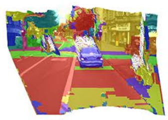
Copy-Pasting Coherent Depth Regions Improves Contrastive Learning for Urban-Scene Segmentation

Liang Zeng   Attila Lengyel   Nergis Tömen   Jan van Gemert

Delft University of Technology

1. Why contrastive learning on urban scene?

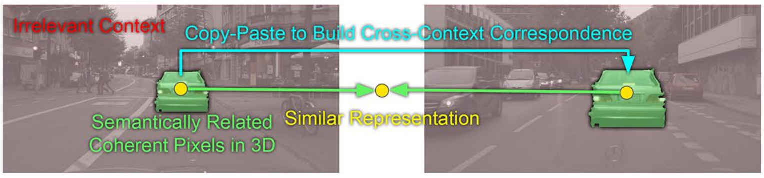
1. applying contrastive learning to complex, non-object-centric urban scenes segmentation is a non-trivial and often overlooked research topic
2. self-supervised depth estimation on urban scenes is well-addressed in literature
3. semantic relatedness of pixels correlates with their coherence in 3D space



grouping coherent, semantically related pixels into coherent depth regions

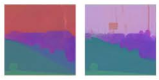
2. What would be good(robust) pixel representations?

1. dependent on related pixels, e.g., a pixel is classified as “car” together with other pixels of the car
2. invariant to irrelevant pixels(contexts), e.g., the “car” pixels representation is constant no matter where the car appears
3. using copy-paste to simulate the scenarios and using contrastive learning to learn such constraints



cars in different context

w/o copy-paste

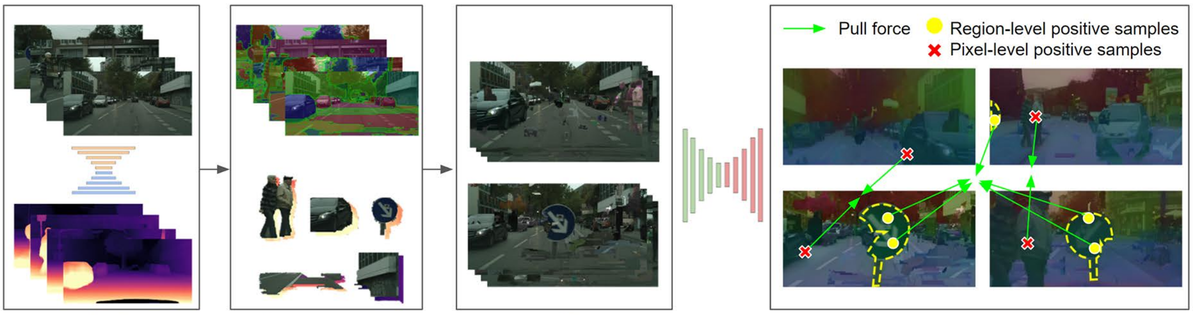


with copy-paste

copy-paste is vital for learning object-specific representations invariant across different contexts

3. Method

Region-level: pixels from identical region under transformations
Pixel-level: identical pixels under transformations
Loss = λ $Loss_{pixel}$ + (1 − λ)$Loss_{region}$



Depth Estimation   Pixel Grouping   Copy-Paste   Contrastive Learning

Our method consists of four steps. 1. Training a *depth estimator* on video clips by *self-supervision*. 2. Grouping pixels coherent in 3D space given the depth by a *heuristic algorithm*. 3. Building cross-context correspondences by *copy-paste*. 4. Pulling together the representations of corresponding pixels and regions using *SwAV contrastive learning framework*.

