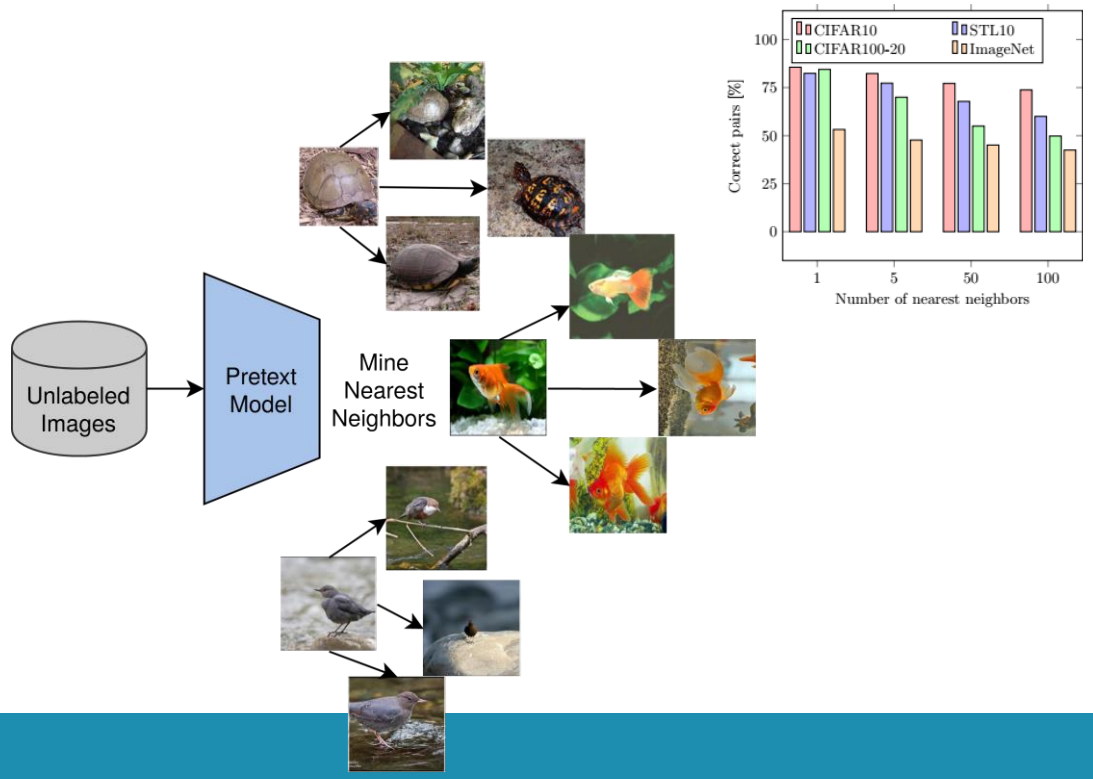
- Special mechanisms required (Sobel, PCA, cluster re-assignments, etc.).

[1] Unsupervised representation learning by predicting image rotations, Gidaris et al. (2018)
[2] Colorful Image Colorization, Richard et al. (2016)
[3] Unsupervised feature learning via non-parametric instance discrimination, Wu et al. (2018)

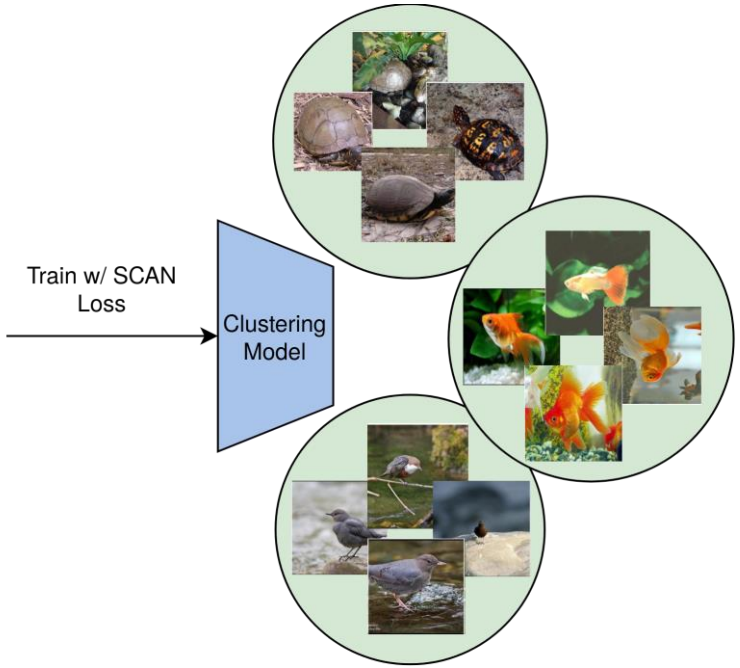
SCAN: Semantic Clustering by Adopting Nearest Neighbors

Approach: A two-step approach where feature learning and clustering are decoupled.

