PooDLe: Pooled and dense self-supervised learning from naturalistic videos

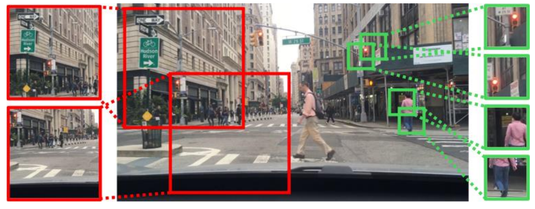
NYU agentic learning ai lab

Alex N. Wang*, Chris Hoang*, Yuwen Xiong, Yann LeCun, Mengye Ren

https://agenticlearning.ai/poodle

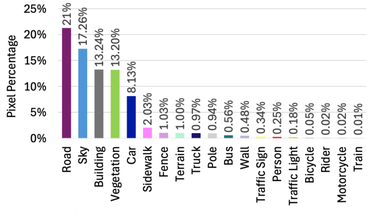
Meta

Challenges of naturalistic data



Dense scene from BDD100K with typical **global crops** versus PooDLe's **subcrops**

- Existing self-supervised learning (SSL) methods often rely on clear, single-subject ("iconic") examples like ImageNet.
- These methods struggle on naturalistic data with dense scenes, imbalanced classes, and objects of varying size.
- Dense SSL methods prioritizes large classes, ignoring important classes like "Person" which occupies <0.3% of pixels in BDD100K.

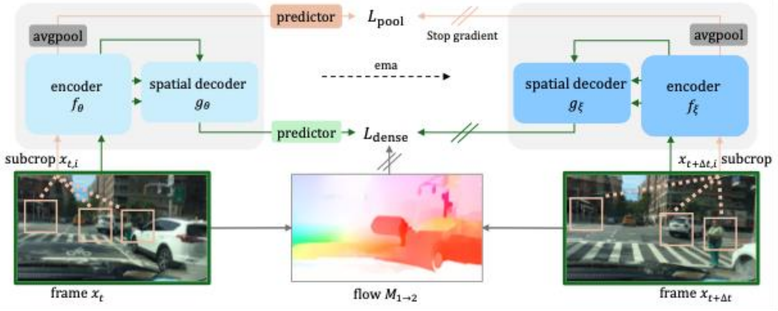


Pixel frequency of each class in BDD100K

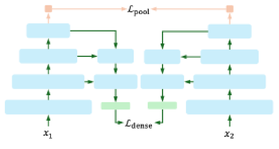
Contributions

- Combine SSL training on a dense, flow-equivariance objective on larger crops with a pooled, invariance objective on small, iconic-like subcrops via a spatial decoder module (SDM).
- Achieve state-of-the-art results on semantic segmentation and object detection pretraining on BDD100K and Walking Tours.
- Study the effects of crop area and input resolution for dense and pooled objectives, and temporal stride for paired frame sampling.

Pooled and dense SSL architecture



- PooDLe unifies two joint embedding objectives, L_{pool} and L_{dense}, with a spatial decoder architecture.
- L_{pool} is an invariance objective on pseudo-iconic subcrops and directly trains the encoder features.
- The UNet-inspired SDM upsamples top-level encoder features for L_{dense}, which is a flow-equivariance objective on global crops.
- Using lateral connections in the SDM to incorporate low-level features improves small class understanding.



SDM architecture

Variant	Dense	Pool	Top-Down	Lateral	All	Small	Large
FlowE	✓				28.8	8.7	40.5
	✓	✓			28.9	7.2	41.6
	✓	✓			30.3	6.8	44.0
	✓	✓		✓	30.3	10.9	41.7
	✓		✓	✓	31.8	12.8	42.8
PooDLe	✓	✓	✓	✓	34.2	15.0	45.5

Ablations on losses and architecture. PooDLe performs best on mIoU across the "All", "Small", "Large" class breakdowns

Semantic segmentation results

- We pretrain PooDLe and baselines on BDD100K and WalkingTours (WT)
- Models are evaluated on in-distribution tasks as well as Cityscapes

		BDD100K				Cityscapes			
		Linear		UperNet		Linear		UperNet	
Method	Arch	mIoU	Acc	mIoU	Acc	mIoU	Acc	mIoU	Acc
Scratch	R50	9.7	55.0	26.1	81.2	9.8	58.0	30.7	84.1
DINO	ViT-S	29.6	86.8	41.1	90.1	35.1	87.9	51.5	91.9
iBOT	ViT-S	27.2	85.4	35.5	88.7	32.0	86.2	44.0	90.3
DoRA	ViT-S	33.2	88.1	43.3	90.7	37.4	88.7	50.8	92.0
DINO	R50	13.1	64.7	25.6	80.3	14.9	69.4	29.2	81.4
DenseCL	R50	24.2	84.9	41.8	90.0	26.6	85.6	53.2	91.9
FlowE	R50	35.7	88.5	47.3	91.5	43.1	89.5	57.7	93.1
PooDLe	R50	**39.2**	**89.2**	**49.9**	**91.8**	**47.2**	**90.2**	**60.7**	**93.5**
Supervised	R50	36.7	84.7	***55.2***	92.0	46.8	87.4	63.4	93.7
PooDLe*	R50	**44.7**	**90.7**	54.1	**92.7**	***52.0***	**91.5**	**65.1**	**94.4**

Experiments pretrained on BDD100K. *init with supervised IN1K weights

- Models are also evaluated on BDD100K with classes separated by size and rarity

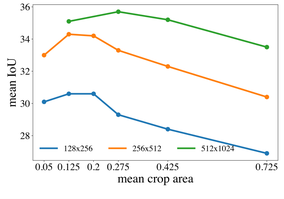
Method	Data	All	Small	Large	Rare	Common
DINO	BDD	29.6	8.4	42.0	1.0	42.7
DenseCL	BDD	24.2	1.6	37.4	0.0	35.4
DoRA	BDD	33.2	11.9	45.6	2.8	47.3
FlowE	BDD	35.7	12.2	49.3	10.7	47.2
PooDLe	BDD	39.2	18.3	51.4	12.0	51.8
Supervised	IN1K	36.7	27.2	42.2	26.6	46.2
PooDLe	BDD*	44.7	25.2	56.1	43.1	57.1

Semantic Segmentation mIoU on BDD100K classes separated by small-large, rare-common classes

Study of crop area, resolution, temporal stride

	classes		
Subcrop area	small	large	all
0.04- 0.18	16.0	41.6	34.6
0.18- 0.36	14.2	41.2	33.7
0.36 - 0.54	13.6	39.5	32.4
0.54 - 1.00	12.3	38.8	31.5

*Subcrop area is a fraction of global crops



- Smaller subcrops are better as pseudo-iconic views, especially for smaller classes
- Larger input resolutions peak for larger crop sizes