4. Experiment

Dataset: Cityscapes and KITTI
Platform: one 16GB V100 GPU

4.1 Unsupervised semantic segmentation(clustering)

For unsupervised semantic segmentation on Cityscapes and KITTI, our method surpasses the previous state-of-the-art baseline by +7.14% in mIoU on Cityscapes and +6.65% on KITTI

Method	Init.	Training Data	CS-Sem.		KT-Sem.	
			Acc	*mIoU*	*Acc*	*mIoU*
PiCIE	scratch	Cityscapes	33.56	8.33	32.20	6.52
Ours($\lambda = 0.5$)	scratch	Cityscapes	65.42	20.49	**68.37**	21.03
PiCIE	scratch	KITTI	30.28	6.81	30.62	7.54
Ours($\lambda = 0.5$)	scratch	KITTI	49.18	17.20	49.58	18.22
PiCIE	ImageNet	Cityscapes	**68.50**	16.24	56.74	13.54
Ours($\lambda = 0.5$)	ImageNet	Cityscapes	66.70	**23.38**	68.25	**22.50**
PiCIE	ImageNet	KITTI	53.24	12.55	61.74	12.92
Ours($\lambda = 0.5$)	ImageNet	KITTI	56.96	18.85	59.11	19.57

Table 1: Unsupervised semantic segmentation performance on Cityscapes *val* set and *train* set. We retrained PiCIE with equivalent setting to ours.

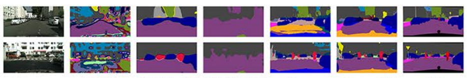
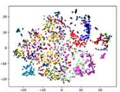
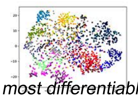


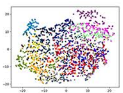
Image   PiCIE   PiCIE-IN 5 epochs   PiCIE-IN 20 epochs   Ours   Ours-IN   GT



$\lambda = 0$



most differentiable
$\lambda = 0.5$



$\lambda = 1$

Feature distribution by t-sne w.r.t class label

4.2 Semi-supervised segmentation(fine-tuning)

For semi-supervised semantic and instance segmentation on Cityscapes and KITTI, our method is competitive with recent method pre-trained on the larger ImageNet and COCO using more GPUs by their authors. Training on Cityscapes and KITTI in same condition, our method surpasses SwAV and PixPro

Pre-training Method	Pre-training Dataset	CS-Sem.	CS-Inst.		KT-Sem.	KT-Inst.	
		mIoU	*AP*	AP_{50}	*mIoU*	*AP*	AP_{50}
scratch	-	65.11	24.30	46.98	32.99	8.15	15.79
supervised	ImageNet	70.54	27.34	50.59	40.09	12.38	23.42
SwAV	ImageNet	71.07	28.08	52.25	40.52	**13.78**	**27.90**
DenseCL	ImageNet	72.09	28.97	51.93	40.88	12.63	22.74
PixPro	ImageNet	72.66	29.04	52.59	40.50	13.04	24.95
ORL	COCO	72.32	**29.94**	52.55	41.88	12.02	23.48
CAST	COCO	69.92	27.33	51.31	38.78	10.67	20.13
SwAV	Cityscapes	61.69	23.62	46.21	36.10	9.22	18.01
PixPro	Cityscapes	61.64	23.78	46.45	36.99	9.61	18.73
SwAV	KITTI	60.74	23.51	46.08	36.90	9.42	18.11
PixPro	KITTI	61.25	23.23	46.23	37.28	9.45	18.57
Ours($\lambda = 1$)	Cityscapes	**73.55**	**29.94**	**52.88**	**42.70**	12.58	24.98
Ours($\lambda = 0.5$)	Cityscapes	73.03	29.11	51.87	42.32	12.22	23.16
Ours($\lambda = 1$)	KITTI	71.62	28.77	52.71	41.17	11.74	20.57
Ours($\lambda = 0.5$)	KITTI	71.45	27.86	51.16	41.03	11.36	20.49

Table 2: Segmentation performance over Cityscapes *val* set and 5-fold validation of KITTI *train* set. SwAV and PixPro on Cityscapes and KITTI are trained with limited GPU compute as ours.

5. Discussion

1. investigating the transferrability on datasets other than urban scenes
2. exploring data-driven method instead of heuristic pixel grouping algorithm