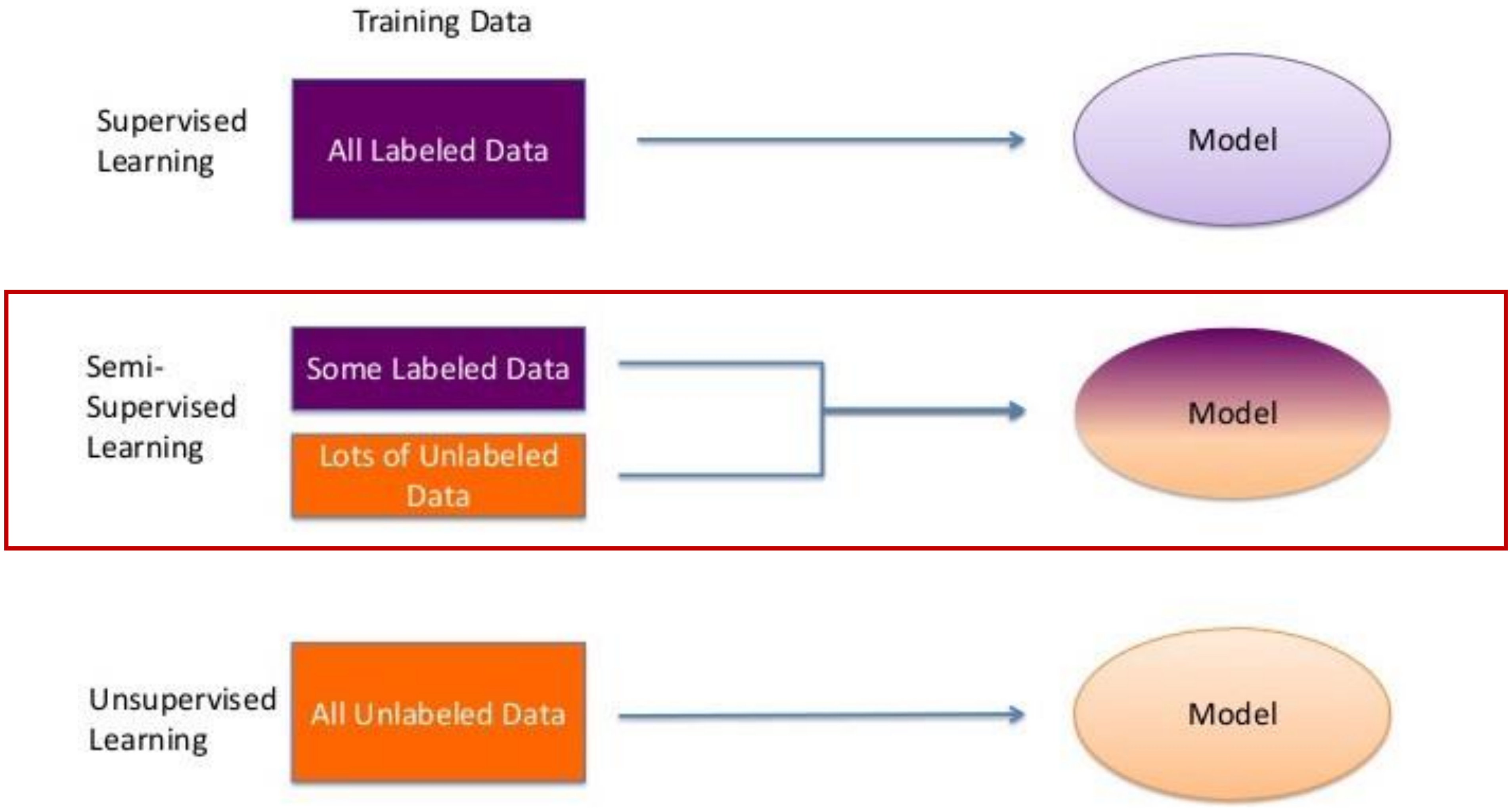
Learning from Future: A Novel Self-Training Framework for Semantic Segmentation

Ye Du[1,2] **Yujun Shen**[3] **Haochen Wang**[4] **Jingjing Fei**[5] **Wei Li**[5]
Liwei Wu[5] **Rui Zhao**[5,6] **Zehua Fu**[1,2] **Qingjie Liu**[1,2*]

[1] State Key Laboratory of Virtual Reality Technology and Systems, Beihang University
[2] Hangzhou Innovation Institute, Beihang University
[3] The Chinese University of Hong Kong
[4] Institute of Automation, Chinese Academy of Sciences [5] SenseTime Research
[6] Qing Yuan Research Institute, Shanghai Jiao Tong University, Shanghai, China

NeurIPS 2022

Semi-Supervised Learning



Unsupervised Domain Adaptive(UDA)

