香港中文大學
The Chinese University of Hong Kong

Mix-and-Match Tuning for Self-Supervised Semantic Segmentation

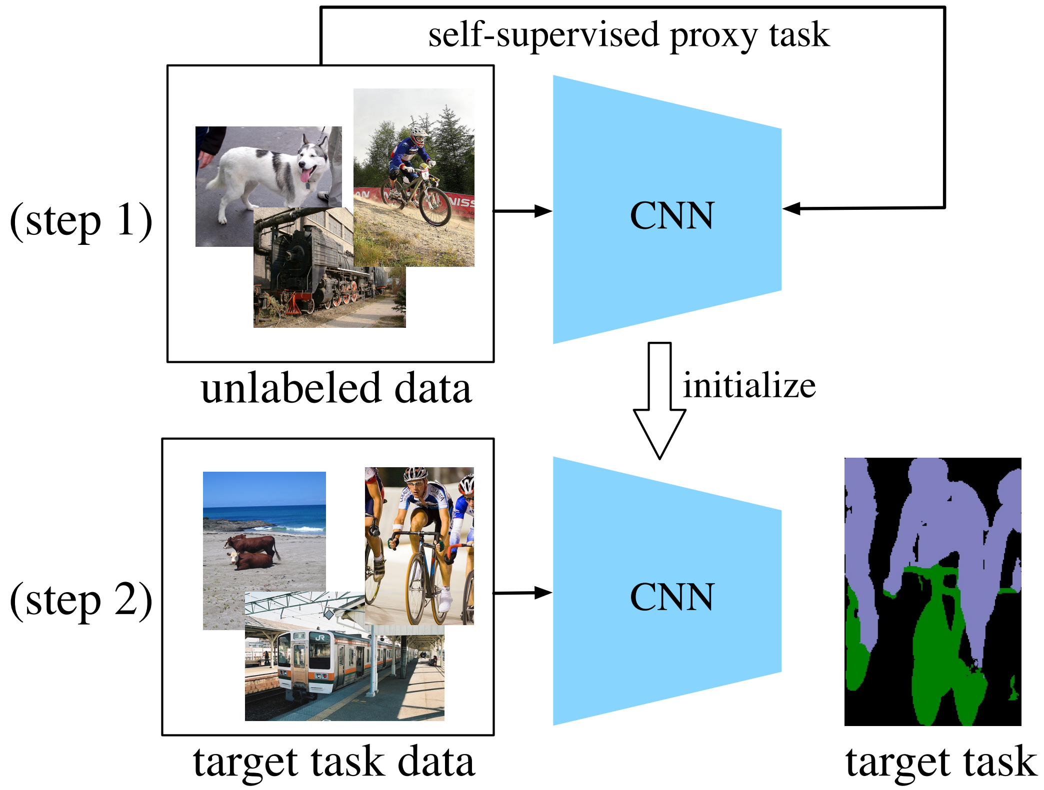
Xiaohang Zhan, Ziwei Liu, Ping Luo, Xiaoou Tang, Chen Change Loy