Step 1: Solve a pretext task + Mine k-NN

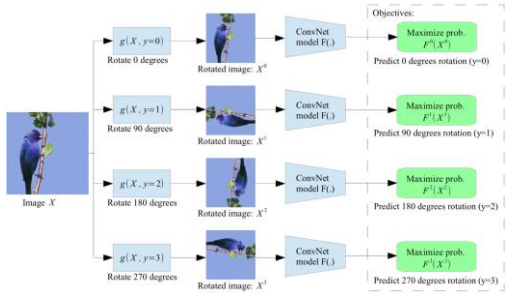


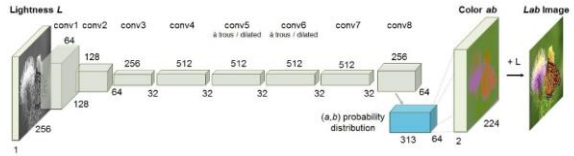
Step 2: Train clustering model by imposing consistent predictions among neighbors

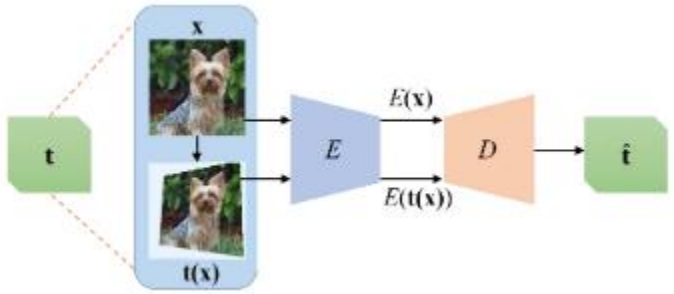


Step 1: Solve a pretext task + Mine k-NN

Question: *How to select a pretext task appropriate for the down-stream task of semantic clustering?*







Problem: Pretext tasks which try to predict image transformations result in a feature representation that is covariant to the applied transformation.

→ Undesired for the down-stream task of semantic clustering.

→ ***Solution:*** Pretext model should minimize the distance between an image and its augmentations.

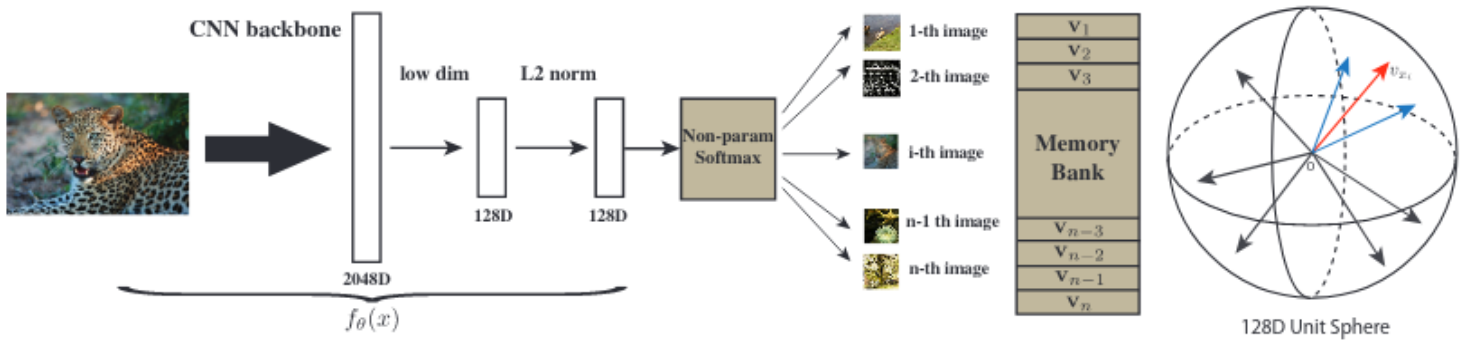
$$\min_{\theta} d(\Phi_\theta(X_i), \Phi_\theta(T[X_i]))$$

[1] Unsupervised representation learning by predicting image rotations, Gidaris et al. (2018)
[2] Colorful Image Colorization, Richard et al. (2016)
[3] AET vs AED, Zhang et al. (2019)

Step 1: Solve a pretext task + Mine k-NN

Question: *How to select a pretext task appropriate for the down-stream task of semantic clustering?*