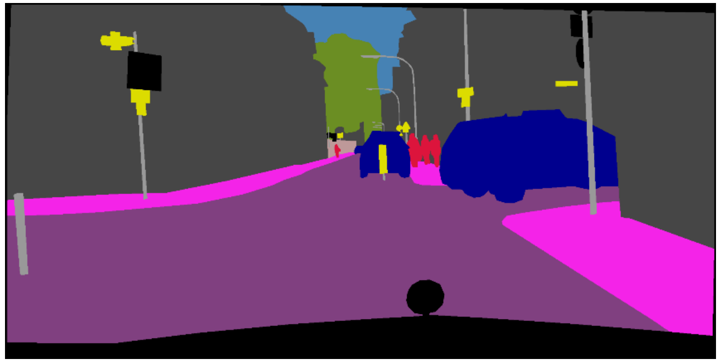
Source Domain

Target Domain

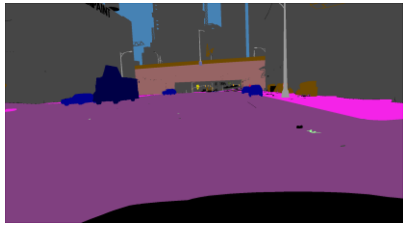
test

Test on target domain

predict



available



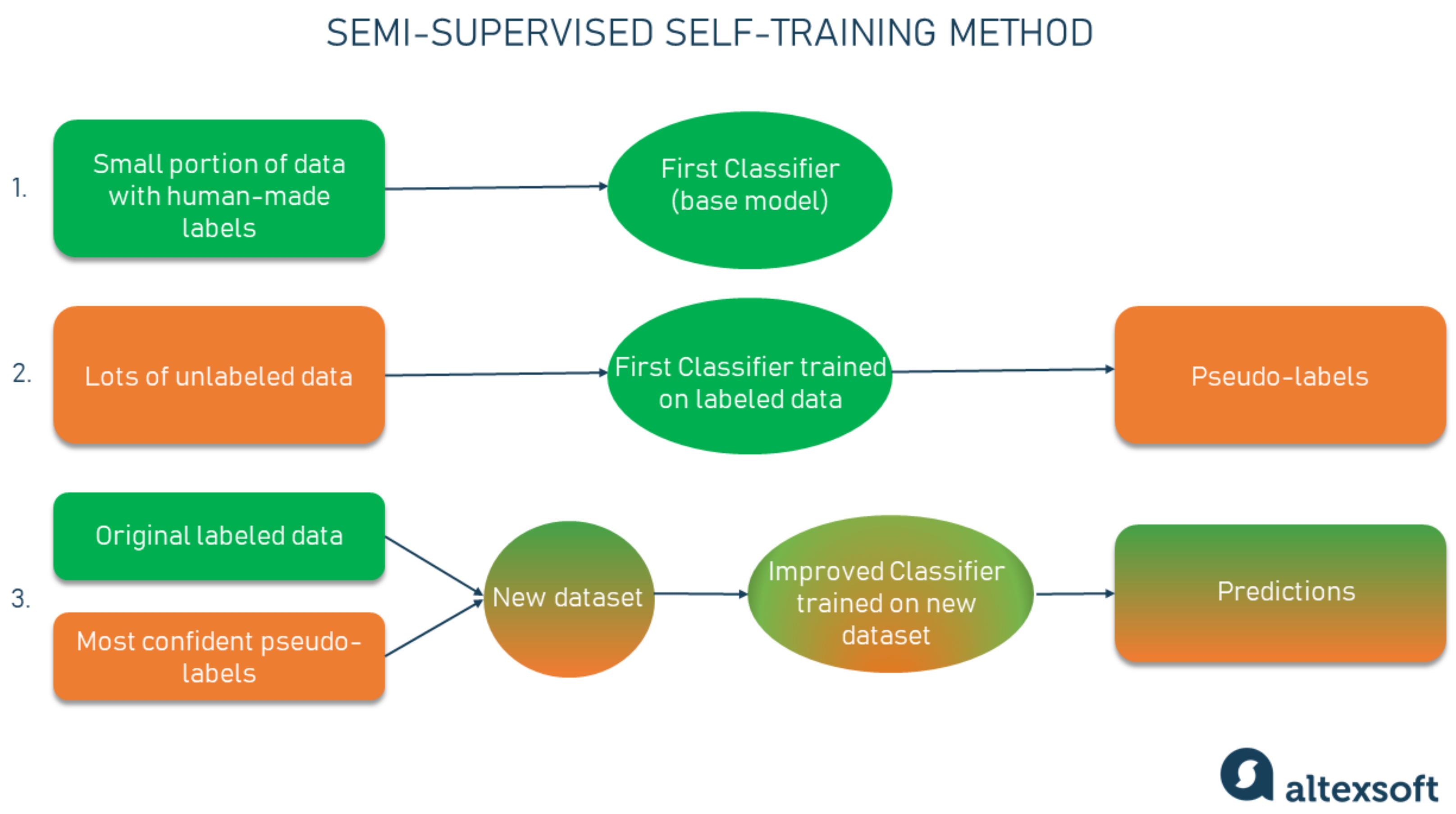
unavailable

Distribution P $\neq$ Distribution Q

Self-Training

teacher

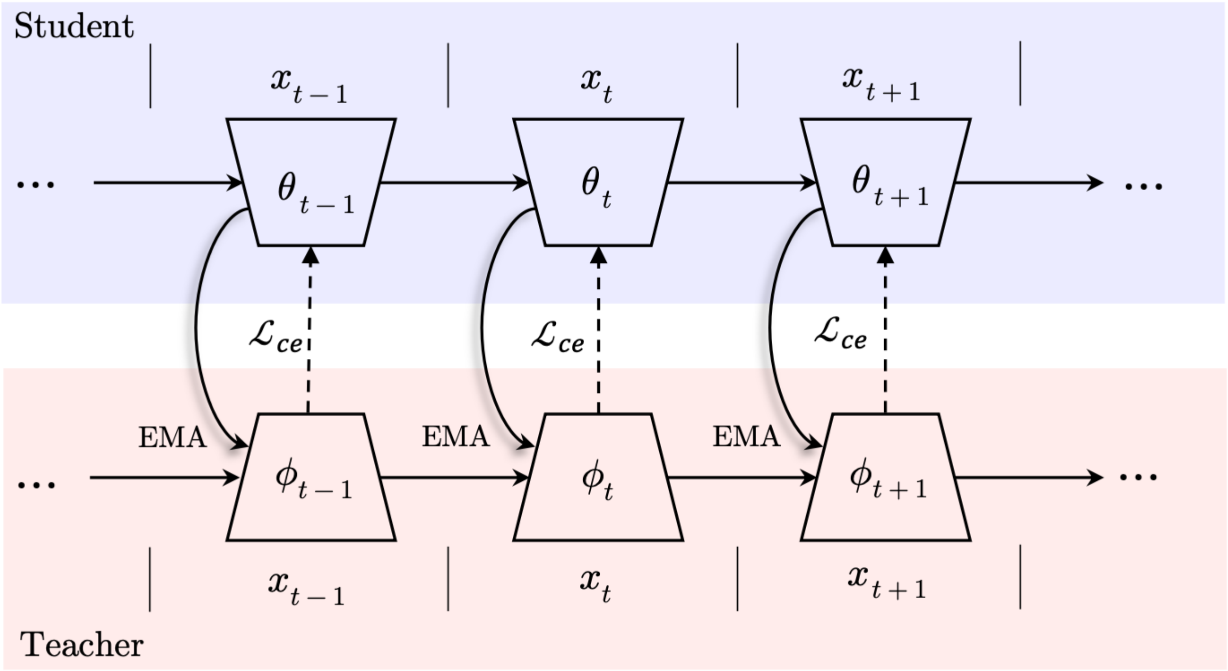
student



An example of Self-Training

Basic Self-Training

这一形式的self-training一般称作mean-teacher