Department of Information Engineering, The Chinese University of Hong Kong

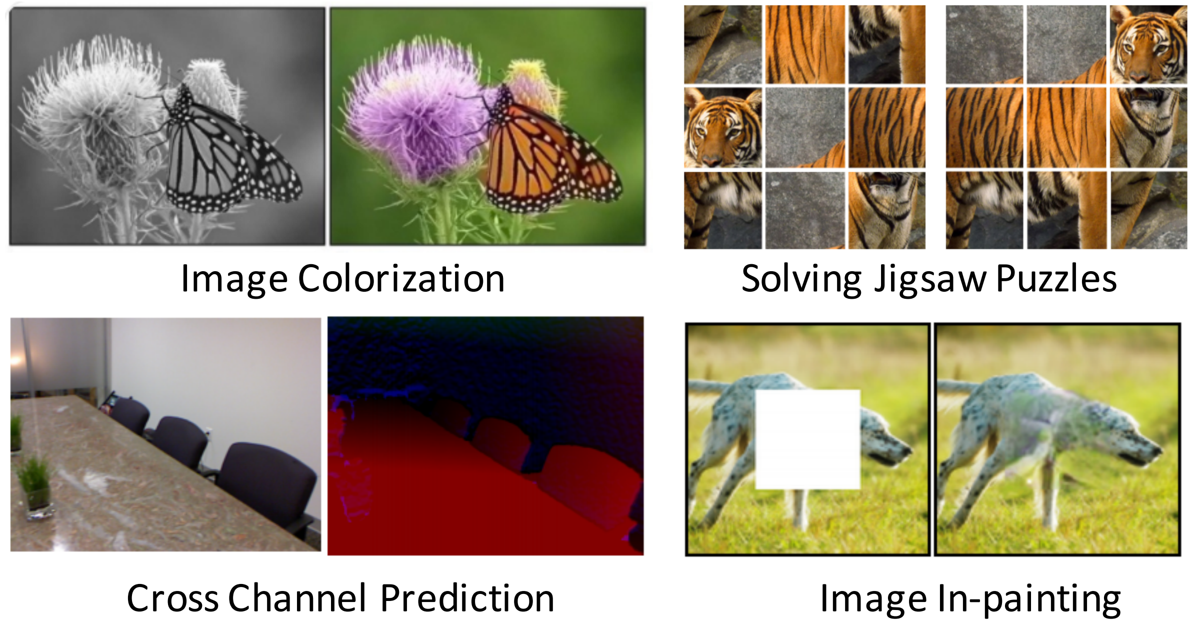
{zx017, lz013, pluo, xtang, ccloy}@ie.cuhk.edu.hk

Self-supervised Learning

香港中文大學
The Chinese University of Hong Kong

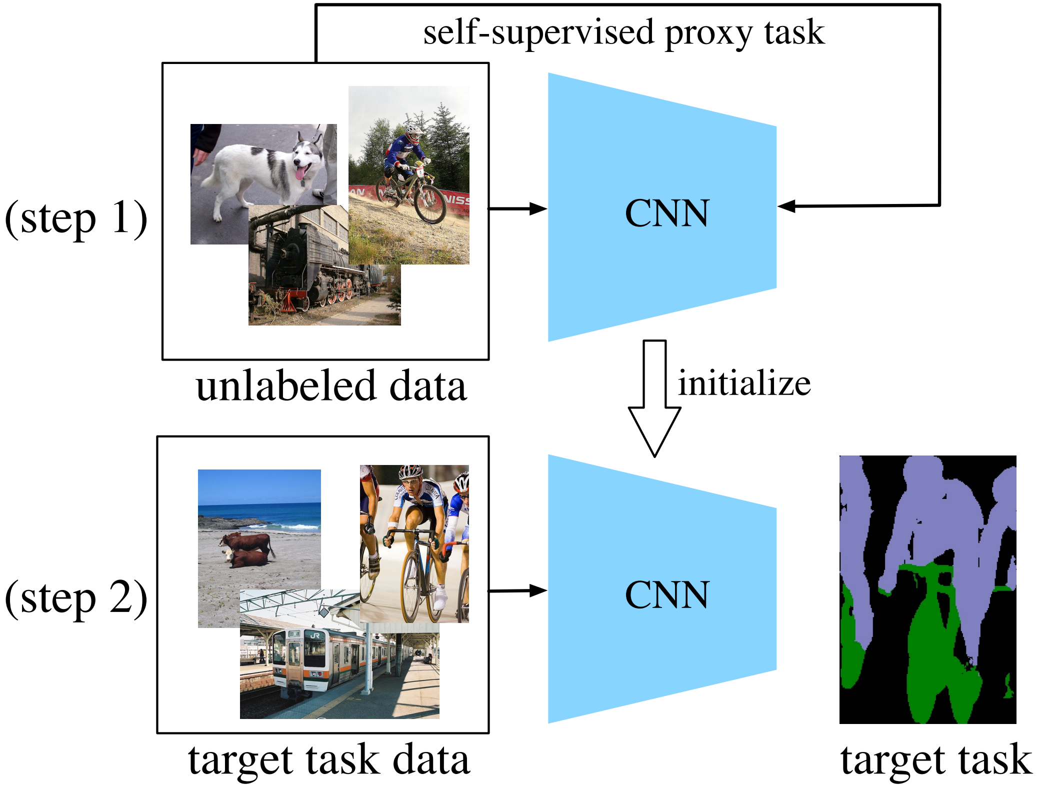


Self-supervised proxy task, e.g.:



香港中文大學
The Chinese University of Hong Kong

Self-supervised Learning



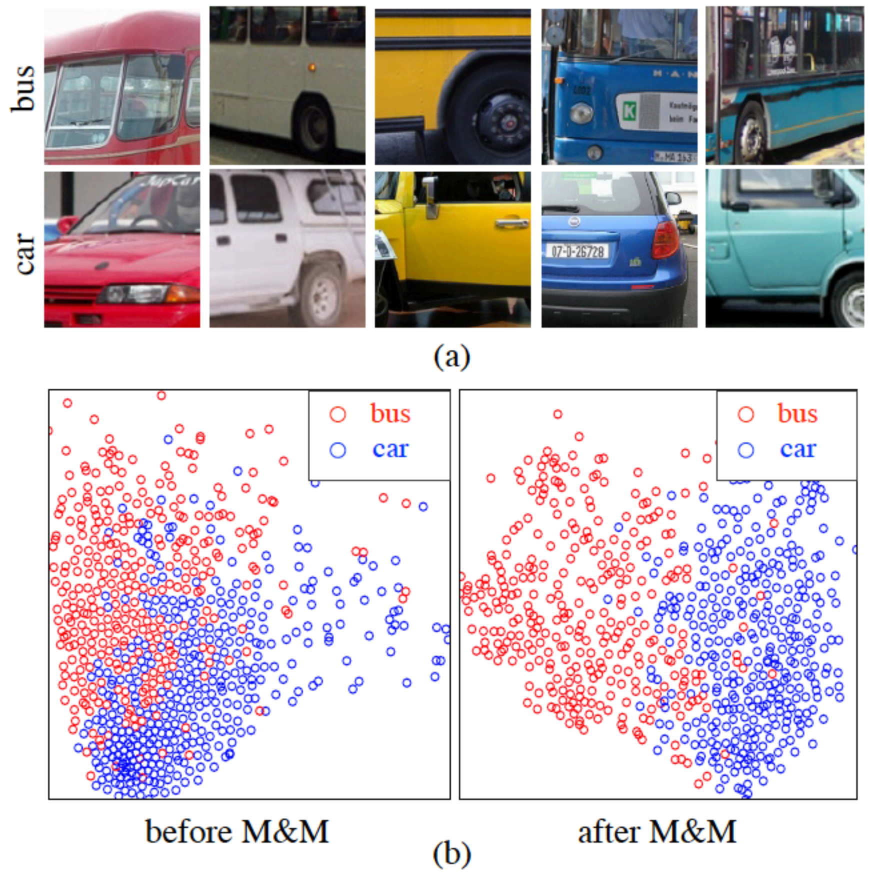
Method	Arch.	VOC12 %mIoU.
ImageNet	VGG-16	64.2
Random	VGG-16	35.0
Larsson *et al.* (Larsson, Maire, and Shakhnarovich 2016)	VGG-16	50.2
Larsson *et al.* (Larsson, Maire, and Shakhnarovich 2017)	VGG-16	56.0
Ours (M&M + Graph, colorization pre-trained)	VGG-16	**64.5**
ImageNet	AlexNet	48.0
Random	AlexNet	23.5
k-means (Krähenbühl et al. 2015)	AlexNet	32.6
Pathak *et al.* (Pathak et al. 2016b)	AlexNet	29.7
Donahue *et al.* (Donahue, Krähenbühl, and Darrell 2016)	AlexNet	35.2
Zhang *et al.* (Zhang, Isola, and Efros 2016a)	AlexNet	35.6
Zhang *et al.* (Zhang, Isola, and Efros 2016b)	AlexNet	36.0
Noroozi et al. (Noroozi and Favaro 2016)	AlexNet	37.6
Larsson *et al.* (Larsson, Maire, and Shakhnarovich 2017)	AlexNet	38.4
Ours (M&M + Random Triplets, colorization pre-trained)	AlexNet	40.9
Ours (M&M + Graph, colorization pre-trained)	AlexNet	42.8
Ours (M&M + Graph, randomly initialized)	AlexNet	**43.6**