$$\min_{\theta} d(\Phi_\theta(X_i), \Phi_\theta(T[X_i]))$$

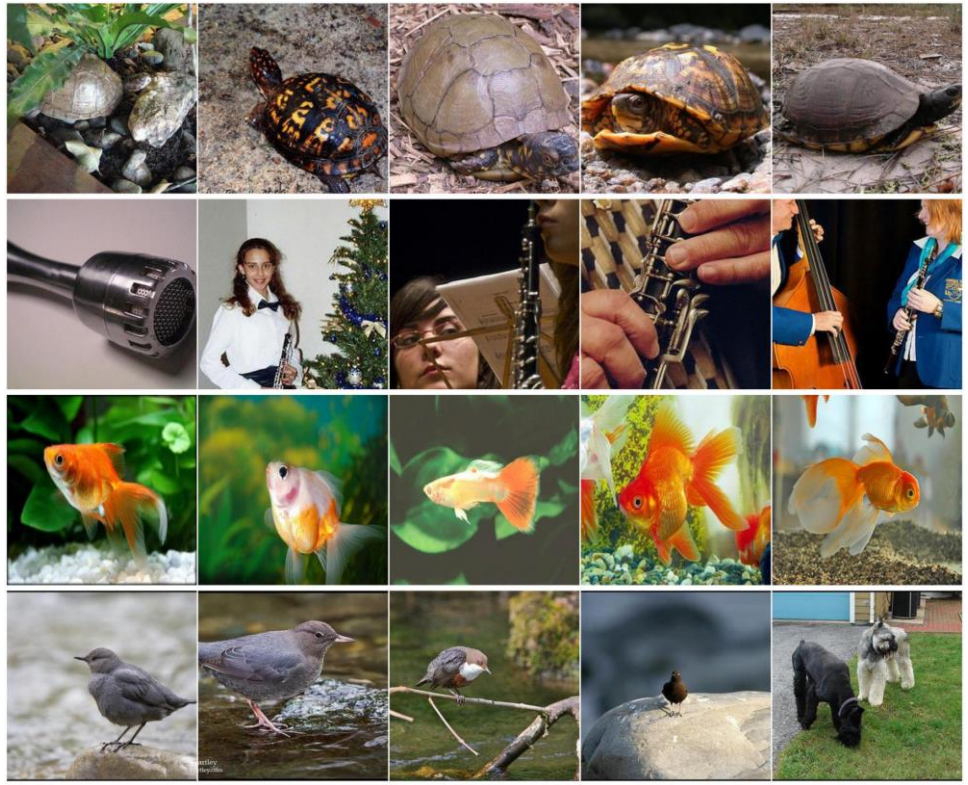


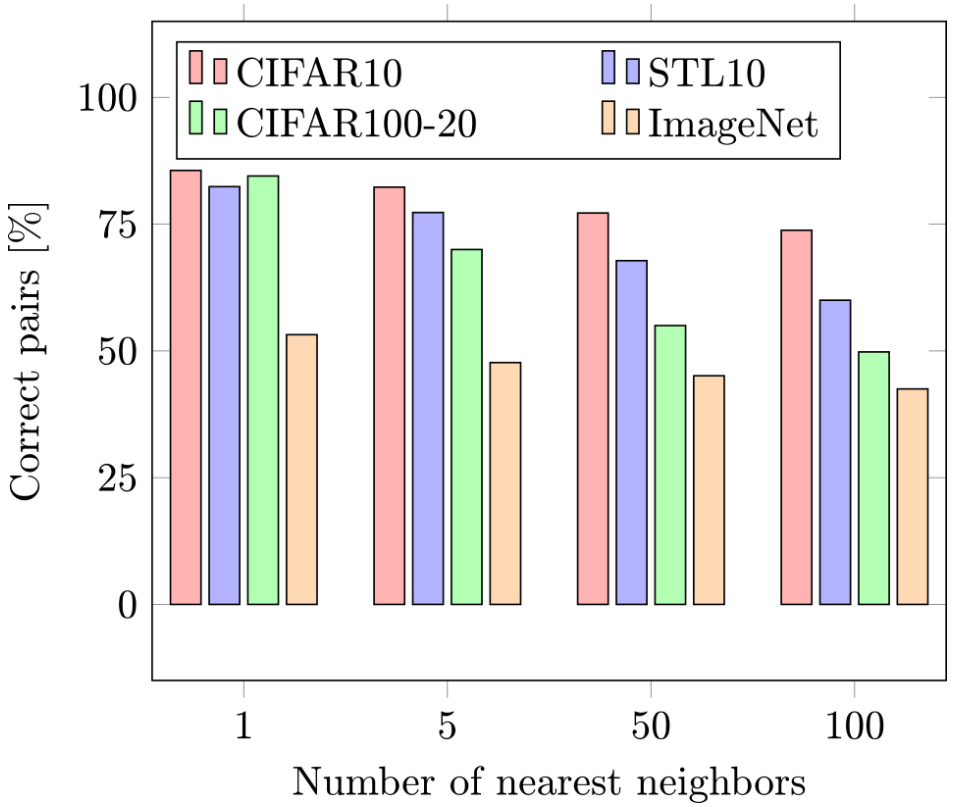
Instance discrimination satisfies the invariance criterion w.r.t. augmentations applied during training.

[1] Unsupervised feature learning via non-parametric instance discrimination, Wu et al. (2018)

Step 1: Solve a pretext task + Mine k-NN

The nearest neighbors tend to belong to the same semantic class.