(a) Self-training

每次迭代:

1. 对教师进行EMA更新
2. 教师网络产生伪标签
3. 学生网络监督式训练更新

学生网络的累积——教师

Given student θ_t and teacher ϕ_t at time t,

Confirmation bias

$$\phi_{t+1} = \mu\phi_t + (1-\mu)\theta_t,$$

EMA更新当前teacher

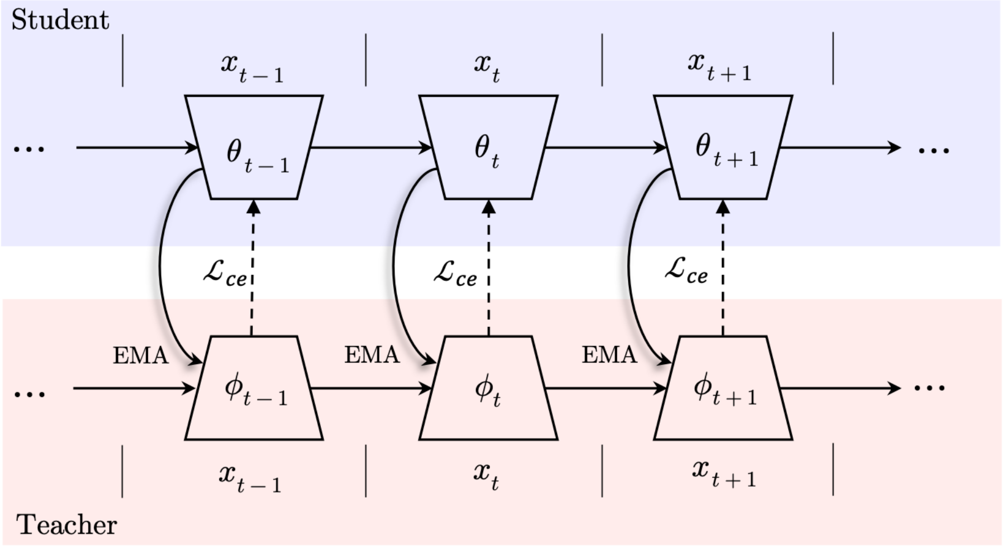
$$\theta_{t+1} = \theta_t - \gamma\nabla_\theta\left[\mathcal{L}(g_{\theta_t}(x_l), y_l) + \lambda\mathcal{L}(g_{\theta_t}(x_u), \boxed{\hat{y}_u|\phi_{t+1}})\right],$$

