CVPR2022

DAFormer: Improving Network Architectures and Training Strategies for Domain-Adaptive Semantic Segmentation

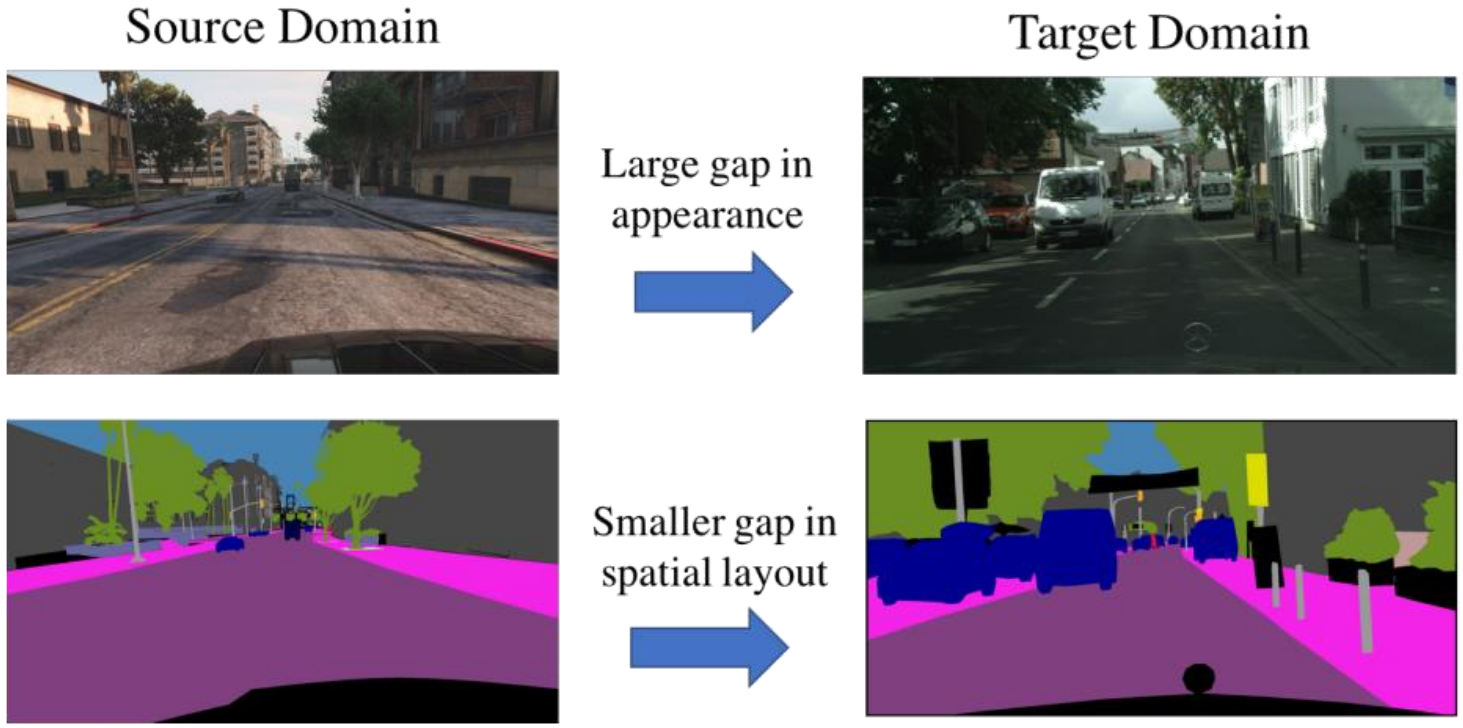
Lukas Hoyer
ETH Zurich
lhoyer@vision.ee.ethz.ch

