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

Alex N. Wang[1,*], Christopher Hoang[1,*], Yuwen Xiong, Yann LeCun[1,2], Mengye Ren[1]

[1] New York University [2] Meta

Iconic image data as implicit supervision

ImageNet (iconic)

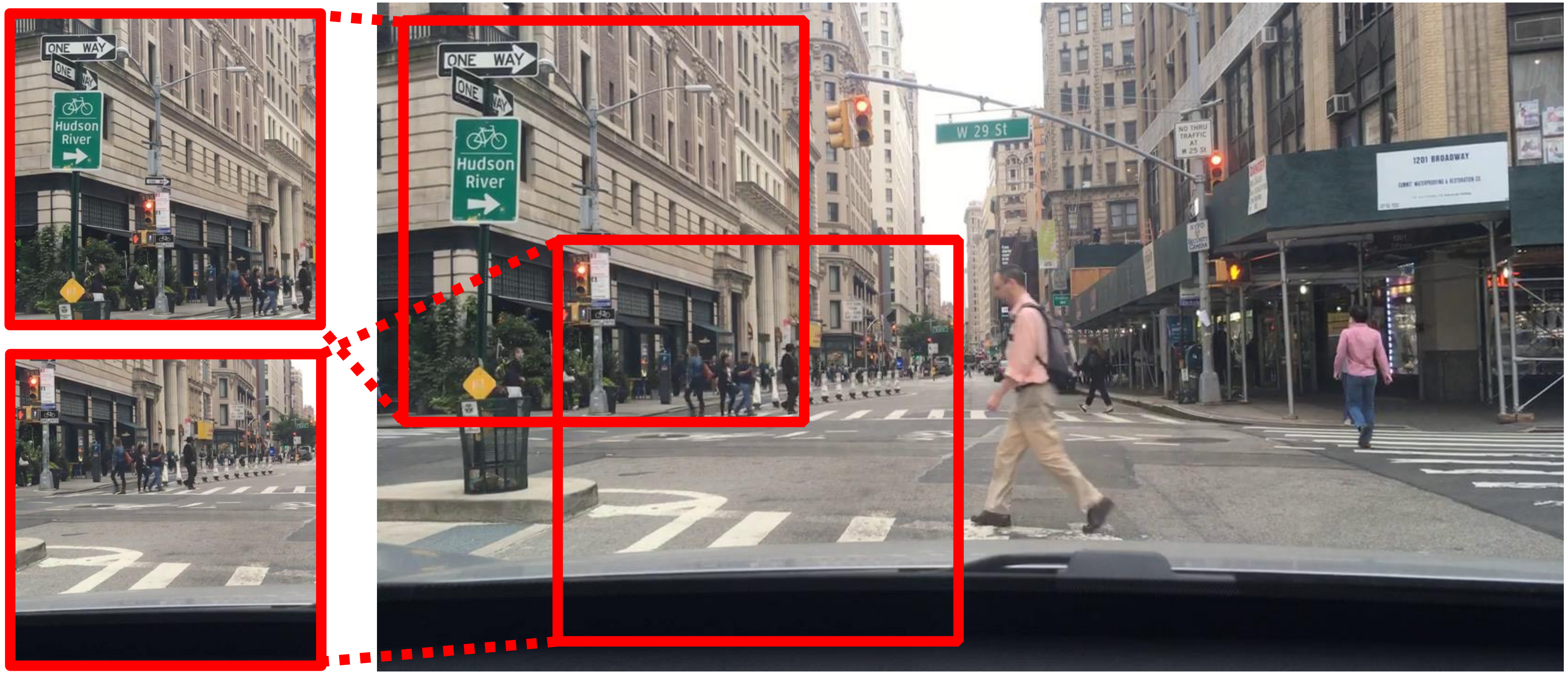
Iconic image data as implicit supervision

ImageNet (iconic)