Student监督训练更新

teacher产生的伪标签

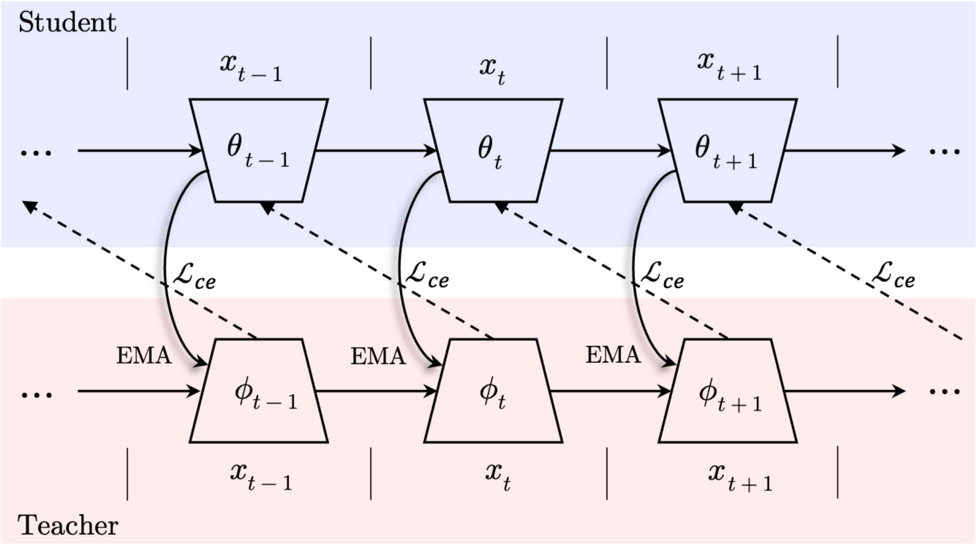
Self-Training

Teacher, a temporal ensemble of the supervised student.



(a) Self-training

Supervision signals from the current teacher

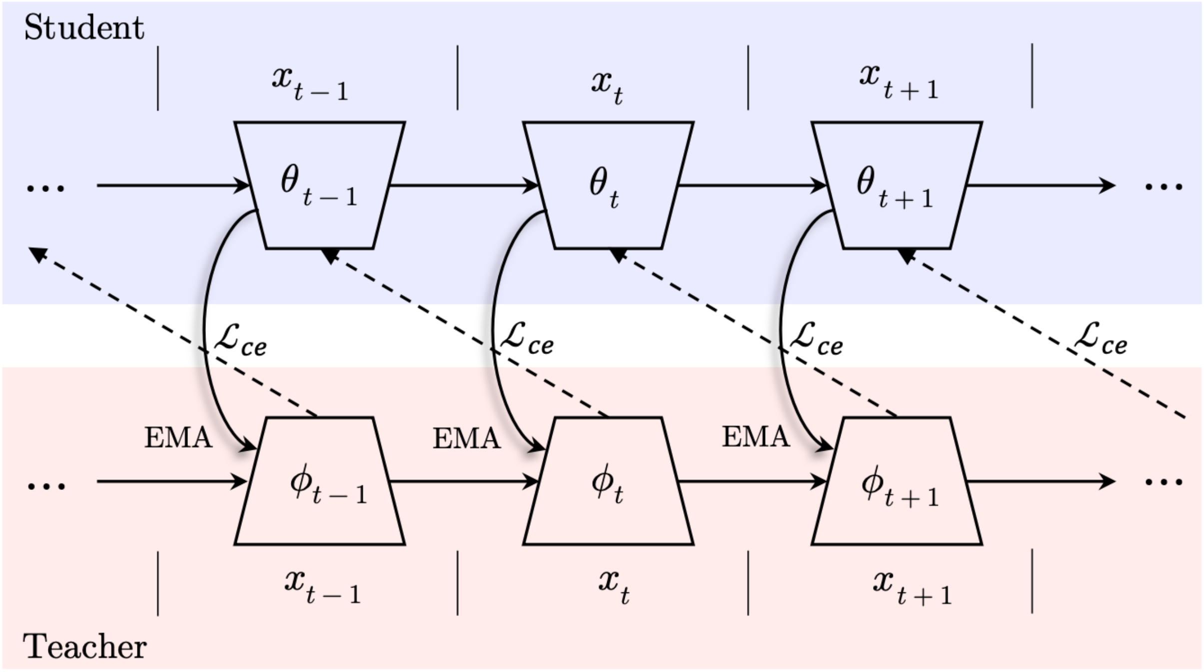


(b) Future-self-training

Supervision signals come from the future teacher

Future Self-Training

Naïve-FST



(b) Future-self-training

每次迭代:

1. 学生网络模拟更新一次
2. 教师网络EMA更新
3. 教师网络产生伪标签
4. 学生网络监督式训练更新

Given student θ_t and teacher ϕ_t at time t,

t 时刻的伪标签　模拟学生网络更新

教师网络"提前"更新一次

$$\phi_{t+1} = \mu\phi_t + (1-\mu)\left(\theta_t - \gamma\nabla_\theta\left[\mathcal{L}(g_{\theta_t}(x_l), y_l) + \lambda\mathcal{L}(g_{\theta_t}(x_u), \boxed{\hat{y}_u|\phi_t})\right]\right),$$

$$\theta_{t+1} = \theta_t - \gamma\nabla_\theta\left[\mathcal{L}(g_{\theta_t}(x_l), y_l) + \lambda\mathcal{L}(g_{\theta_t}(x_u), \boxed{\hat{y}_u|\phi_{t+1}})\right].$$

t+1 时刻的伪标签

Future Self-Training

Naïve-FST

$$\phi_{t+1} = \mu\phi_t + (1-\mu)\left(\theta_t - \gamma\nabla_\theta\left[\mathcal{L}(g_{\theta_t}(x_l), y_l) + \lambda\mathcal{L}(g_{\theta_t}(x_u), \hat{y}_u|\phi_t)\right]\right),$$
$$\theta_{t+1} = \theta_t - \gamma\nabla_\theta\left[\mathcal{L}(g_{\theta_t}(x_l), y_l) + \lambda\mathcal{L}(g_{\theta_t}(x_u), \hat{y}_u|\phi_{t+1})\right].$$

对t时刻学生网络的EMA消失了!

ST	56.3 ± 0.4	-
-	-	-
Naive-FST	56.4 ± 0.4	$\uparrow 0.1$
Improved-FST	57.7 ± 0.6	$\uparrow 1.4$

$$\phi'_{t+1} = \mu\phi_t + (1-\mu)\theta_t,$$

恢复对t时刻学生模型的EMA

Improved-FST

$$\phi_{t+1} = \mu'\phi'_{t+1} + (1-\mu')(\theta_t - \gamma\nabla_\theta[\mathcal{L}(g_{\theta_t}(x_l), y_l) + \lambda\mathcal{L}(g_{\theta_t}(x_u), \hat{y}_u|\phi'_{t+1})]),$$
$$\theta_{t+1} = \theta_t - \gamma\nabla_\theta\left[\mathcal{L}(g_{\theta_t}(x_l), y_l) + \lambda\mathcal{L}(g_{\theta_t}(x_u), \hat{y}_u|\phi_{t+1})\right],$$

Future Self-Training

two variants: FST-D & FST-W

当前时刻虚拟的学生和教师
$$\widetilde{\phi}_t = \mu\phi_t + (1-\mu)\theta_t$$
$$\widetilde{\theta}_t = \theta_t$$

$$\widetilde{\theta}_{t+k+1} = \widetilde{\theta}_{t+k} - \gamma\nabla_{\widetilde{\theta}}[\mathcal{L}(g_{\widetilde{\theta}_{t+k}}(x_l), y_l) + \lambda\mathcal{L}(g_{\widetilde{\theta}_{t+k}}(x_u), \hat{y}_u|\widetilde{\phi}_{t+k})],$$
$$\widetilde{\phi}_{t+k+1} = \mu'\widetilde{\phi}_{t+k} + (1-\mu')(\widetilde{\theta}_{t+k+1}),$$

FST-D
D-deeper

使用同样的训练样本对t时刻的学生和教师虚拟更新k次得到t+k时刻的学生和教师

$$\phi_{t+1} = \widetilde{\phi}_{t+K},$$

使用上述虚拟的t+k时刻的教师

FST-W
W-wider

$$\phi_{t+1} = \mu'\{\mu\phi_t + (1-\mu)\theta_t\} + (1-\mu')(\theta_t - \frac{1}{N}\sum_{i=1}^{N}\gamma\nabla_\theta[\mathcal{L}(g_{\theta_t}(x_l^i), y_l^i) + \lambda\mathcal{L}(g_{\theta_t}(x_u^i), \hat{y}_u^i|\phi_t)]),$$