Step 2: Train clustering model

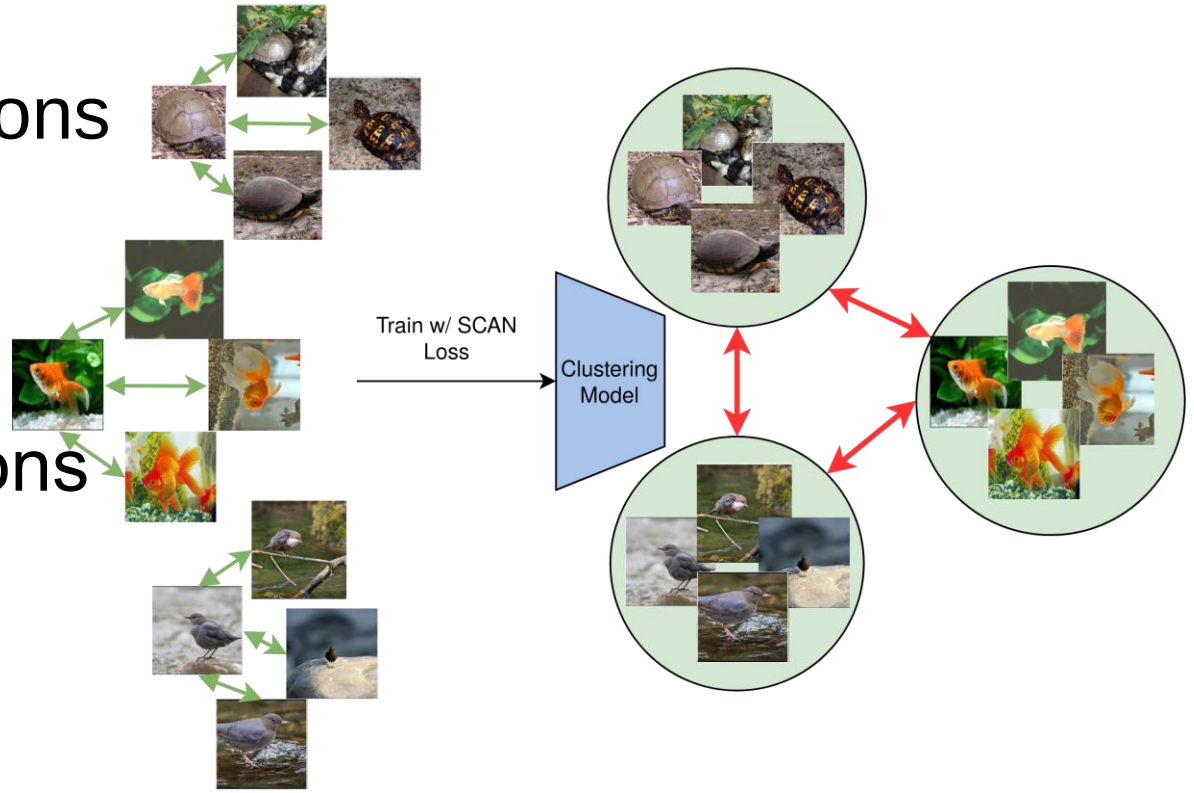
- SCAN-Loss:

(1) Enforce consistent predictions among neighbors. Maximize:

$$\log \langle \Phi_\eta(X), \Phi_\eta(k) \rangle$$

→ Dot product forces predictions to be one-hot (confident)

(2) Maximize entropy to avoid all samples being assigned to the same cluster.

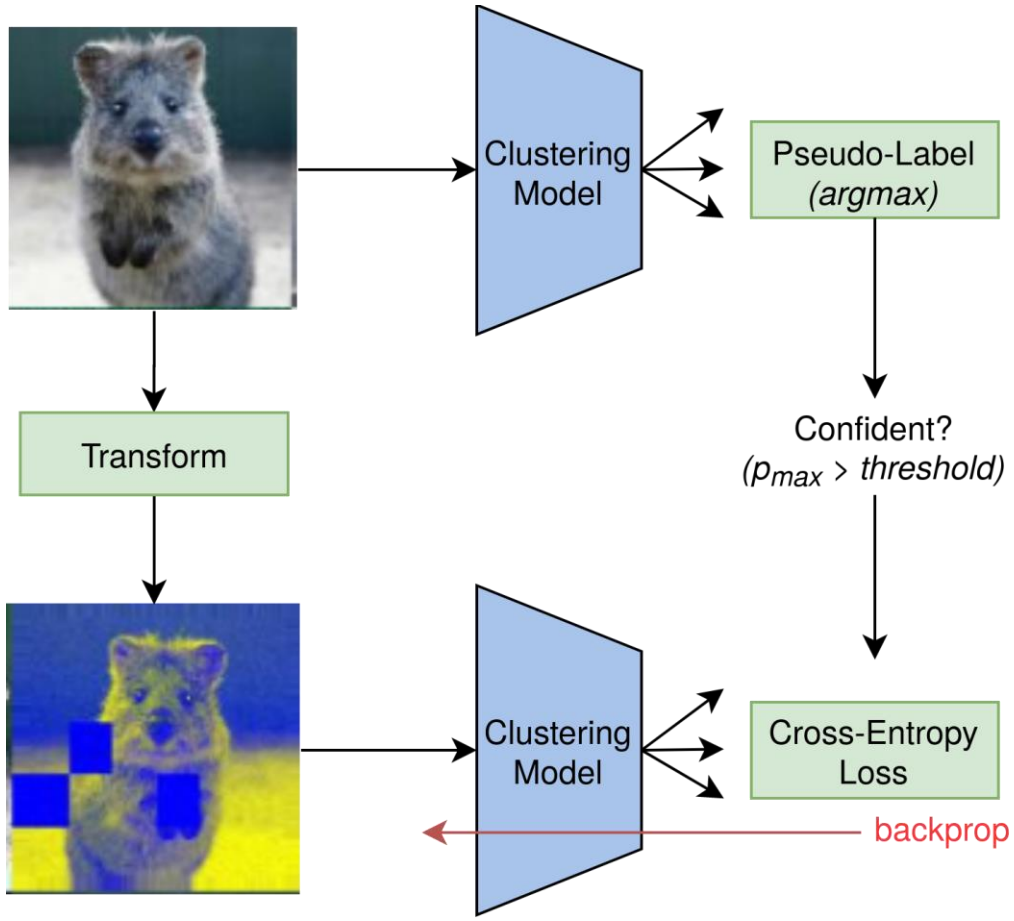


Step 2b: Refinement through self-labeling

- Refine the model through self-labeling

- Apply a cross-entropy loss on strongly augmented [1] versions of confident samples.