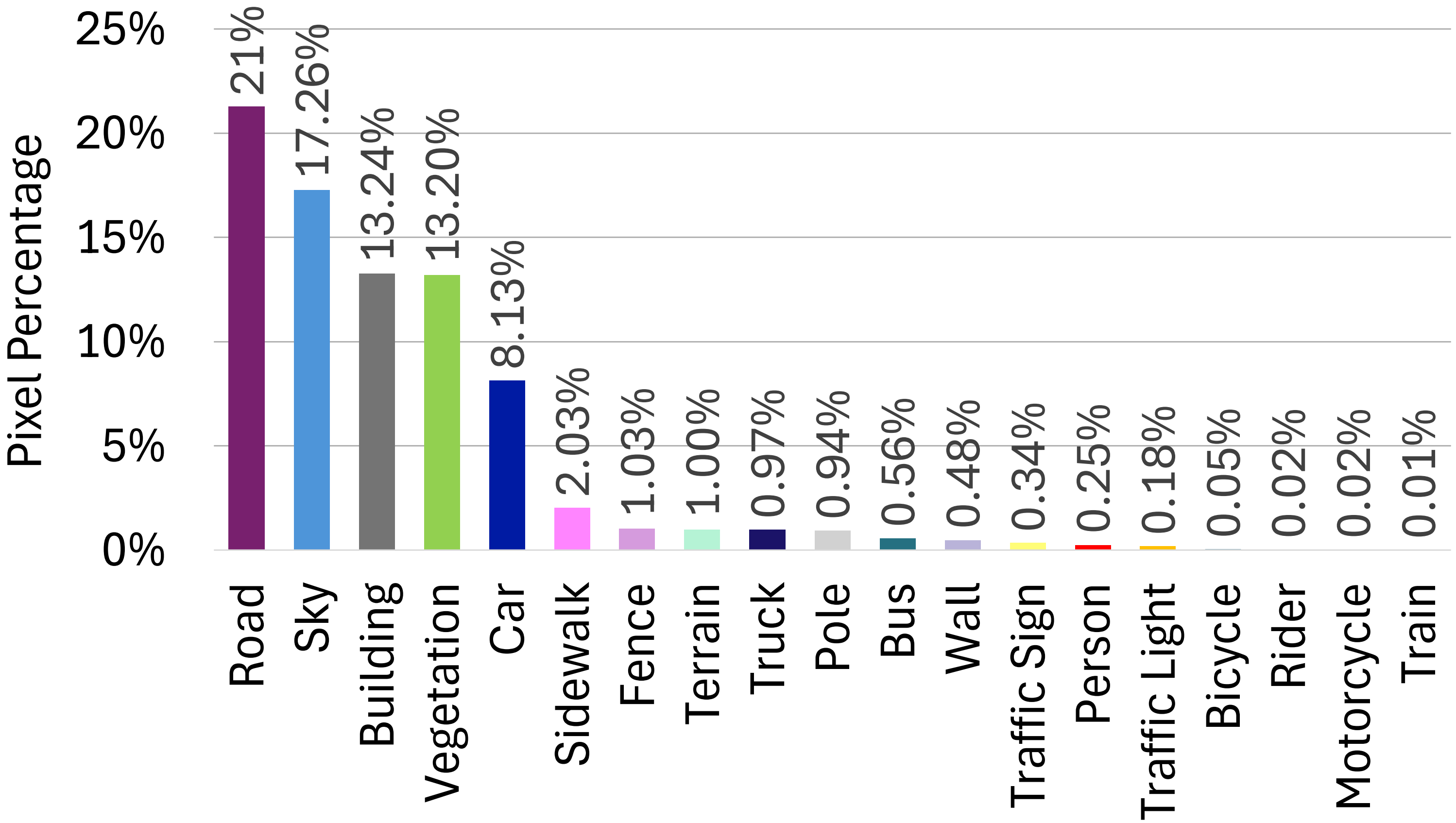
BDD100K (naturalistic, dense)

