SlotCon: Self-Supervised Visual Representation Learning with Semantic Grouping

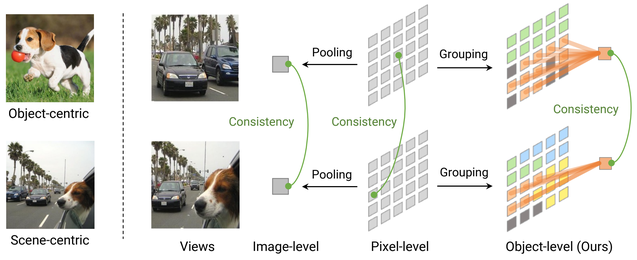
Xin Wen[1] Bingchen Zhao[2] Anlin Zheng[1,3] Xiangyu Zhang[3] Xiaojuan Qi[1]

[1]The University of Hong Kong [2]University of Edinburgh [3]MEGVII Technology

Code

Twitter Thread

Scene-centric data are easier to collect, and contain more information per image.

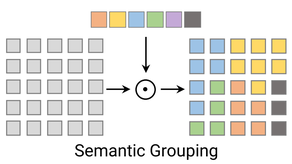


But effective pre-training on them requires more than instance discrimination. We endorse **object-level contrastive learning**.

Then, how to find the objects w/o supervision?

Hand-crafted object-proposal methods like saliency, selective-search, k-means clustering may limit the upper bound of learned representations.

*We propose **learnable semantic grouping** for online object discovery.*



- First learn semantic prototypes from the whole dataset, where each prototype can represent a semantic class (e.g., cat, dog).
- Then assign a nearest-neighbor prototype to each pixel.
- Pixels with the same pseudo-label forms a group (object).

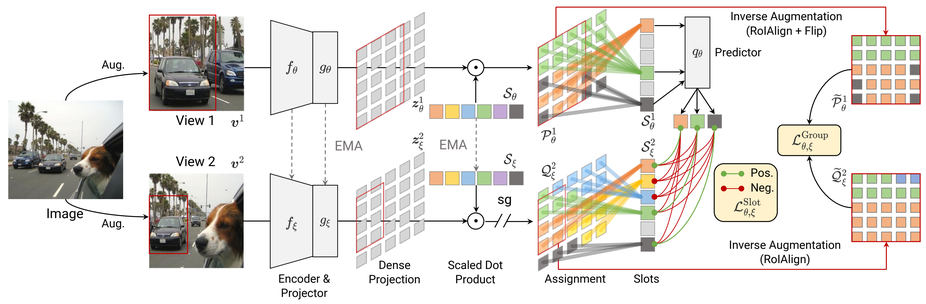
Priors for objectness

- Prior 1: geometric-covariance and photometric-invariance
- Prior 2: small prototype number
- Prior 3: meaningful grouping. i.e., avoiding collapse

Table 1: Details of the datasets used for pre-training.

Dateset	#Img.	#Obj./Img.	#Class
ImageNet-1K [15]	1.28M	1.7	1000
COCO [48]	118K	7.3	80
COCO+ [48]	241K	N/A	N/A

Solving object discovery & representation learning jointly



Online Object discovery

- Rand. init. the prototypes
- Train w/ pixel-level deep clustering
- Pixels w/ same assign. form an object

Object-level contrastive learning

- Pool obj-level rep.(slots) from each view
- Slots w/o assigned pixels are filtered out
- Contrastive learning between slots

New SOTA on scene-centric pre-training

Table 2: **Main transfer results with COCO pre-training.** We report the results in COCO [48] object detection, COCO instance segmentation, and semantic segmentation in Cityscapes [13], PASCAL VOC [20] and ADE20K [86]. Compared with other image-, pixel-, and object-level self-supervised learning methods, our method shows consistent improvements over different tasks without leveraging multi-crop [6] and objectness priors. (†: re-impl. w/ official weights; ‡: full re-impl.)