8.2%

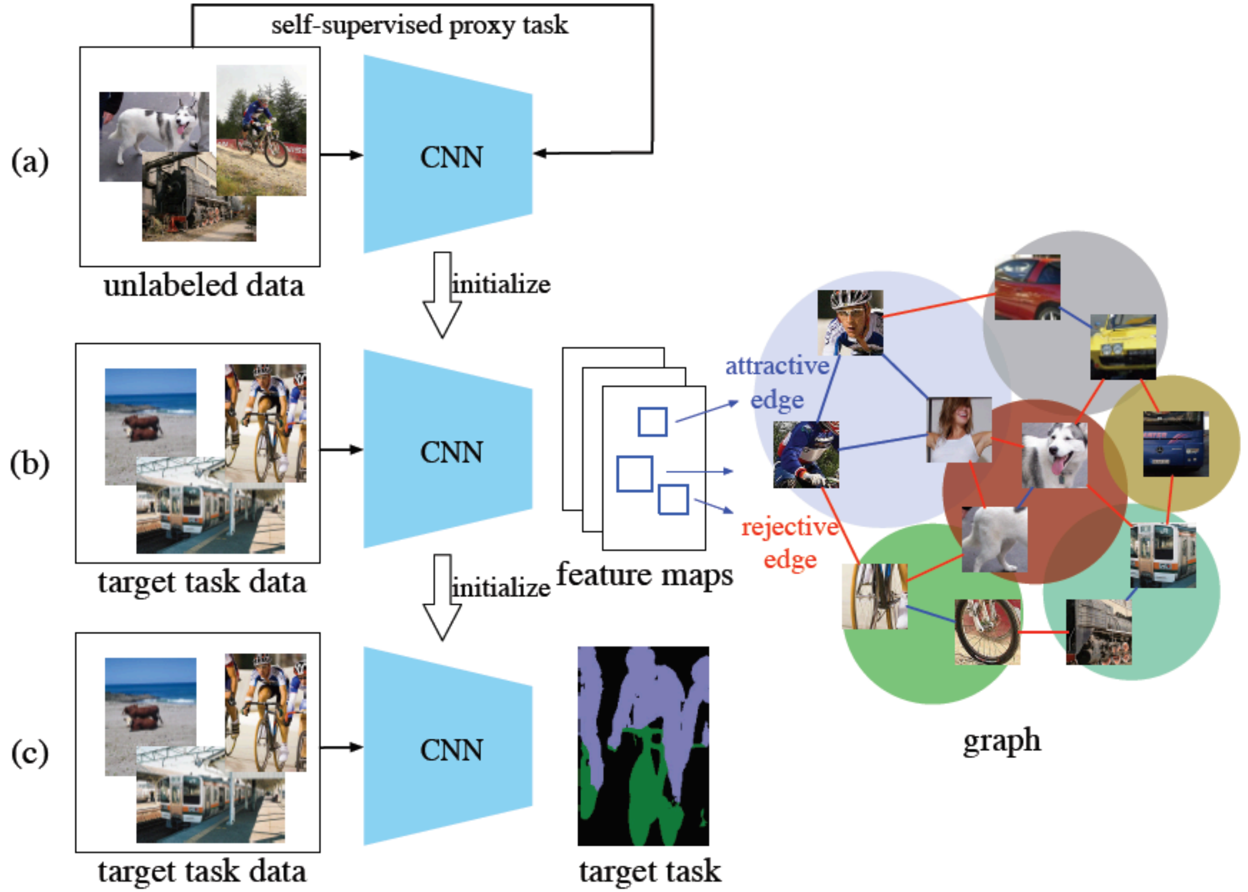
9.6%

Sematic Segmentation benchmark (PASCAL VOC 2012 validation set)