Learning from dense scenes

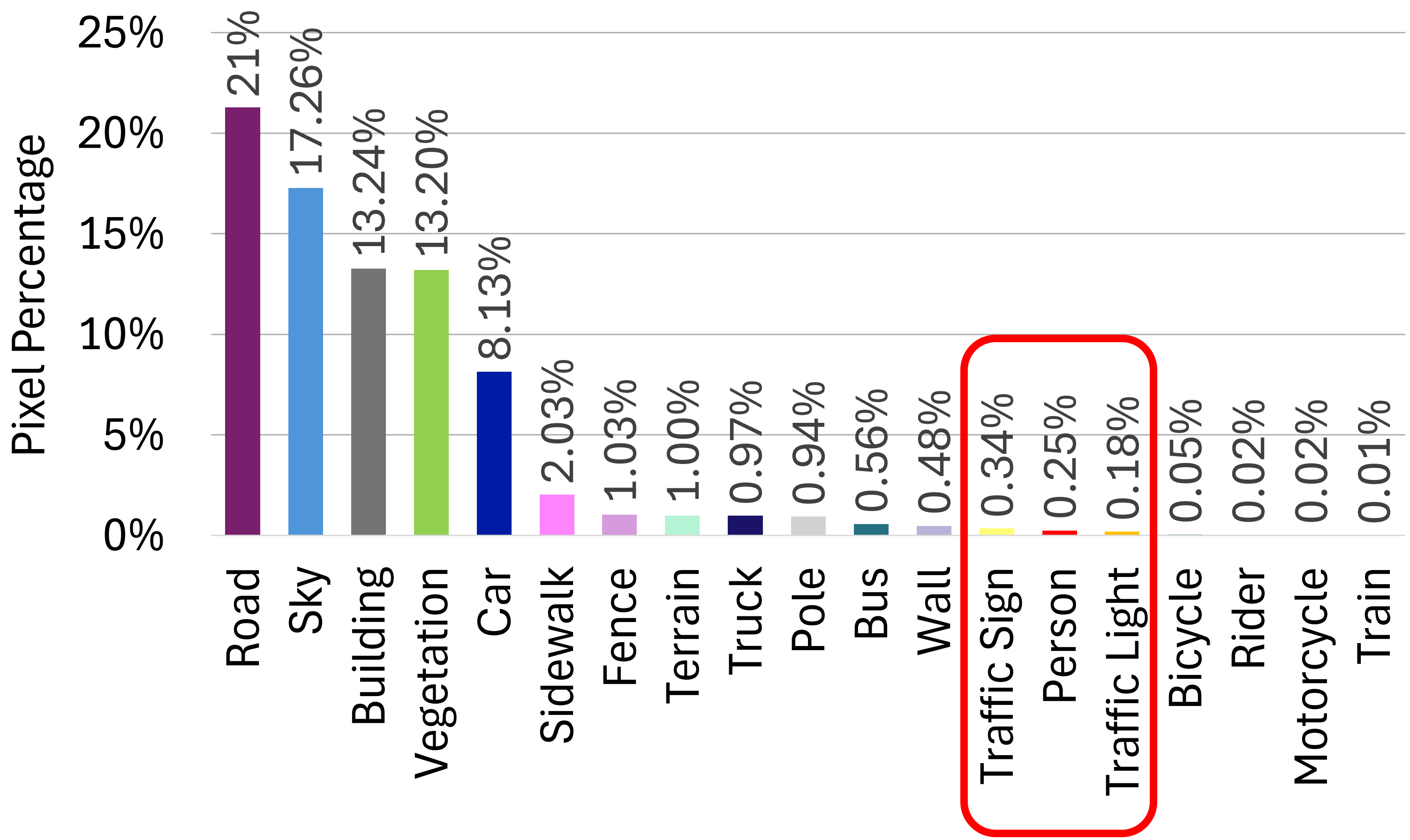
Learning from dense scenes



Size imbalance in dense, naturalistic data

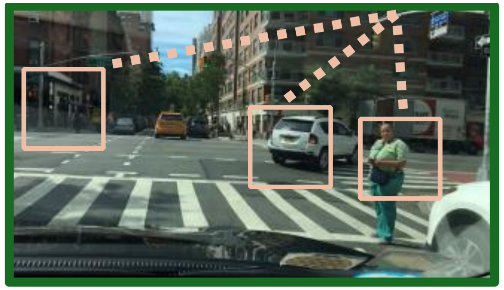


Size imbalance in dense, naturalistic data



Pooled and Dense Self-supervised Learning

frame x_t

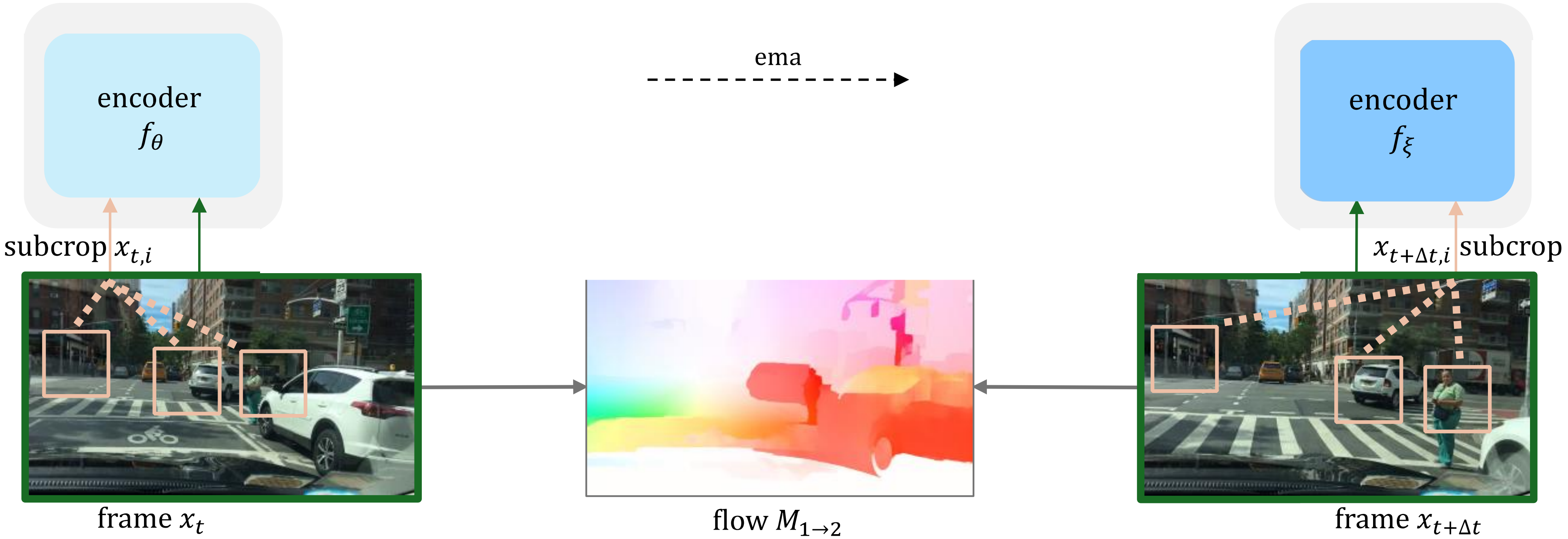