模拟N个不同学生网络，教师对N个虚拟学生进行EMA（虚拟学生网络初始都是θ_t，但当前用于更新的样本不同，产生不同的梯度）

Future Self-Training

Pseudo code in pytorch style

```
g_t.params = mu*g_t.params+(1-mu)*g_s.params
# cache the current student
g_tmp = g_s.copy()
# pseudo label prediction: for temp network
with no_grad():
    y_u = argmax(g_t.forward(x_u))

# train the temp model
loss_l = CrossEntropyLoss(g_tmp.forward(x_l), y_l)
loss_u = CrossEntropyLoss(g_tmp.forward(x_u), y_u)
loss_virtual = loss_l + Lambda * loss_u   # calculate the loss for temp model

loss_virtual.backward()
update(g_tmp.params)    # SGD update: temp network

# momentum update with future student states
g_t.params = mu_prime * g_t.params + (1-mu_prime) * g_tmp.params
# pseudo label prediction: for student network
with no_grad():
    y_u = argmax(g_t.forward(x_u))

# train the student
loss_l = CrossEntropyLoss(g_s.forward(x_l), y_l)
loss_u = CrossEntropyLoss(g_s.forward(x_u), y_u)
loss = loss_l + Lambda * loss_u   # calculate loss for student model

loss.backward()
update(g_s.params)   # SGD update: student network
```

FST-D implementation

```
self._record_model()
for _ in range(self.ahead_step):  # look ahead
    self._update_ema(self.local_iter)
    optimizer.zero_grad()
    log_vars = self(**data_batch)
    optimizer.step()
    log_vars.pop('loss', None)
```

Experiments

ablation study of FST-D & FST-W on UDA

Task: SYNTHIA -> Cityscapes

Method	mIoU	Δ
SourceOnly	34.3 ± 2.2	-
ST	56.3 ± 0.4	-
-	-	-
Naive-FST	56.4 ± 0.4	$\uparrow 0.1$
Improved-FST	57.7 ± 0.6	$\uparrow 1.4$
FST-W	56.8 ± 0.1	$\uparrow 0.5$
FST-D	$\mathbf{59.8 \pm 0.1}$	$\uparrow \mathbf{3.5}$

future from same data batch

Method	Batch	mIoU	Δ
SourceOnly	$1\times$	34.3 ± 2.2	-
ST	$1\times$	56.3 ± 0.4	-
ST	$4\times$	55.5 ± 0.4	$\downarrow 0.8$
Naive-FST	$1\times$	58.7 ± 2.3	$\uparrow 2.3$
Improved-FST	$1\times$	58.7 ± 0.7	$\uparrow 2.4$
FST-W	$1\times$	59.3 ± 0.5	$\uparrow 3.0$
FST-D	$1\times$	$\mathbf{59.6 \pm 1.4}$	$\uparrow \mathbf{3.3}$

future from different data batch

Discussing how to implement virtual update, using the same data or different data

Experiments

Generalization on different backbones

Task: SYNTHIA -> Cityscapes

Method	K	mIoU	Δ
ST	-	55.0 ± 0.9	-
FST	2	56.3 ± 1.0	$\uparrow 1.3$
FST	3	$\mathbf{56.9 \pm 0.5}$	$\uparrow \mathbf{1.9}$
FST	4	56.4 ± 0.9	$\uparrow 1.4$

(a) DeepLabV2 [11] *w/* ResNet-50 [26].

Method	K	mIoU	Δ
ST	-	56.3 ± 0.4	-
FST	2	57.8 ± 1.3	$\uparrow 1.5$
FST	3	$\mathbf{59.8 \pm 0.1}$	$\uparrow \mathbf{3.5}$
FST	4	59.7 ± 0.8	$\uparrow 3.4$

(b) DeepLabV2 [11] *w/* ResNet-101 [26].

Method	K	mIoU	Δ
ST	-	56.3 ± 0.8	-
FST	2	58.1 ± 3.1	$\uparrow 1.8$
FST	3	58.5 ± 0.7	$\uparrow 2.2$
FST	4	$\mathbf{58.8 \pm 1.0}$	$\uparrow \mathbf{2.5}$

(c) PSPNet [75] *w/* ResNet-101 [26].

Method	K	mIoU	Δ
ST	-	61.3 ± 0.7	-
FST	2	63.7 ± 2.0	$\uparrow 2.4$
FST	3	64.3 ± 2.3	$\uparrow 3.0$
FST	4	$\mathbf{64.4 \pm 2.0}$	$\uparrow \mathbf{3.1}$

(d) UPerNet [66] *w/* Swin-B [42].

Method	K	mIoU	Δ
ST	-	59.9 ± 2.0	-
FST	2	62.5 ± 1.2	$\uparrow 2.6$
FST	3	62.5 ± 1.9	$\uparrow 2.6$
FST	4	$\mathbf{62.6 \pm 1.8}$	$\uparrow \mathbf{2.7}$

(e) UPerNet [66] *w/* BEiT-B [6].