Dengxin Dai
MPI for Informatics &
ETH Zurich
dai@vision.ee.ethz.ch

Luc Van Gool
ETH Zurich &
KU Leuven
vangool@vision.ee.ethz.ch

01 DAformer

- **Background**

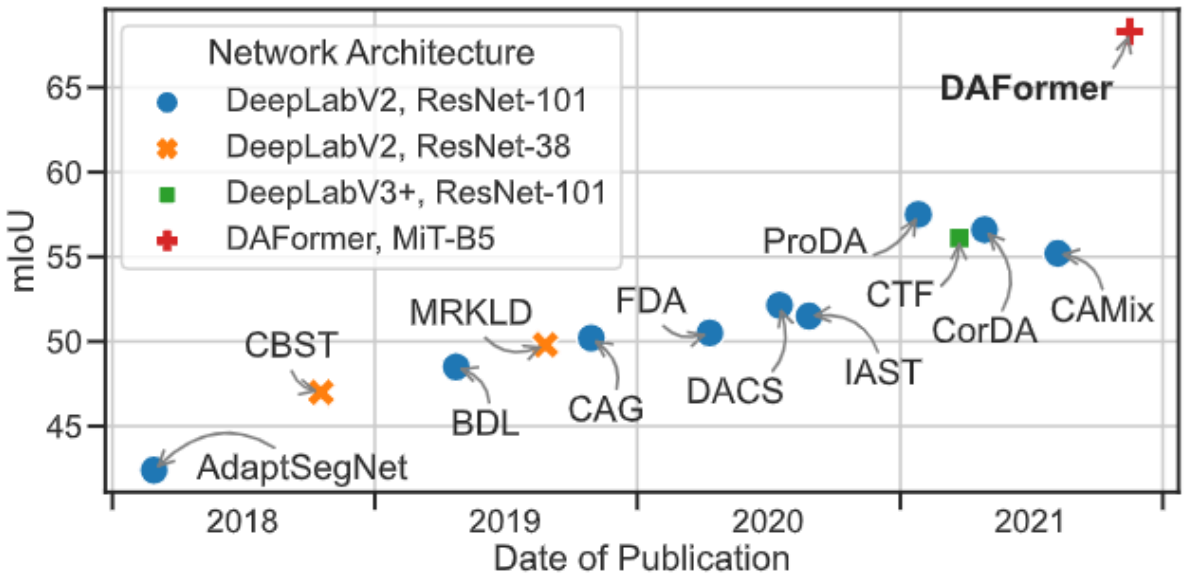


Target sample　　Adaptation result　　Ground truth

01 DAformer

- **CVPR (2022)**
- **Motivation**

(1)Previous UDA methods mostly evaluated their contributions using a DeepLabV2 [6] or FCN8s [48] network architecture with ResNet or VGG backbone in order to be comparable to previously published works. However, even their strongest architecture (DeepLabV2+ResNet101) is outdated in the area of supervised semantic segmentation.
(2) As the potential of a more capable architecture, such as DAFormer, can be impaired by unstable training and over-fitting to the source domain



01 DAformer

- **Contribution**
  - Compile DAFormer, a network architecture tailored for UDA
  - Introduce three training strategies to UDA to avoid overfitting to the source domain