frame $x_{t+\Delta t}$

Method

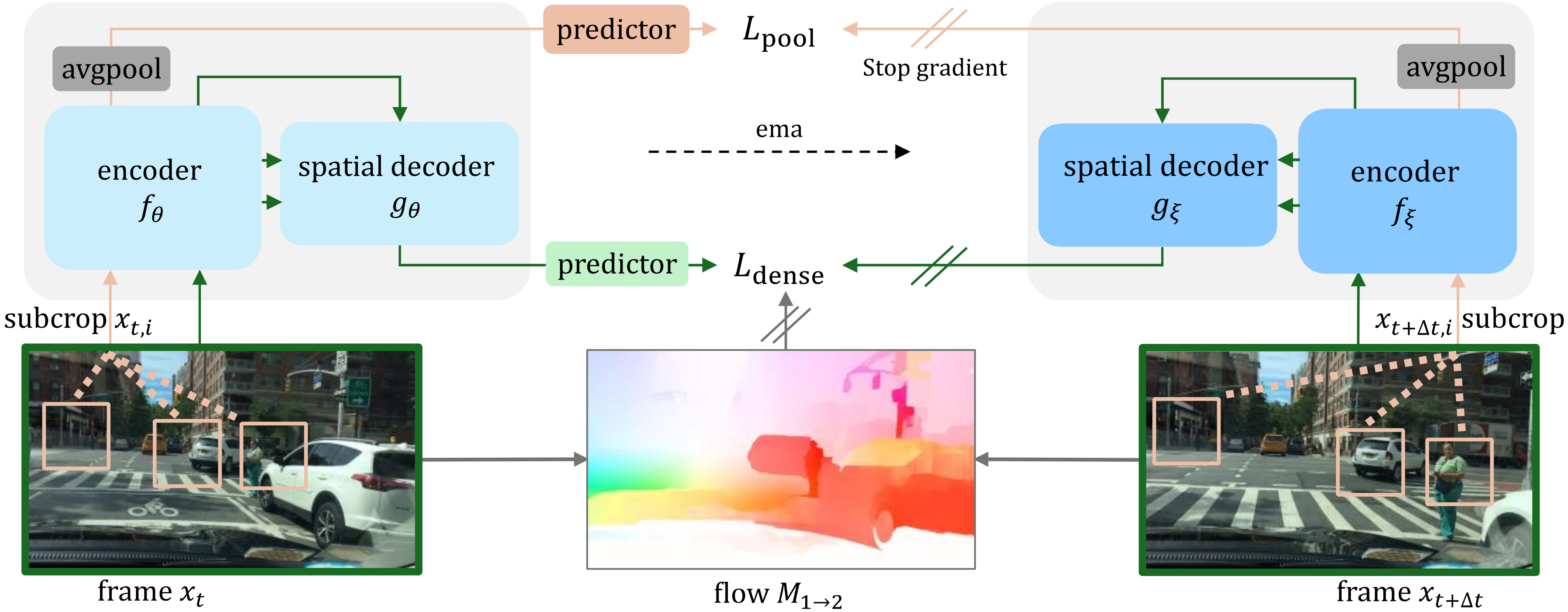
frame x_t flow $M_{1\rightarrow 2}$ frame $x_{t+\Delta t}$

Method



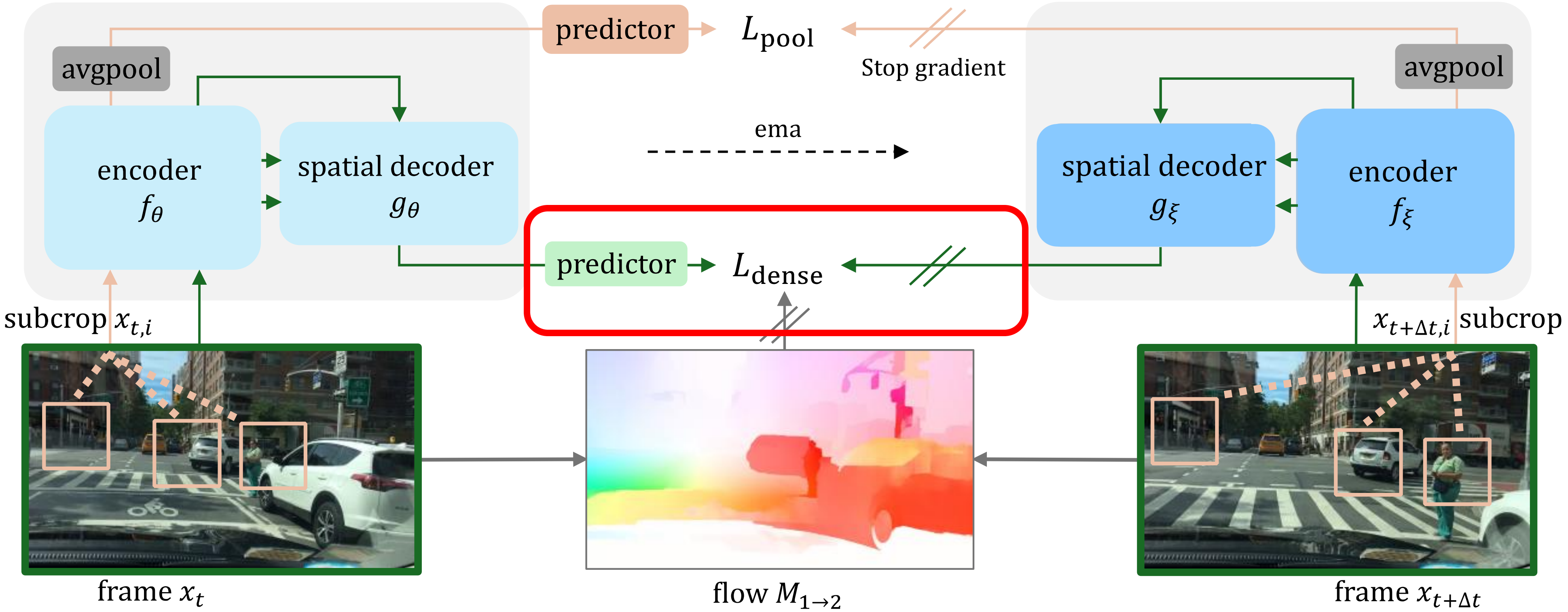
encoder
f_θ
ema
encoder
f_ξ
subcrop $x_{t,i}$
$x_{t+\Delta t,i}$ subcrop
frame x_t
flow $M_{1\to2}$
frame $x_{t+\Delta t}$