- Applying strong augmentations avoids overfitting.



[1] RandAugment, Cubuk et al. (2020)
[2] FixMatch, Sohn et al. (2020)
[3] Probability of error, Scudder H. (1965)

Experimental setup

- ResNet backbone + Identical hyperparameters.

- SimCLR and MoCo implementation for the pretext task.

- Experiments on four datasets

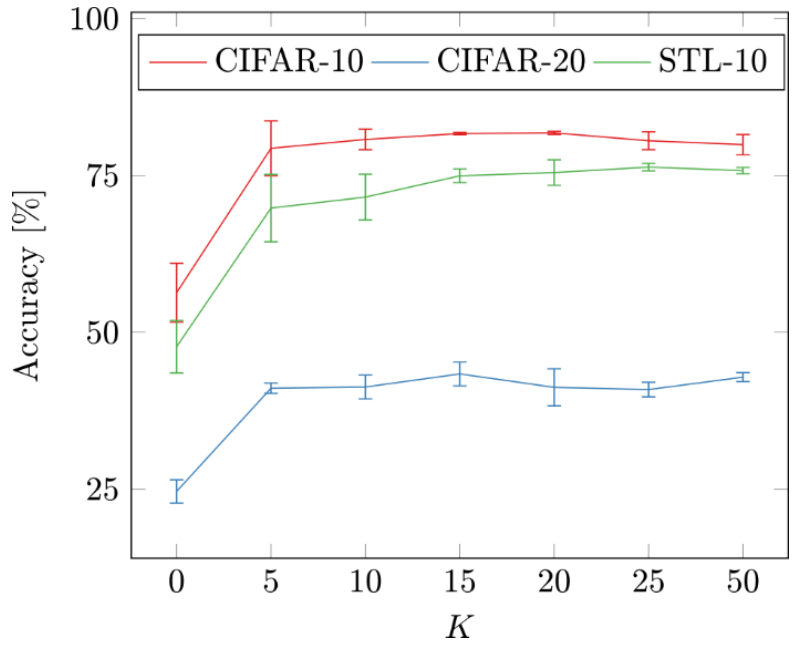
Dataset	Classes	Train images	Val images	Aspect ratio
CIFAR10	10	50,000	10,000	32 x 32
CIFAR100-20	20	50,000	10,000	32 x 32
STL10	10	5,000	8,000	96 x 96
ImageNet	1000	1,281,167	50,000	224 x 224

Ablation studies - SCAN

- Pretext task

Pretext Task	ACC (Avg +- Std)
Rotation Prediction	74.3 +- 3.9
Instance Discrimination	87.6 +- 0.4

- Number of NNs (K)

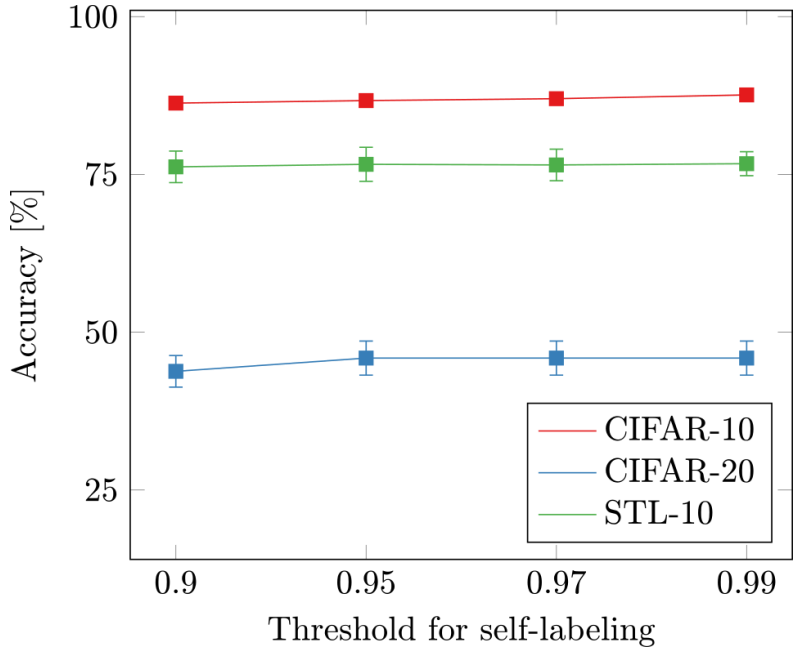


Ablation studies - Self-label

Self-labeling (CIFAR-10)