01 DAformer

- **Overview of the Proposed Model**

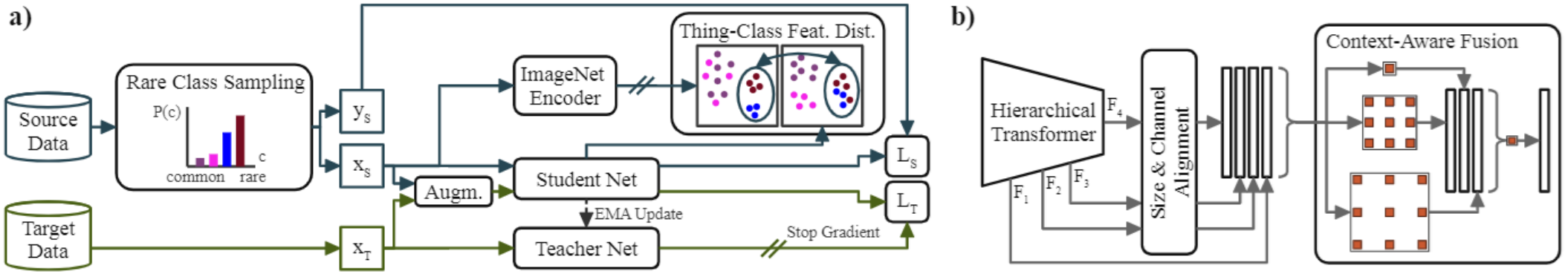
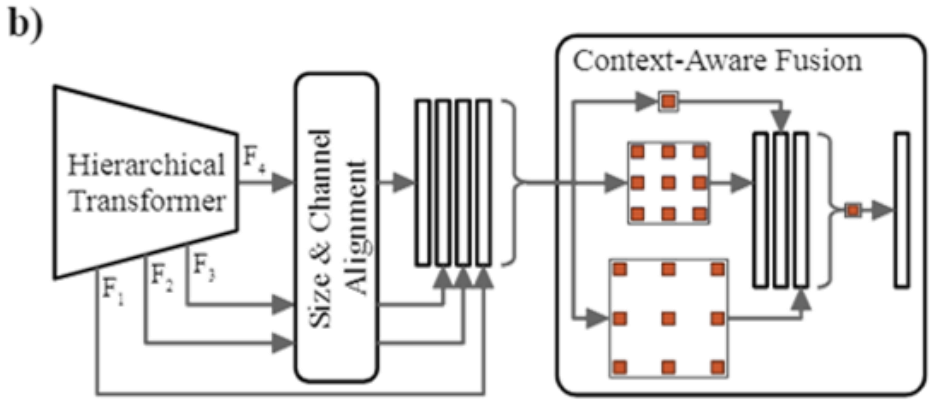


Figure 2. Overview of our UDA framework with Rare Class Sampling, Thing-Class Feature Distance, and DAFormer network.

- **(1)** $\mathcal{L}_S^{(i)} = -\sum_{j=1}^{H\times W}\sum_{c=1}^{C} y_S^{(i,j,c)} \log g_\theta(x_S^{(i)})^{(j,c)}$
- **(2)** $p_T^{(i,j,c)} = [c = \arg\max_{c'} h_\phi(x_T^{(i)})^{(j,c')}]$, $q_T^{(i)} = \frac{\sum_{j=1}^{H\times W}[\max_{c'} h_\phi(x_T^{(i)})^{(j,c')} > \tau]}{H\cdot W}$. $\mathcal{L}_T^{(i)} = -\sum_{j=1}^{H\times W}\sum_{c=1}^{C} q_T^{(i)} p_T^{(i,j,c)} \log g_\theta(x_T^{(i)})^{(j,c)}$.
- **(3) Exponentially moving average:** $\phi_{t+1} \leftarrow \alpha\phi_t + (1-\alpha)\theta_t$.
- **(4) overall UDA loss:** $\mathcal{L} = \mathcal{L}_S + \mathcal{L}_T + \lambda_{FD}\mathcal{L}_{FD}$.