Method



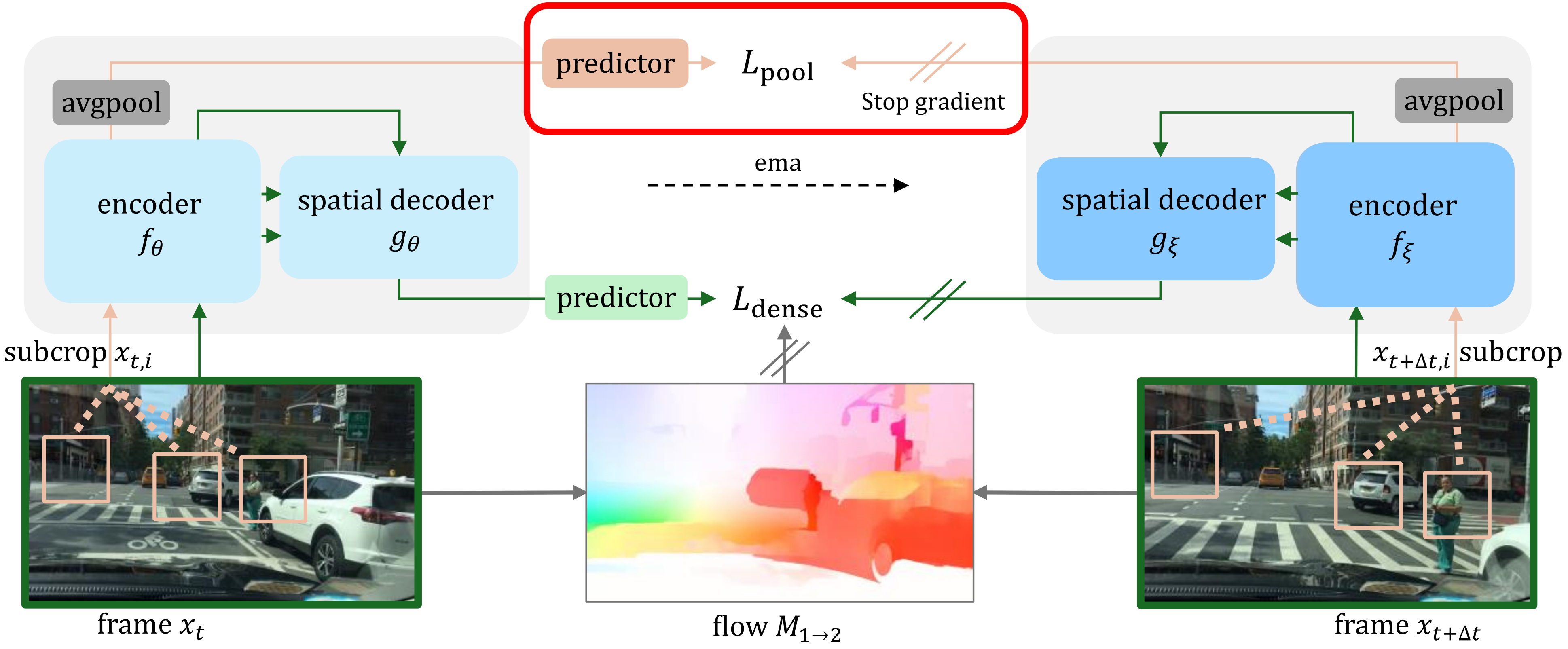
predictor
L_{pool}
Stop gradient
avgpool
avgpool
encoder f_θ
spatial decoder g_θ
ema
spatial decoder g_ξ
encoder f_ξ
predictor
L_{dense}
subcrop $x_{t,i}$
$x_{t+\Delta t,i}$ subcrop
frame x_t
flow $M_{1\to 2}$
frame $x_{t+\Delta t}$