Method	K	mIoU	Δ
ST	-	68.3 ± 0.5	-
FST	2	69.1 ± 0.3	$\uparrow 0.8$
FST	3	$\mathbf{69.3 \pm 0.3}$	$\uparrow \mathbf{1.0}$
FST	4	68.8 ± 0.9	$\uparrow 0.5$

(f) DAFormer [29] *w/* MiT-B5 [67].

FST here are the variant FST-D, K is the ahead steps

Experiments

Superparameter analysis of FST-D and FST-W

Task: SYNTHIA -> Cityscapes

Method	Backbone	K	mIoU	Δ
ST	ResNet-101	-	56.3 ± 0.4	-
FST-D	ResNet-101	2	58.6 ± 0.4	$\uparrow 2.3$
FST-D	ResNet-101	3	59.6 ± 1.4	$\uparrow 3.3$
FST-D	ResNet-101	4	$\mathbf{59.8 \pm 2.0}$	$\uparrow \mathbf{3.5}$

FST-D using different K

K means the steps ahead

Method	Backbone	N	mIoU	Δ
ST	ResNet-101	-	56.3 ± 0.4	-
FST-W	ResNet-101	2	58.5 ± 1.6	$\uparrow 2.2$
FST-W	ResNet-101	3	$\mathbf{59.3 \pm 0.5}$	$\uparrow \mathbf{3.0}$
FST-W	ResNet-101	4	58.6 ± 2.0	$\uparrow 2.3$

FST-W using different N

N means the num of different student ensembled

Experiments

Semi-supervised semantic segmentation on Pascal VOC 2012

Method	PSPNet [75]			DeepLabV2 [11]			DeepLabV3+ [12]		
	1/16	1/8	1/4	1/16	1/8	1/4	1/16	1/8	1/4
ST	65.47	72.24	75.47	68.45	72.54	76.21	73.31	74.20	77.78
FST (ours)	68.35	72.77	75.90	69.43	73.18	76.32	73.88	76.07	78.10
Δ	2.88 ↑	0.53 ↑	0.43 ↑	0.98 ↑	0.64 ↑	0.11 ↑	0.57 ↑	1.87 ↑	0.32 ↑

UDA semantic segmentation

Method	Road	S.walk	Build.	Wall	Fence	Pole	T.light	Sign	Veget.	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	M.bike	Bike	mIoU
GTAV [48] → Cityscapes [15]																				
SourceOnly	76.1	18.7	84.6	29.8	31.4	34.5	44.8	23.4	87.5	42.6	87.3	63.4	21.2	81.1	39.3	44.6	2.9	33.2	29.7	46.1
ProDA [73]	87.8	56.0	79.7	46.3	44.8	45.6	53.5	53.5	88.6	45.2	82.1	70.7	39.2	88.8	45.5	59.4	1.0	48.9	56.4	57.5
CPSL [35]	92.3	59.9	84.9	45.7	29.7	**52.8**	**61.5**	**59.5**	87.9	41.5	85.0	**73.0**	35.5	90.4	48.7	73.9	26.3	53.8	53.9	60.8
DAFormer [29]	**95.7**	**70.2**	**89.4**	53.5	**48.1**	49.6	55.8	59.4	**89.9**	47.9	92.5	72.2	44.7	92.3	74.5	78.2	65.1	55.9	61.8	68.3
FST (ours)	95.3	67.7	89.3	**55.5**	47.1	50.1	57.2	58.6	**89.9**	**51.0**	**92.9**	72.7	**46.3**	**92.5**	**78.0**	**81.6**	**74.4**	**57.7**	**62.6**	**69.3**
SYNTHIA [49] → Cityscapes [15]																				
SourceOnly	56.5	23.3	81.3	16.0	1.3	41.0	30.0	24.1	82.4	–	82.5	62.3	23.8	77.7	–	38.1	–	15.0	23.7	42.4
ProDA [73]	87.8	45.7	84.6	37.1	0.6	44.0	54.6	37.0	**88.1**	–	84.4	74.2	24.3	88.2	–	51.1	–	40.5	45.6	55.5
CPSL [35]	87.2	43.9	85.5	33.6	0.3	47.7	**57.4**	37.2	87.8	–	88.5	**79.0**	32.0	**90.6**	–	49.4	–	50.8	59.8	57.9
DAFormer [29]	84.5	40.7	**88.4**	41.5	6.5	50.0	55.0	**54.6**	86.0	–	89.8	73.2	**48.2**	87.2	–	53.2	–	53.9	61.7	60.9
FST (ours)	**88.3**	**46.1**	88.0	**41.7**	**7.3**	**50.1**	53.6	52.5	87.4	–	**91.5**	73.9	48.1	85.3	–	**58.6**	–	**55.9**	**63.4**	**61.9**

Experiments

Performance on semi-supervised semantic segmentation

Method	1/16	1/8	1/4
SupOnly†	67.87	71.55	75.80
CutMix† [18]	71.66	75.51	77.33
CCT [47]	71.86	73.68	76.51
GCT [32]	70.90	73.29	76.66
CPS [13]	<u>72.18</u>	<u>75.83</u>	<u>77.55</u>
FST (ours)	**73.88**	**76.07**	**78.10**

(a) PASCAL VOC 2012 [17].