01 DAformer

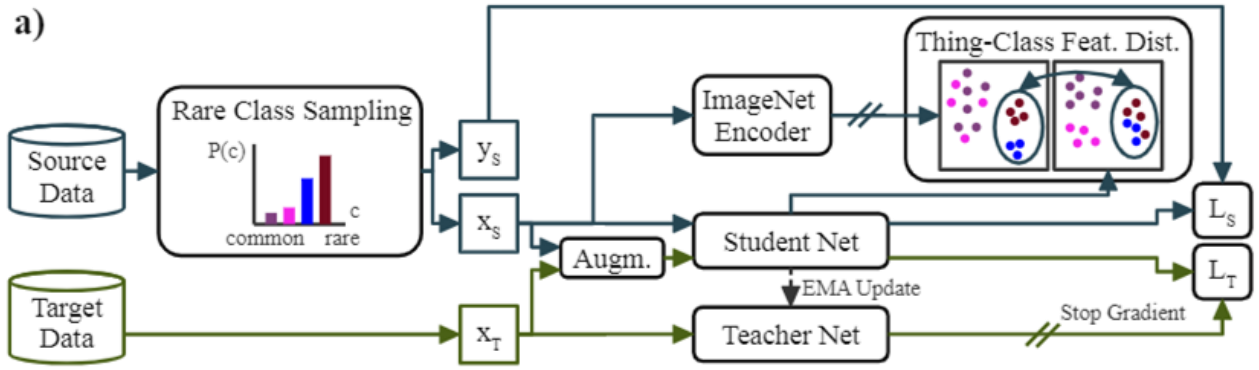
- **DAformer Network Architecture**



- **(1) Follow the design of Mix Transformers(SegFormer)**
- **(2) Utilize additional context information in the decoder**
- **(3) Use multiple parallel 3×3 depthwise separable convolutions with different dilation rates and a 1×1 convolution to fuse them**

01 DAformer

- **Training Strategies for UDA**



- **(1) Rare Class Sampling (RCS):** $f_c = \frac{\sum_{i=1}^{N_S} \sum_{j=1}^{H \times W} [y_S^{(i,j,c)}]}{N_S \cdot H \cdot W}$. $P(c) = \frac{e^{(1-f_c)/T}}{\sum_{c'=1}^{C} e^{(1-f_{c'})/T}}$.
- **(2) Thing-Class ImageNet Feature Distance (FD):**

$$d^{(i,j)} = ||F_{ImageNet}(x_S^{(i)})^{(j)} - F_\theta(x_S^{(i)})^{(j)}||_2 . \quad \mathcal{L}_{FD}^{(i)} = \frac{\sum_{j=1}^{H_F \times W_F} d^{(i,j)} \cdot M_{things}^{(i,j)}}{\sum_j [M_{things}^{(i,j)}]},$$

$$M_{things}^{(i,j)} = \sum_{c'=1}^{C} y_{S,small}^{i,j,c'} \cdot [c' \in \mathcal{C}_{things}] . \qquad y_{S,small}^{c} = [\text{AdaptAvgPool}(y_S^c, H_F, W_F) > r] .$$

- **(3) Learning Rate Warmup for UDA:the learning rate at iteration t is set** $\eta_t = \eta_{base} \cdot t/t_{warm}$.

01 DAfomer

- **Experimental Results：**

	Road	S.walk	Build.	Wall	Fence	Pole	Tr.Light	Sign	Veget.	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	M.bike	Bike	mIoU
GTA5 → Cityscapes																				
CBST [99]	91.8	53.5	80.5	32.7	21.0	34.0	28.9	20.4	83.9	34.2	80.9	53.1	24.0	82.7	30.3	35.9	16.0	25.9	42.8	45.9
DACS [71]	89.9	39.7	87.9	30.7	39.5	38.5	46.4	52.8	88.0	44.0	88.8	67.2	35.8	84.5	45.7	50.2	0.0	27.3	34.0	52.1
CorDA [79]	94.7	63.1	87.6	30.7	40.6	40.2	47.8	51.6	87.6	47.0	89.7	66.7	35.9	90.2	48.9	57.5	0.0	39.8	56.0	56.6
ProDA [93]	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
DAFormer	**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**
Synthia → Cityscapes																				
CBST [99]	68.0	29.9	76.3	10.8	1.4	33.9	22.8	29.5	77.6	–	78.3	60.6	28.3	81.6	–	23.5	–	18.8	39.8	42.6
DACS [71]	80.6	25.1	81.9	21.5	2.9	37.2	22.7	24.0	83.7	–	**90.8**	67.6	38.3	82.9	–	38.9	–	28.5	47.6	48.3
CorDA [79]	**93.3**	**61.6**	85.3	19.6	5.1	37.8	36.6	42.8	84.9	–	90.4	69.7	41.8	85.6	–	38.4	–	32.6	53.9	55.0
ProDA [93]	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
DAFormer	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**