Method

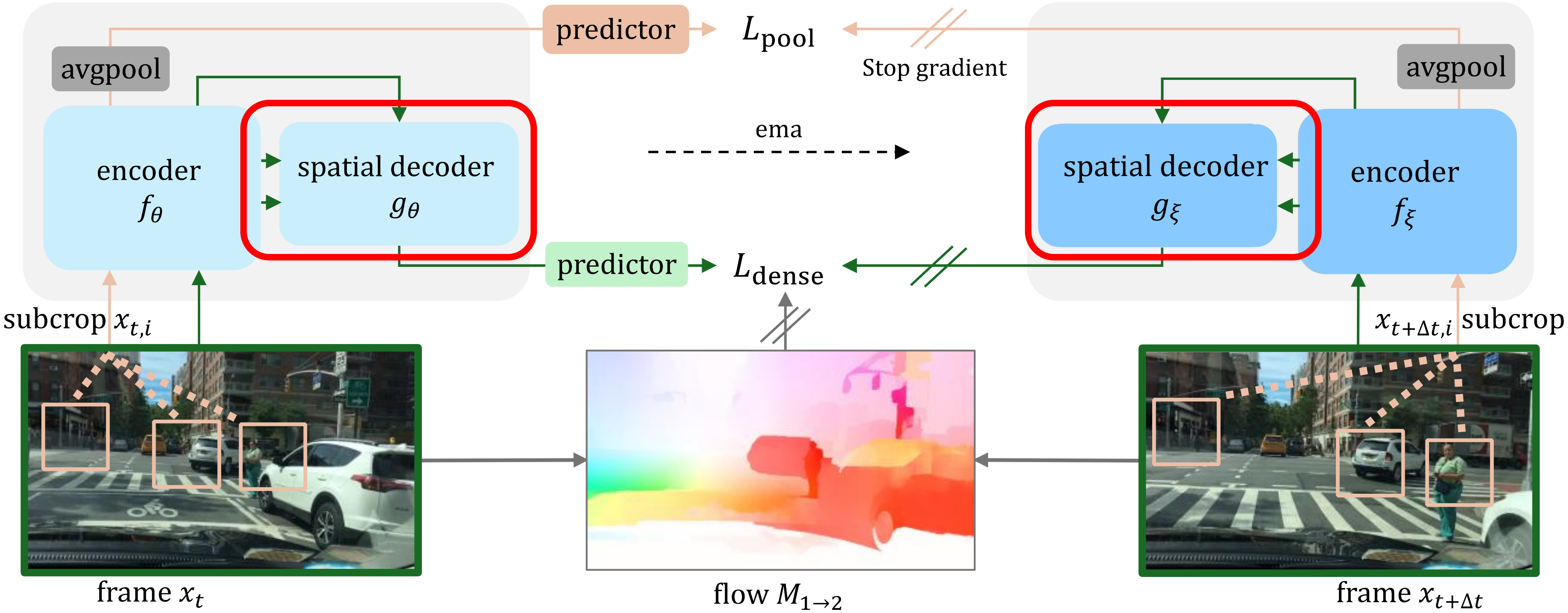


predictor
L_{pool}
Stop gradient
avgpool
avgpool
ema
encoder f_θ
spatial decoder g_θ
spatial decoder g_ξ
encoder f_ξ
predictor
L_{dense}
subcrop $x_{t,i}$
$x_{t+\Delta t,i}$ subcrop
frame x_t
flow $M_{1\to 2}$
frame $x_{t+\Delta t}$