Step	ACC (Avg +- Std)
SCAN	81.8 +- 0.3
Self-labeling	87.6 +- 0.4

Threshold self-labeling

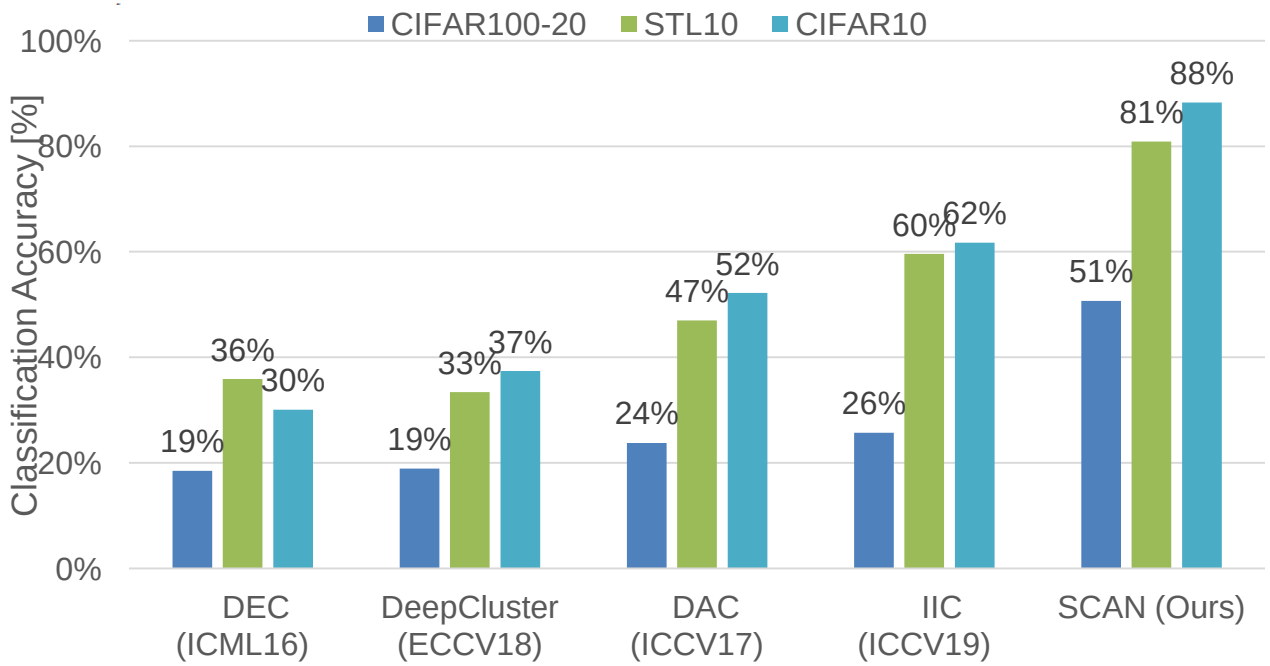


Comparison with SOTA

Dataset	CIFAR10			CIFAR100-20			STL10		
Metric	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
K-means	22.9	8.7	4.9	13.0	8.4	2.8	19.2	12.5	6.1
SC	24.7	10.3	8.5	13.6	9.0	2.2	15.9	9.8	4.8
Triplets	20.5	–	–	9.94	–	–	24.4	–	–
JULE	27.2	19.2	13.8	13.7	10.3	3.3	27.7	18.2	16.4
AEVB	29.1	24.5	16.8	15.2	10.8	4.0	28.2	20.0	14.6
SAE	29.7	24.7	15.6	15.7	10.9	4.4	32.0	25.2	16.1
DAE	29.7	25.1	16.3	15.1	11.1	4.6	30.2	22.4	15.2
SWWAE	28.4	23.3	16.4	14.7	10.3	3.9	27.0	19.6	13.6
AE	31.4	23.4	16.9	16.5	10.0	4.7	30.3	25.0	16.1
GAN	31.5	26.5	17.6	15.1	12.0	4.5	29.8	21.0	13.9
DEC	30.1	25.7	16.1	18.5	13.6	5.0	35.9	27.6	18.6
ADC	32.5	–	–	16.0	–	–	53.0	–	–
DeepCluster	37.4	–	–	18.9	–	–	33.4	–	–
DAC	52.2	40.0	30.1	23.8	18.5	8.8	47.0	36.6	25.6
IIC	61.7	51.1	41.1	25.7	22.5	11.7	59.6	49.6	39.7
SCAN† (Avg $\pm$ Std)	87.6 ± 0.4	78.7 ± 0.5	75.8 ± 0.7	45.9 ± 2.7	46.8 ± 1.3	30.1 ± 2.1	76.7 ± 1.9	68.0 ± 1.2	61.6 ± 1.8
SCAN† (Best)	**88.3**	**79.7**	**77.2**	**50.7**	**48.6**	**33.3**	**80.9**	**69.8**	**64.6**