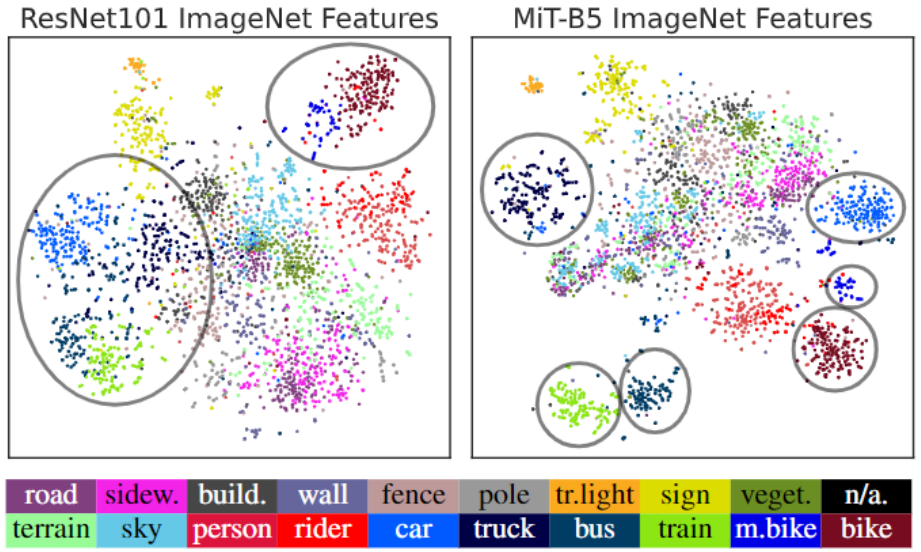
01 DAfomer

- **Comparison of Network Architectures for UDA：**

Architecture	Src-Only	UDA	Oracle	Rel.
DeepLabV2 [6]	34.3 ±2.2	54.2 ±1.7	72.1 ±0.5	75.2%
DA Net [18]	30.9 ±2.1	53.7 ±0.2	72.6 ±0.2	74.0%
ISA Net [35]	32.3 ±2.1	53.3 ±0.4	72.0 ±0.5	74.0%
DeepLabV3+ [7]	31.0 ±1.4	53.7 ±1.0	75.6 ±0.9	71.0%
SegFormer [86]	**45.6** ±0.6	**58.2** ±0.9	**76.4** ±0.2	**76.2%**

Encoder	Decoder	UDA	Oracle	Rel.
MiT-B5 [86]	SegF. [86]	58.1 ±0.9	76.4 ±0.2	76.1%
MiT-B5 [86]	DLv3+ [7]	56.8 ±1.8	75.5 ±0.5	75.2%
R101 [25]	SegF. [86]	50.9 ±1.1	71.3 ±1.3	71.4%
R101 [25]	DLv3+ [7]	53.7 ±1.0	75.6 ±0.9	71.0%

Enc.	Dec.	Src-Only	UDA	Oracle	Rel.
R50 [25]	DLv2 [6]	29.3	52.1	70.8	73.6%
R101 [25]	DLv2 [6]	36.9	53.3	72.5	73.5%
S50 [92]	DLv2 [6]	27.9	48.0	67.7	70.9%
S101 [92]	DLv2 [6]	35.5	53.5	72.2	74.1%
S200 [92]	DLv2 [6]	35.9	56.9	73.5	77.4%
MiT-B3 [86]	SegF. [86]	42.2	50.8	76.5	66.4%
MiT-B4 [86]	SegF. [86]	44.7	57.5	77.1	74.6%
MiT-B5 [86]	SegF. [86]	46.2	58.8	76.2	77.2%



01 DAfomer

- **Ablation study：**

	Network	Warmup	RCS	FD	Misc.	UDA
1	SegF. [86]	–	–	–	–	51.8 ±0.8
2	SegF. [86]	✓	–	–	–	58.2 ±0.9
3	SegF. [86]	✓	✓(T=∞)	–	–	62.0 ±1.5
4	SegF. [86]	✓	✓	–	–	64.0 ±2.4
5	SegF. [86]	✓	–	✓(all $\mathcal{C}$)	–	58.8 ±0.4
6	SegF. [86]	✓	–	✓	–	61.7 ±2.6
7	SegF. [86]	✓	✓	✓	–	66.2 ±1.0
8	SegF. [86]	✓	✓	✓	Crop PL, $\alpha\uparrow$	67.0 ±0.4
9	DLv2 [6]	–	–	–	–	49.1 ±2.0
10	DLv2 [6]	✓	✓	✓	Crop PL, $\alpha\uparrow$	56.0 ±0.5