Task Gap



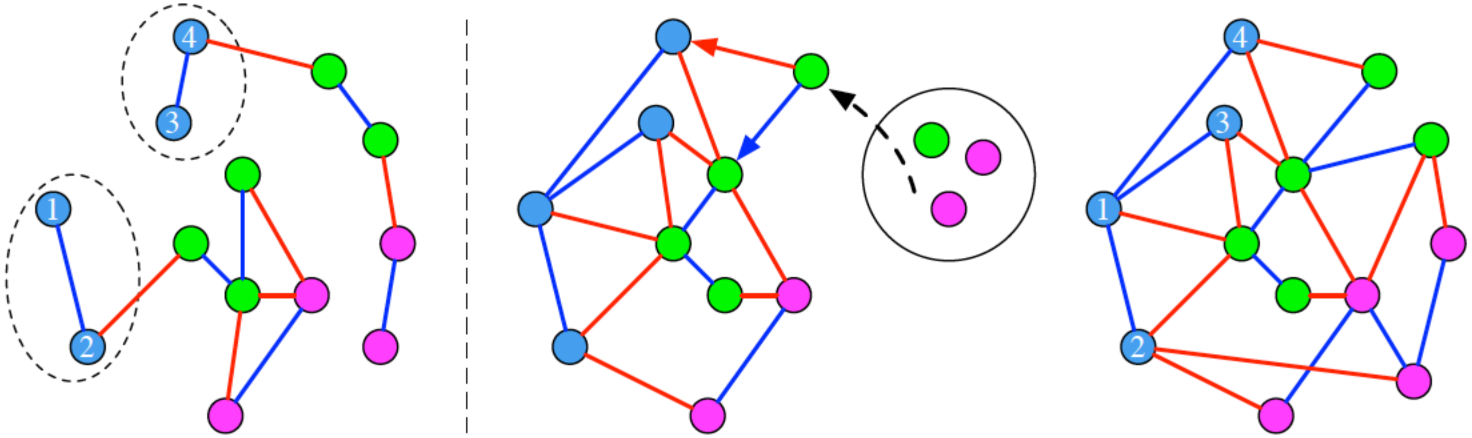
Self-supervised representations sensitive to the designed proxy task rather than the target task.



Mix-and-match tuning to narrow the gap.