Method	Epochs	Multi crop	Obj. Prior	COCO detection			COCO segmentation			Semantic seg. (mIoU)		
				AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}	City.	VOC	ADE
random init.	-	✗	✗	32.8	50.9	35.3	29.9	47.9	32.0	65.3	39.5	29.4
Image-level approaches												
MoCo v2‡ [9]	800	✗	✗	38.5	58.1	42.1	34.8	55.3	37.3	73.8	69.2	36.2
Revisit.† [63]	800	✓	✗	40.1	60.2	43.6	36.3	57.3	38.9	75.3	70.6	37.0
Pixel-level approaches												
Self-EMD [49]	800	✗	✗	39.3	60.1	42.8	-	-	-	-	-	-
DenseCL† [68]	800	✗	✗	39.6	59.3	43.3	35.7	56.5	38.4	75.8	71.6	37.1
PixPro‡ [75]	800	✗	✗	40.5	60.5	44.0	36.6	57.8	39.0	75.2	**72.0**	38.3
Object / Group-level approaches												
DetCon† [35]	1000	✗	✓	39.8	59.5	43.5	35.9	56.4	38.7	76.1	70.2	38.1
ORL† [74]	800	✓	✓	40.3	60.2	44.4	36.3	57.3	38.9	75.6	70.9	36.7
Ours (SlotCon)	800	✗	✗	**41.0**	**61.1**	**45.0**	**37.0**	**58.3**	**39.8**	**76.2**	71.6	**39.0**

Eqv. Performance to ImageNet with only ½ data

Table 3: **Pushing the limit of scene-centric pre-training.** Our method further sees a notable gain in all tasks with extended COCO+ data, showing the great potential of scene-centric pre-training.

Method	Dataset	Epochs	COCO detection			COCO segmentation			Semantic seg. (mIoU)		
			AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}	City.	VOC	ADE
SlotCon	COCO	800	41.0	61.1	45.0	37.0	58.3	39.8	76.2	71.6	39.0
SlotCon	ImageNet	100	41.4	61.6	45.6	37.2	58.5	39.9	75.4	73.1	38.6
SlotCon	ImageNet	200	**41.8**	**62.2**	45.7	**37.8**	59.1	**40.7**	76.3	**75.0**	38.8
ORL [74]	COCO+	800	40.6	60.8	44.5	36.7	57.9	39.3	-	-	-
SlotCon	COCO+	800	**41.8**	**62.2**	**45.8**	**37.8**	**59.4**	40.6	**76.5**	73.9	**39.2**

Evaluation on unsupervised semantic segmentation

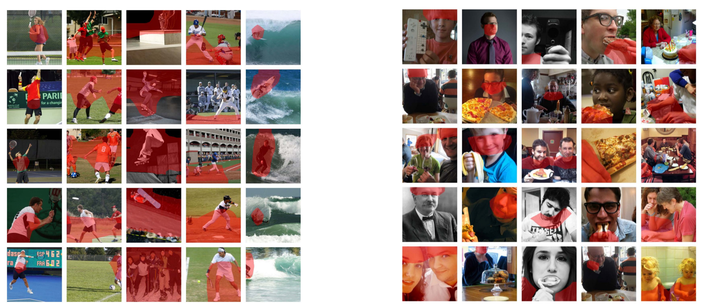
Table 5: Main results in COCO-Stuff unsupervised semantic segmentation.

Method	mIoU	pAcc
MaskContrast [64]	8.86	23.03
PiCIE + H. [40]	14.36	49.99
SegDiscover [38]	14.34	**56.53**
Ours (SlotCon)	**18.26**	42.36

The model managed to discover visual concepts

Figure 3: **Examples of visual concepts discovered by SlotCon from the COCO `val2017` split.** Each column shows the top 5 segments retrieved with the same prototype, marked with reddish masks or arrows. Our method can discover visual concepts across various scenarios and semantic granularities regardless of small object size and occlusion. (*best viewed in color*)

Though it can be biased toward occupying classes, e.g., persons



Set the number of protos close to real class number

K	COCO	City	VOC	ADE
128	40.7	**76.4**	**71.9**	38.5
256	**41.0**	76.2	71.6	39.0
512	40.9	75.6	71.6	38.9
1024	40.7	75.8	70.9	**39.1**

Geometric augs are crucial for sem. grouping

Method	Geometric aug.	VOC mIoU
Random init.	-	39.5
SlotCon	✓	**71.6**
SlotCon	✗	62.6