Comparison with SOTA

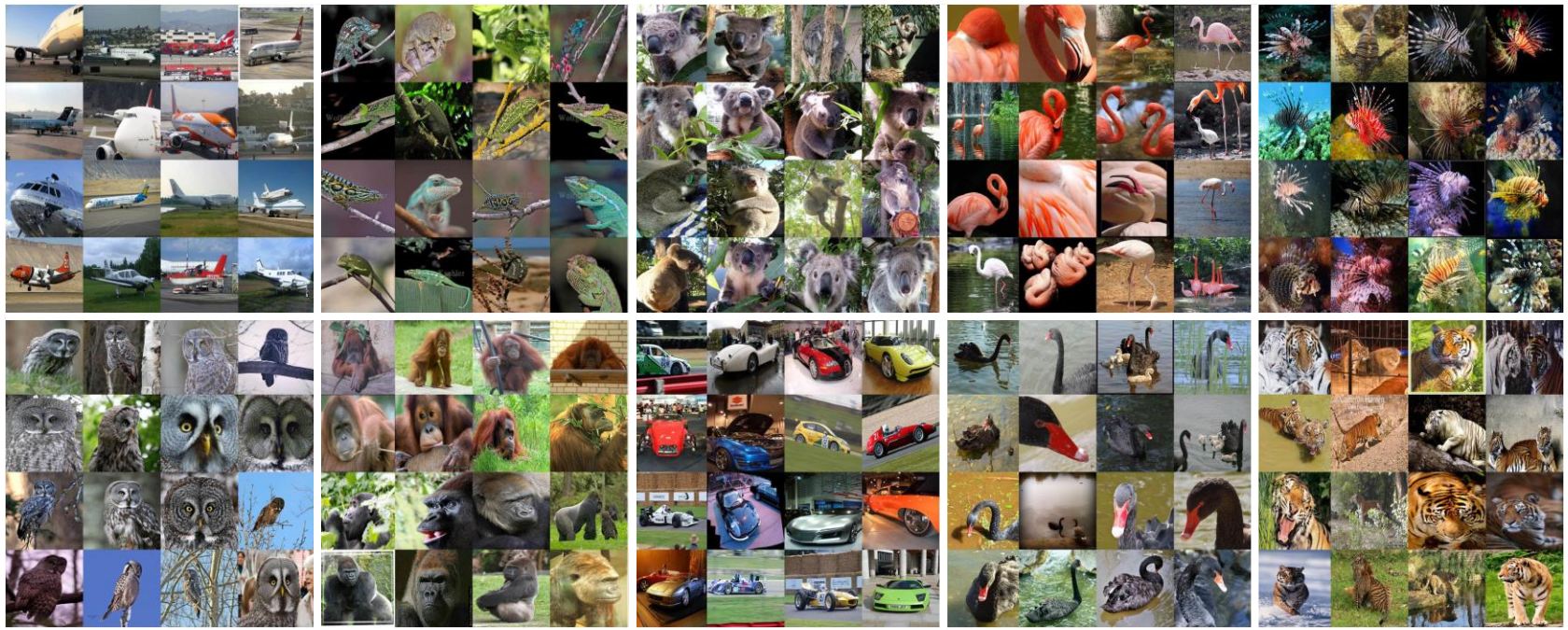
Dataset	CIFAR10			CIFAR100-20			STL10		
Metric	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
Supervised	93.8	86.2	87.0	80.0	68.0	63.2	80.6	65.9	63.1
Pretext + K-means	65.9 ± 5.7	59.8 ± 2.0	50.9 ± 3.7	39.5 ± 1.9	40.2 ± 1.1	23.9 ± 1.1	65.8 ± 5.1	60.4 ± 2.5	50.6 ± 4.1
SCAN* (Avg ± Std)	81.8 ± 0.3	71.2 ± 0.4	66.5 ± 0.4	42.2 ± 3.0	44.1 ± 1.0	26.7 ± 1.3	75.5 ± 2.0	65.4 ± 1.2	59.0 ± 1.6
SCAN† (Avg ± Std)	87.6 ± 0.4	78.7 ± 0.5	75.8 ± 0.7	45.9 ± 2.7	46.8 ± 1.3	30.1 ± 2.1	76.7 ± 1.9	68.0 ± 1.2	61.6 ± 1.8
SCAN† (Best)	**88.3**	**79.7**	**77.2**	**50.7**	**48.6**	**33.3**	**80.9**	**69.8**	**64.6**

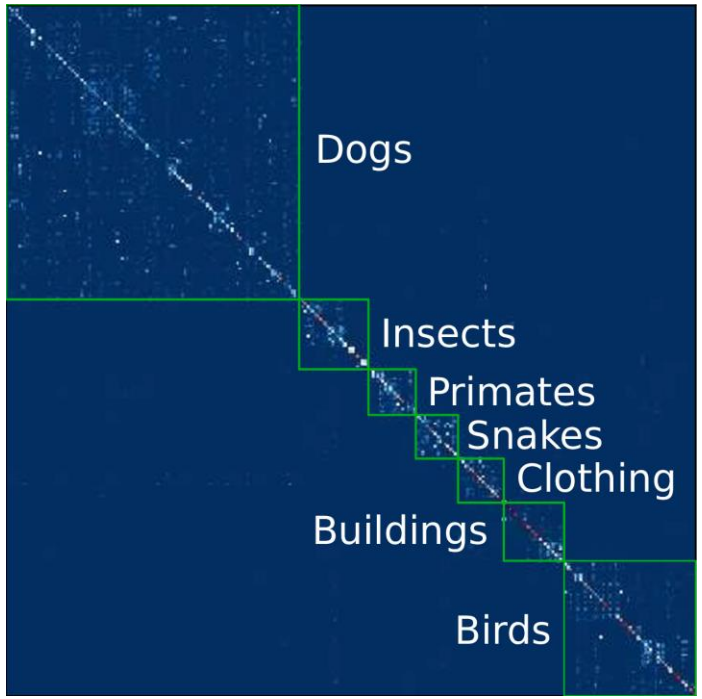


- Large performance gains w.r.t. to prior works: +26:6% on CIFAR10, +25:0% on CIFAR100-20 and +21:3% on STL10
- **SCAN** outperforms SimCLR + K-means
- Close to supervised performance on CIFAR-10 and STL-10

ImageNet Results

- **Scalable:** First method which scales to ImageNet (1000 classes)
- **Semantic clusters:** We observe that the clusters capture a large variety of different backgrounds, viewpoints, etc.
- **Confusion matrix** shows ImageNet hierarchy containing dogs, insects, primates, snakes, clothing, buildings, birds etc.