Class-wise Connected Graph

香港中文大學
The Chinese University of Hong Kong



(a)
random triplets

(b-i)
construction of class-wise connected graph

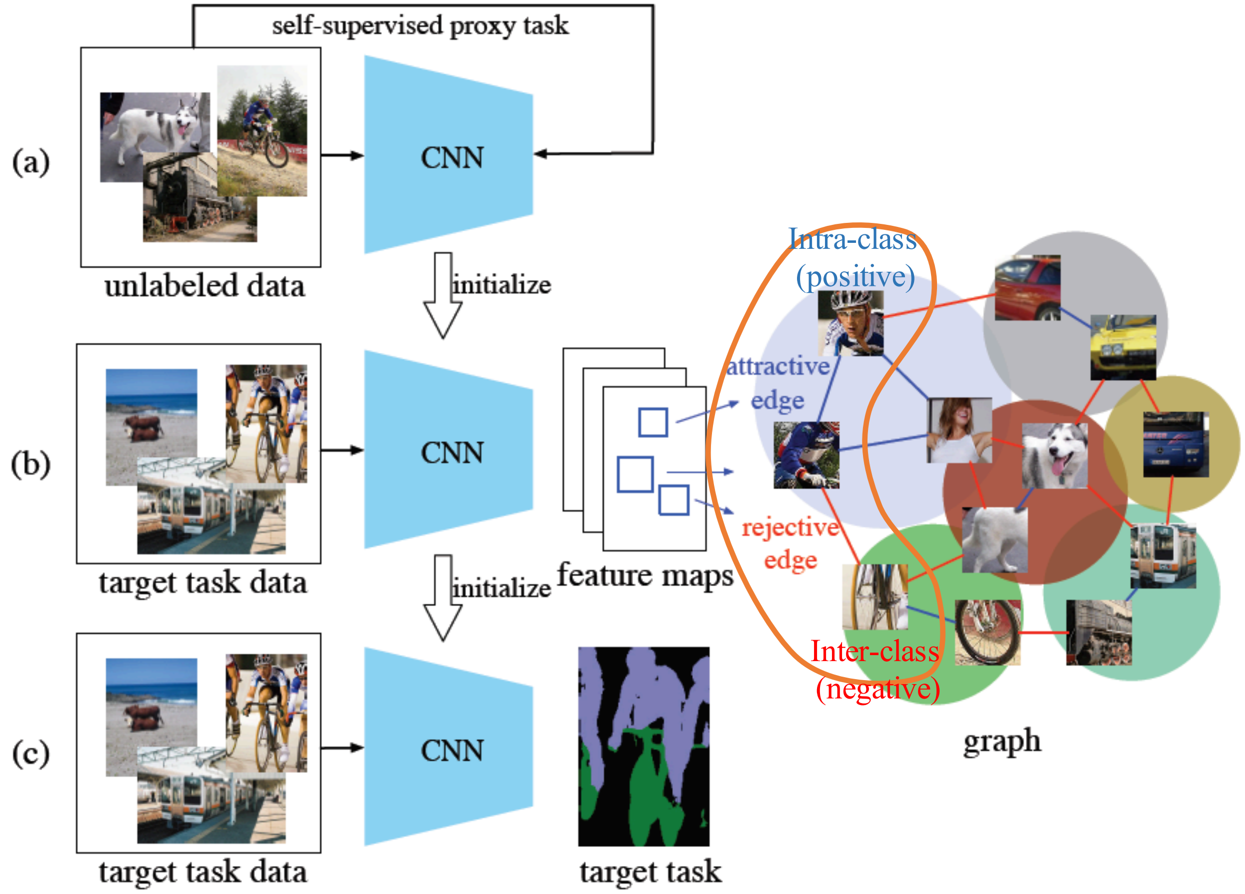
(b-ii)
the final graph

(a) Randomly selected triplets
i. May form multiple centers for each class
ii. Some nodes will never be used

(b) Our class-wise connected graph:
i. All nodes within the same class form a connected graph
ii. Each node can serve as an "anchor" node and to be used for optimization.

Training

香港中文大學
The Chinese University of Hong Kong



Formulate into triplet loss:

$$L = \frac{1}{N}\sum_{i}^{N} \max\left\{D\left(P_a^i, P_p^i\right) - D\left(P_a^i, P_n^i\right) + \alpha, 0\right\},$$

$$D(P_i, P_j) = \|(\mathbf{x}_i/\|\mathbf{x}_i\|_2 - \mathbf{x}_j/\|\mathbf{x}_j\|_2)\|^2,$$

For each node as an anchor (P_a), and randomly selected positive (P_p), randomly selected negative (P_n).
X_i : CNN representation on node i

Experiments

香港中文大學
The Chinese University of Hong Kong

Method	Arch.	VOC12 %mIoU.
ImageNet	VGG-16	64.2
Random	VGG-16	35.0
Larsson *et al.* (Larsson, Maire, and Shakhnarovich 2016)	VGG-16	50.2
Larsson *et al.* (Larsson, Maire, and Shakhnarovich 2017)	VGG-16	56.0
Ours (M&M + Graph, colorization pre-trained)	VGG-16	**64.5**
ImageNet	AlexNet	48.0
Random	AlexNet	23.5
k-means (Krähenbühl et al. 2015)	AlexNet	32.6
Pathak *et al.* (Pathak et al. 2016b)	AlexNet	29.7
Donahue *et al.* (Donahue, Krähenbühl, and Darrell 2016)	AlexNet	35.2
Zhang *et al.* (Zhang, Isola, and Efros 2016a)	AlexNet	35.6
Zhang *et al.* (Zhang, Isola, and Efros 2016b)	AlexNet	36.0
Noroozi et al. (Noroozi and Favaro 2016)	AlexNet	37.6
Larsson *et al.* (Larsson, Maire, and Shakhnarovich 2017)	AlexNet	38.4
Ours (M&M + Random Triplets, colorization pre-trained)	AlexNet	40.9
Ours (M&M + Graph, colorization pre-trained)	AlexNet	42.8
Ours (M&M + Graph, randomly initialized)	AlexNet	**43.6**