Method

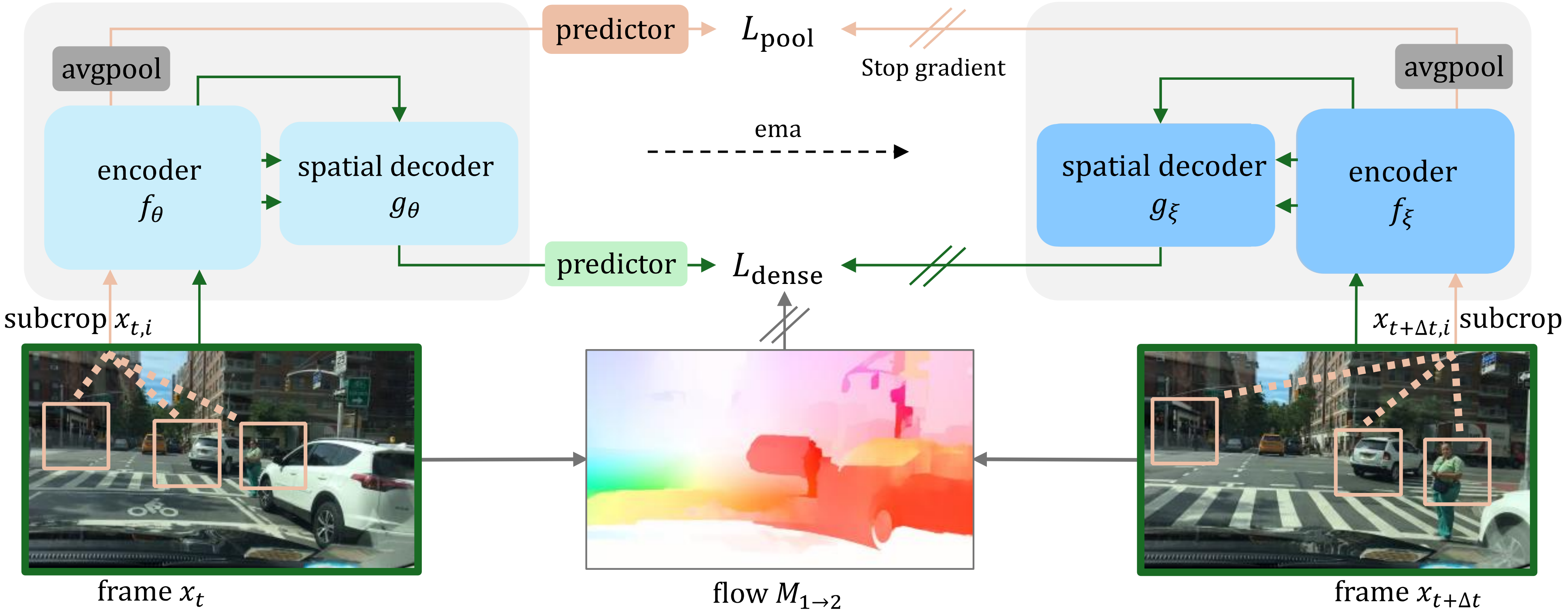

predictor
L_{pool}
Stop gradient
avgpool
avgpool
encoder f_θ
spatial decoder g_θ
ema
spatial decoder g_ξ
encoder f_ξ
predictor
L_{dense}
subcrop $x_{t,i}$
$x_{t+\Delta t,i}$ subcrop
frame x_t
flow $M_{1\to 2}$
frame $x_{t+\Delta t}$

Method



predictor
L_{pool}
Stop gradient
avgpool
avgpool
encoder f_θ
spatial decoder g_θ
ema
spatial decoder g_ξ
encoder f_ξ
predictor
L_{dense}
subcrop $x_{t,i}$
$x_{t+\Delta t,i}$ subcrop
frame x_t
flow $M_{1\to 2}$
frame $x_{t+\Delta t}$

Method



predictor
L_{pool}
Stop gradient
avgpool
avgpool
encoder
f_θ
spatial decoder
g_θ
ema
spatial decoder
g_ξ
encoder
f_ξ
predictor
L_{dense}
subcrop $x_{t,i}$
$x_{t+\Delta t,i}$ subcrop
frame x_t
flow $M_{1\to 2}$
frame $x_{t+\Delta t}$

Results on BDD100K

Method	Arch	Ep.	Pretrain	BDD100K Sem. Seg.			
				Linear		UperNet	
				mIoU	Acc	mIoU	Acc
Scratch	R50	-	-	9.7	55.0	26.1	81.2
DINO (Caron et al., 2021)	ViT-S	300	BDD	29.6	86.8	41.1	90.1
iBOT (Zhou et al., 2021)	ViT-S	800	BDD	27.2	85.4	35.5	88.7
DoRA (Venkataramanan et al., 2024)	ViT-S	200	BDD	33.2	88.1	43.3	90.7
DINO (Caron et al., 2021)	R50	100	BDD	13.1	64.7	25.6	80.3
PixPro (Xie et al., 2021)	R50	100	BDD	21.8	80.0	37.3	88.0
DenseCL (Wang et al., 2021b)	R50	100	BDD	24.2	84.9	41.8	90.0
FlowE (Xiong et al., 2021)	R50	100	BDD	35.7	88.5	47.3	91.5
Supervised	R50	600	IN1K	36.7	84.7	**55.2**	92.0

Results on BDD100K

Method	Arch	Ep.	Pretrain	BDD100K Sem. Seg.			
				Linear		UperNet	
				mIoU	Acc	mIoU	Acc
Scratch	R50	-	-	9.7	55.0	26.1	81.2
DINO (Caron et al., 2021)	ViT-S	300	BDD	29.6	86.8	41.1	90.1
iBOT (Zhou et al., 2021)	ViT-S	800	BDD	27.2	85.4	35.5	88.7
DoRA (Venkataramanan et al., 2024)	ViT-S	200	BDD	33.2	88.1	43.3	90.7
DINO (Caron et al., 2021)	R50	100	BDD	13.1	64.7	25.6	80.3
PixPro (Xie et al., 2021)	R50	100	BDD	21.8	80.0	37.3	88.0
DenseCL (Wang et al., 2021b)	R50	100	BDD	24.2	84.9	41.8	90.0
FlowE (Xiong et al., 2021)	R50	100	BDD	35.7	88.5	47.3	91.5
PooDLe	R50	100	BDD	**39.2**	**89.2**	**49.9**	**91.8**
Supervised	R50	600	IN1K	36.7	84.7	**55.2**	92.0

Results on BDD100K

Method	Arch	Ep.	Pretrain	BDD100K Sem. Seg. Linear mIoU	Linear Acc	UperNet mIoU	UperNet Acc
Scratch	R50	-	-	9.7	55.0	26.1	81.2
DINO (Caron et al., 2021)	ViT-S	300	BDD	29.6	86.8	41.1	90.1
iBOT (Zhou et al., 2021)	ViT-S	800	BDD	27.2	85.4	35.5	88.7
DoRA (Venkataramanan et al., 2024)	ViT-S	200	BDD	33.2	88.1	43.3	90.7
DINO (Caron et al., 2021)	R50	100	BDD	13.1	64.7	25.6	80.3
PixPro (Xie et al., 2021)	R50	100	BDD	21.8	80.0	37.3	88.0
DenseCL (Wang et al., 2021b)	R50	100	BDD	24.2	84.9	41.8	90.0
FlowE (Xiong et al., 2021)	R50	100	BDD	35.7	88.5	47.3	91.5
PooDLe	R50	100	BDD	**39.2**	**89.2**	**49.9**	**91.8**
Supervised	R50	600	IN1K	36.7	84.7	**55.2**	92.0
PooDLe	R50	100	BDD*	**44.7**	**90.7**	54.1	**92.7**