Decoder	C_e	#Params	UDA	Oracle	Rel.
SegF. [86]	768	3.2M	67.0 ±0.4	76.8 ±0.3	87.2%
SegF. [86]	256	0.5M	67.1 ±1.1	76.5 ±0.4	87.7%
UperNet [85]	512	29.6M	67.4 ±1.1	**78.0** ±0.2	86.4%
UperNet [85]	256	8.3M	66.7 ±1.2	77.4 ±0.3	86.2%
ISA [35] Fusion	256	1.1M	66.3 ±0.9	76.3 ±0.4	86.9%
Context only at F_4	256	3.2M	67.0 ±0.6	76.6 ±0.2	87.5%
DAFormer w/o DSC	256	10.0M	67.0 ±1.5	76.7 ±0.6	87.4%
DAFormer	256	3.7M	**68.3** ±0.5	77.6 ±0.2	**88.0%**

Architecture	LR Warmup	UDA	Oracle	Rel.
DeepLabV2 [6]	–	49.1 ±2.0	67.4 ±1.7	72.8%
DeepLabV2 [6]	✓	54.2 ±1.7	72.1 ±0.5	75.2%
SegFormer [86]	–	51.8 ±0.8	72.9 ±1.6	71.1%
SegFormer [86]	✓	58.2 ±0.9	76.4 ±0.2	76.2%

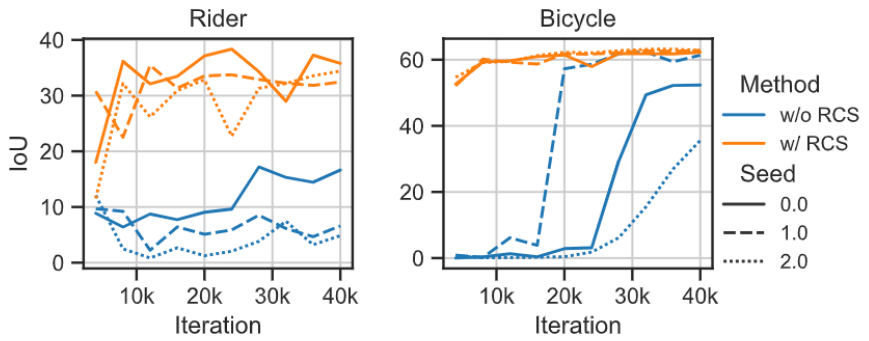


Figure 4. SegFormer UDA performance for the rare classes rider and bicycle without and with Rare Class Sampling (RCS).

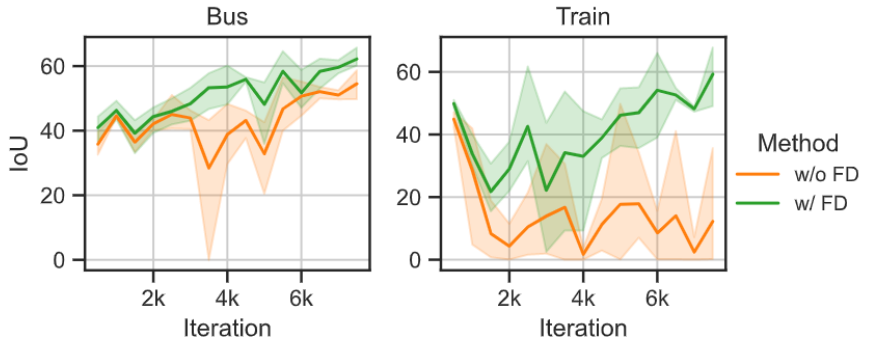


Figure 5. SegFormer UDA performance in the beginning of the training with and without ImageNet Feature Distance (FD).