PASCAL VOC 2012 validation set

Experiments

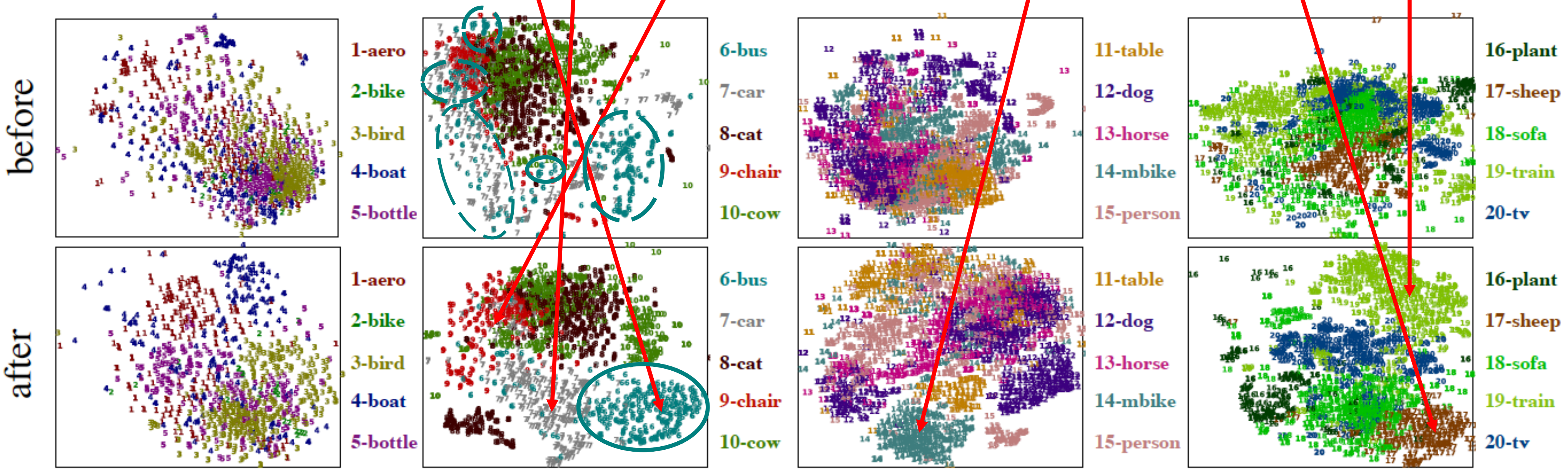
benchmark	PASCAL VOC2012					CityScapes	
pre-train	Random	Jigsaw	Colorize	Random	Colorize	Random	Colorize
backbone	AlexNet			VGG-16		VGG-16	
baseline	19.8	36.5	38.4	35.0	60.2	42.5	57.5
M&M	**43.6**	41.2	42.8	56.7	**64.5 (64.3)**	49.1	**66.4 (65.6)**
ImageNet	48.0			64.2		67.9	

Full results with different baselines and datasets

Experiments

香港中文大學
The Chinese University of Hong Kong

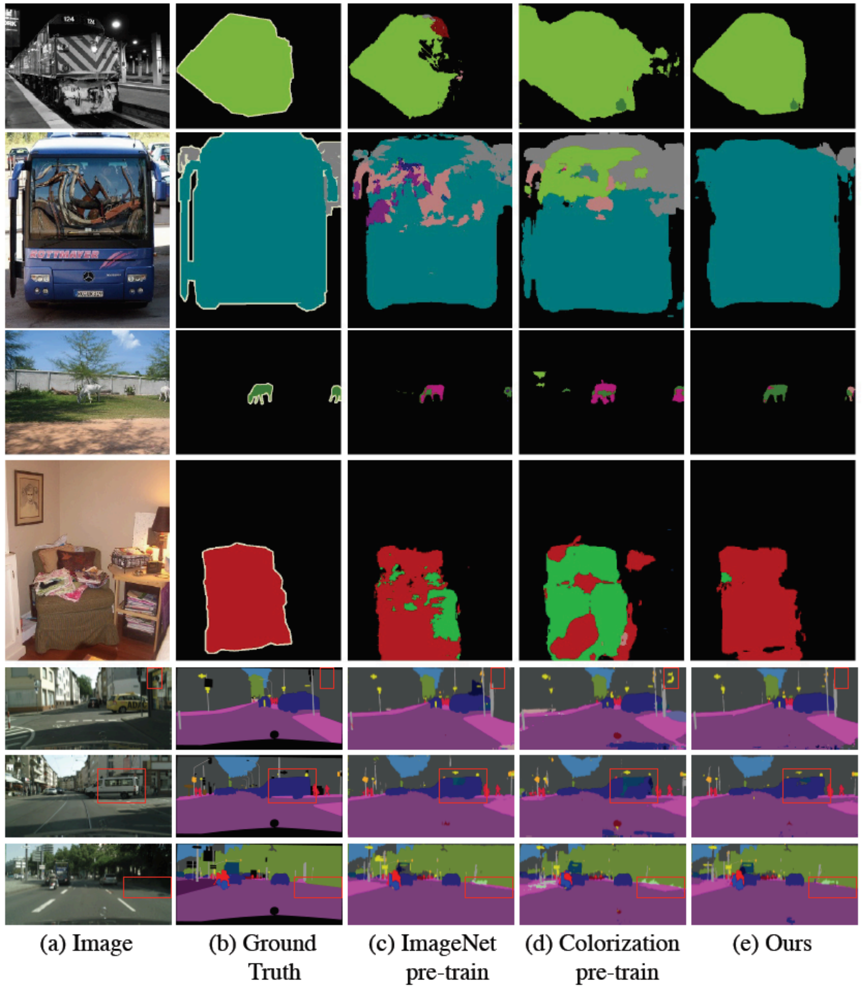
	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mIoU.
ImageNet	81.7	**37.4**	**73.3**	55.8	59.6	82.4	74.7	82.4	30.8	60.3	46.1	**71.4**	65.3	**72.6**	76.7	**49.7**	70.6	34.2	72.7	**60.2**	64.2
Colorization	73.6	28.5	67.5	55.5	50.2	78.3	66.1	78.3	26.8	60.8	**50.6**	70.6	64.9	62.2	73.5	38.2	66.8	**38.8**	68.1	55.1	60.2
M&M	**83.1**	37.0	69.6	**56.1**	**62.9**	**84.4**	**76.4**	**82.8**	**33.4**	**61.5**	44.7	67.3	**68.5**	68.0	**78.5**	42.2	**72.7**	37.2	**75.7**	58.6	**64.5**
Ensemble ImageNet+M&M	84.5	39.4	76.3	60.3	64.6	85.4	77.7	84.1	35.6	63.6	50.4	70.6	72.0	73.6	80.1	50.2	73.7	37.6	77.8	66.6	67.4