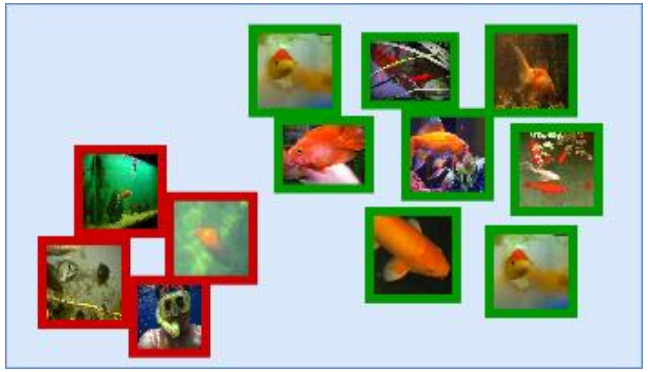
Comparison with supervised methods

- Trained with 1% of the labels
- **SCAN**: Top-1: 39.9%, Top-5: 60.0%, NMI: 72.0%, ARI: 27.5%

Method	Backbone	Labels	Top-1	Top-5
Supervised Baseline	ResNet-50	✓	25.4	48.4
Pseudo-Label	ResNet-50	✓	-	51.6
VAT + Entropy Min.	ResNet-50	✓	-	47.0
InstDisc	ResNet-50	✓	-	39.2
BigBiGAN	ResNet-50(4x)	✓	-	55.2
PIR	ResNet-50	✓	-	57.2
CPC v2	ResNet-161	✓	52.7	77.9
SimCLR	ResNet-50	✓	48.3	75.5
SCAN (Ours)	ResNet-50	✗	39.9	60.0

Prototypical behavior

Prototype: The closest sample to the mean embedding of the high confident samples of a certain class.



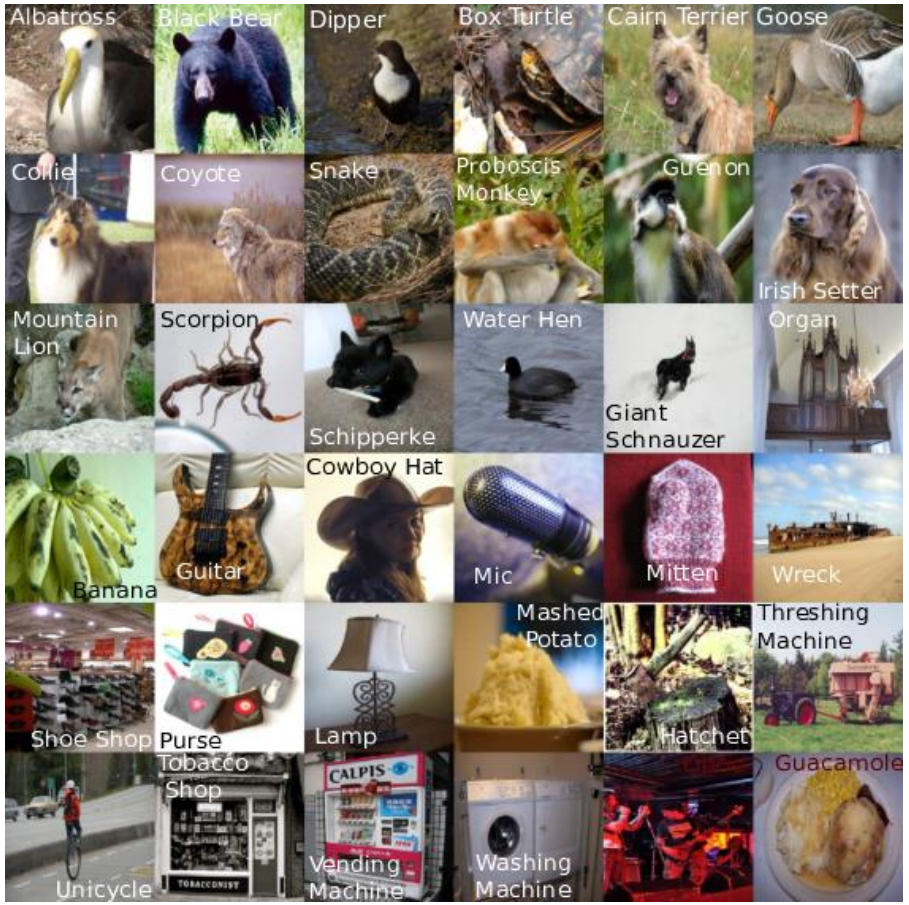
Prototypes:

- show what each cluster represents
- are often more **pure**

STL10

CIFAR10

ImageNet



Conclusion

- Two step approach: decouple feature learning and clustering
- Nearest neighbors capture variance in viewpoints and backgrounds
- Promising results on large scale datasets

Future directions

- Extension to other modalities, e.g. video, audio
- Other domains, e.g. segmentation, semi-supervised, etc.

Code is available on Github