Method	1/16	1/8	1/4
SupOnly†	65.74	72.53	74.43
CutMix† [18]	67.06	71.83	76.36
CCT [47]	69.32	74.12	75.99
GCT [32]	66.75	72.66	76.11
CPS [13]	<u>70.50</u>	**75.71**	**77.41**
FST (ours)	**71.03**	<u>75.36</u>	<u>76.61</u>

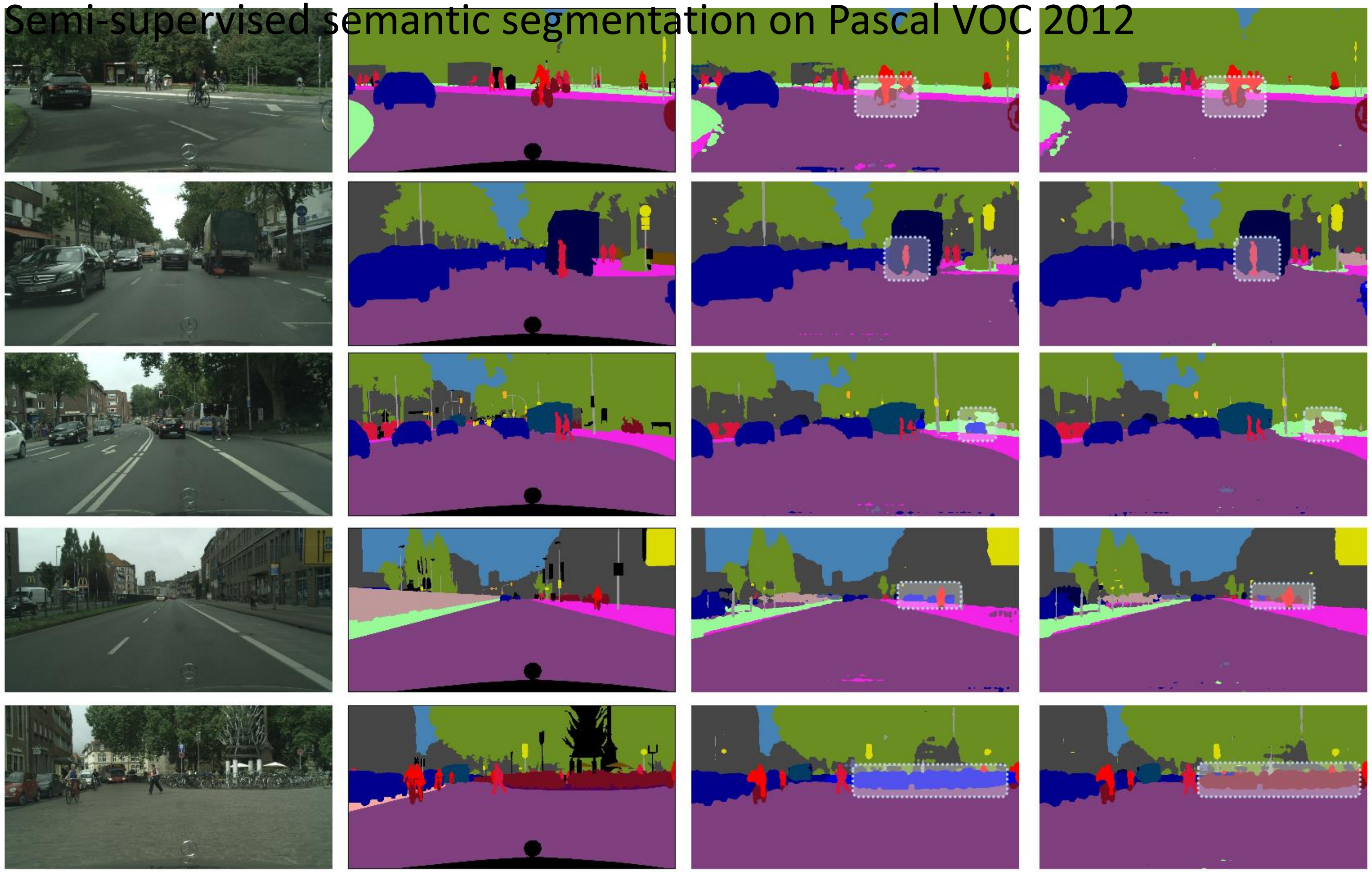
(b) Cityscapes [15].

All competitors are methods with improvement on basic framework. With any improvement tricks like strong data augmentation or contrastive learning.

Effect of FST on improving pseudo-label quality and performance.

Visualization

Semi-supervised semantic segmentation on Pascal VOC 2012



(a) Image (b) Ground Truth (c) ST (d) FST (ours)