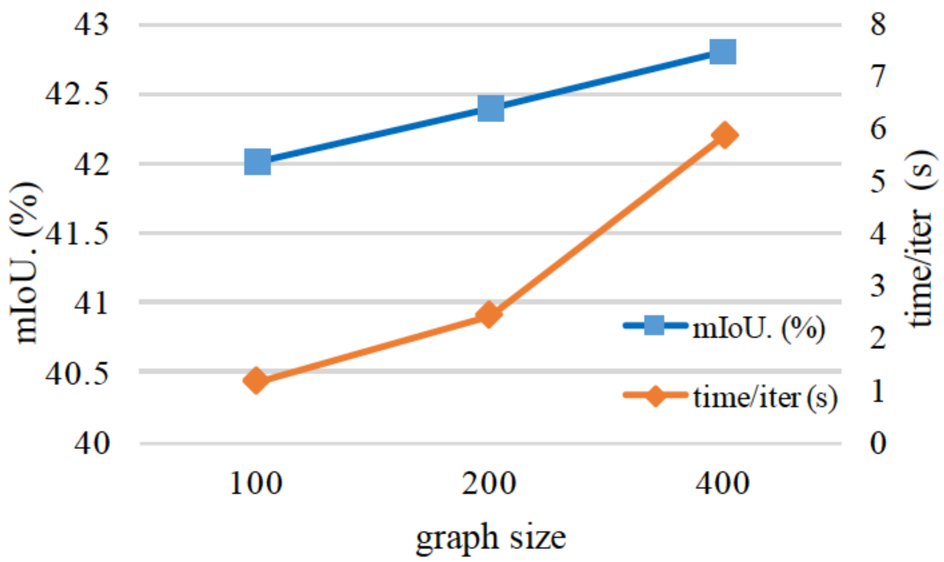
Per-class results and T-SNE feature visualization

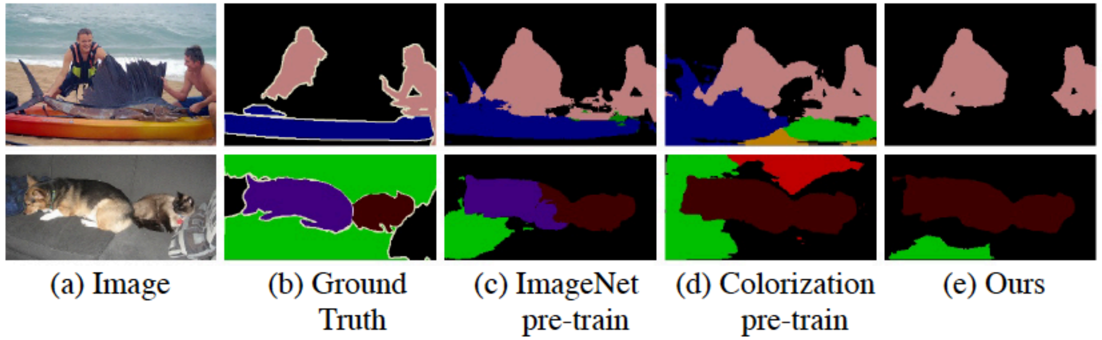
Experiments

香港中文大學
The Chinese University of Hong Kong



Visualization results





Failure cases

Thank You!