Computer Vision Lab
ETH zürich

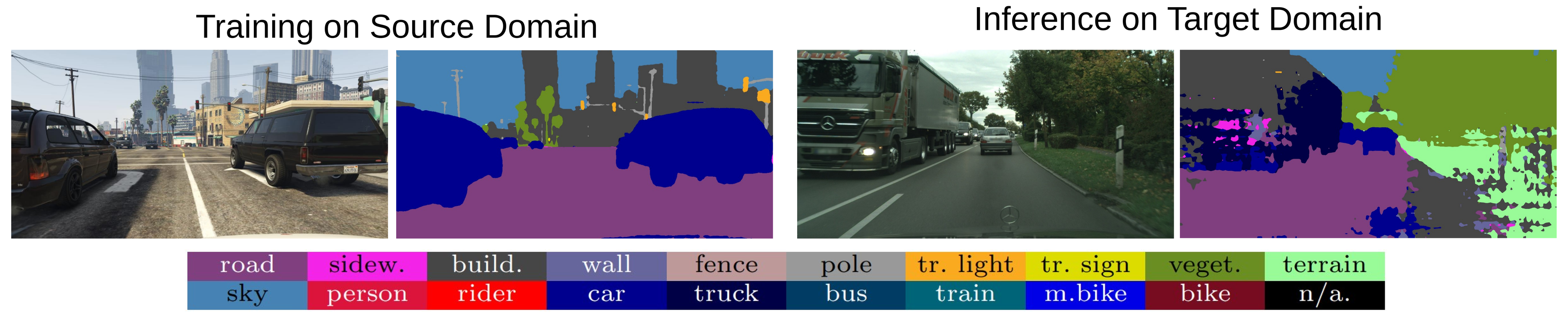
DAFormer: Improving Network Architectures and Training Strategies for Domain-Adaptive Semantic Segmentation

Lukas Hoyer, Dengxin Dai, Luc Van Gool • github.com/lhoyer/DAFormer

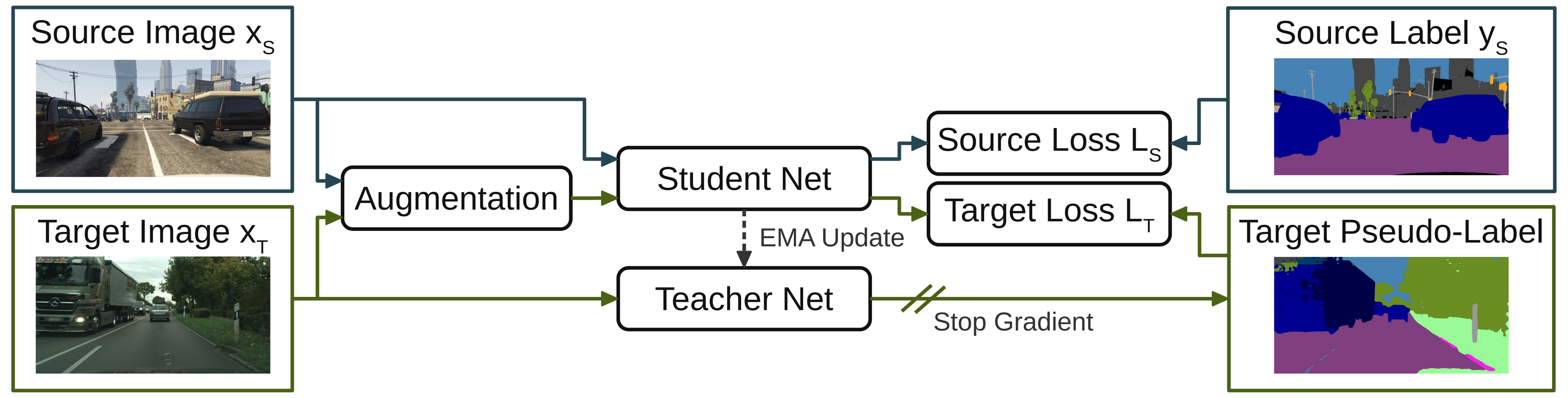
CVPR JUNE 19-24 2022 NEW ORLEANS • LOUISIANA

1 Unsupervised Domain Adaptation (UDA)

Problem: Networks trained on one domain often experience a performance drop on another domain



Solution: Adapt network to unlabeled target images, e.g. using self-training [2]