Performance by class subgrouping

Method	Pretrain	All	Small	Large	Rare	Common
DINO	BDD	29.6	8.4	42.0	1.0	42.8
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
Supervised	IN1K	36.7	**27.2**	42.2	16.1	46.2

Performance by class subgrouping

Method	Pretrain	All	Small	Large	Rare	Common
DINO	BDD	29.6	8.4	42.0	1.0	42.8
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	16.1	46.2

Performance by class subgrouping

Method	Pretrain	All	Small	Large	Rare	Common
DINO	BDD	29.6	8.4	42.0	1.0	42.8
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	16.1	46.2
PooDLe	BDD*	**44.7**	25.2	**56.1**	**17.9**	**57.1**

BDD100K segmentations

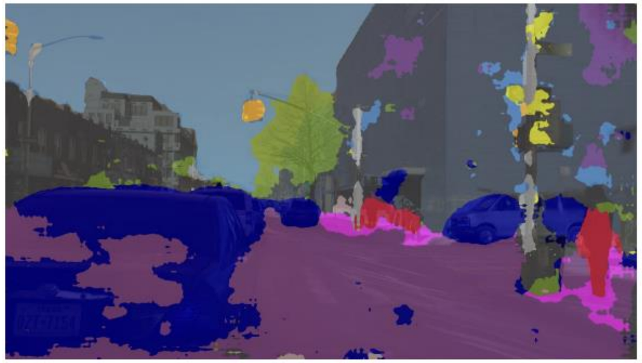
Image

Ground Truth

BDD100K segmentations

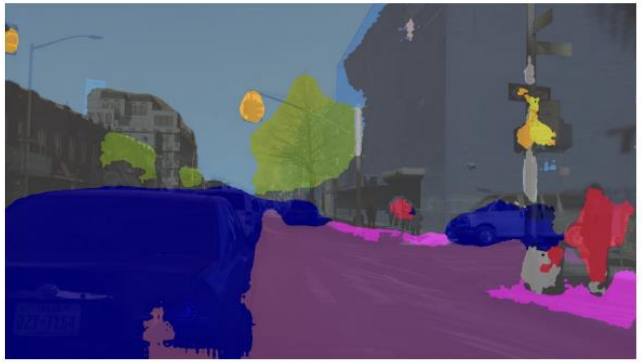
Image

Ground Truth



Supervised IN1K

36.7 mIoU



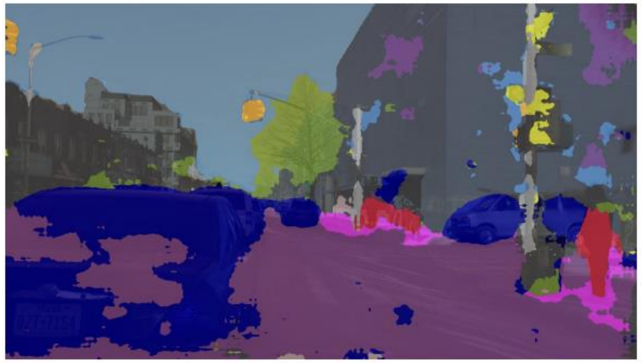
FlowE

35.7 mIoU

BDD100K segmentations

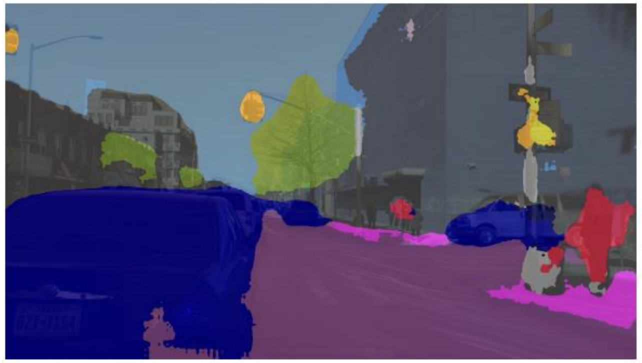
Image

Ground Truth



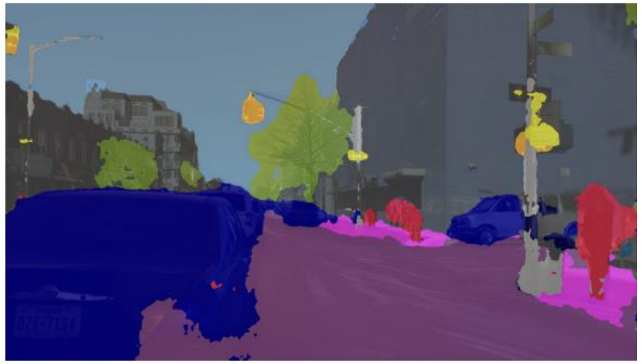
Supervised IN1K

36.7 mIoU



FlowE

35.7 mIoU



PooDLe

39.2 mIoU

See our paper for details on

- Training on WalkingTours
- Our new WalkingTours-Semantic benchmark
- Ablation studies
- Details on our subcropping strategies
- And more