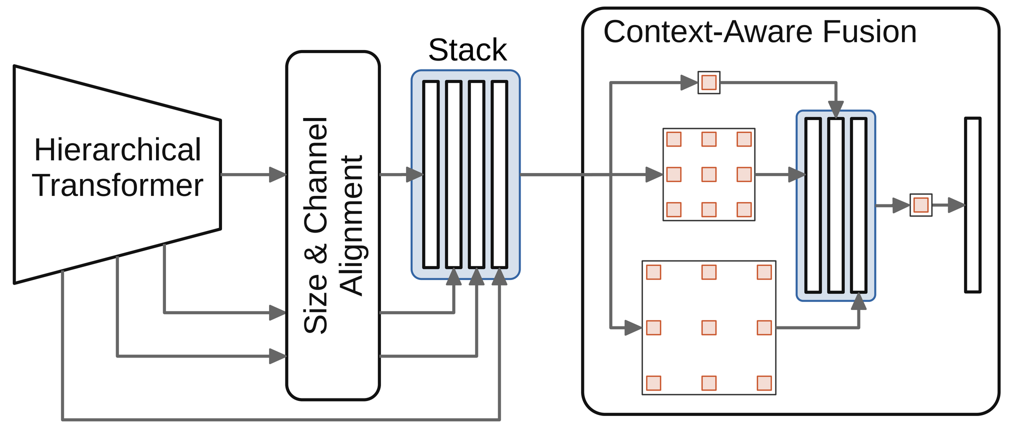


2 Improved Network Architecture for UDA

Problem: UDA methods were only evaluated on outdated network architectures (e.g. DeepLabV2 [1])

Solution: Design of an architecture tailored for UDA

- Hierarchical Transformer encoder [3]
- Context-aware multi-level feature fusion decoder



Network Architecture	UDA	Oracle	UDA / Oracle
DeepLabV2 [1]	56.0	72.1	77.7%
Hierarchical Transformer [3]	67.0	76.8	87.2%
Full DAFormer	68.3	77.6	88.0%

➔ DAFormer newly reveals the high potential of Transformers for UDA

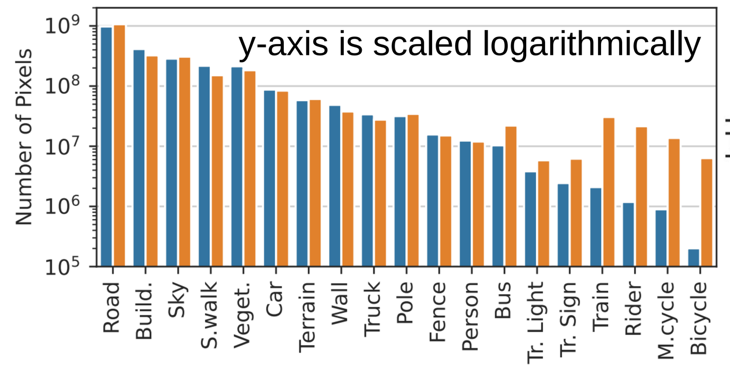
3 Improved Training Strategies for UDA

Rare Class Sampling (RCS)

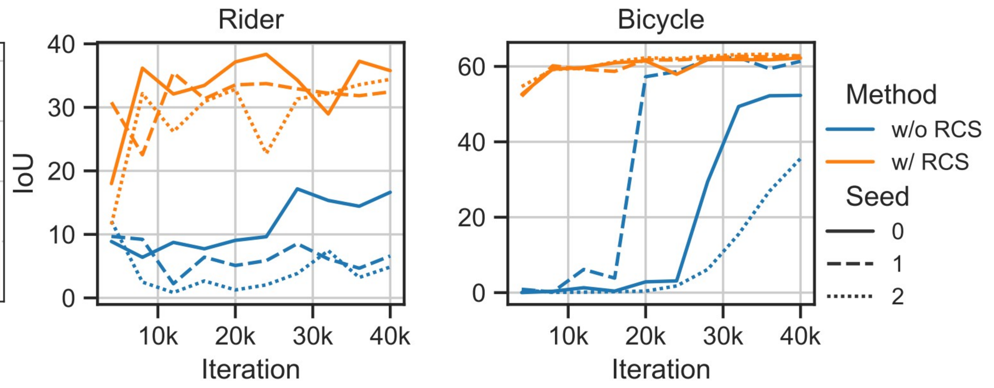
Problem: Source domain class distribution is heavily imbalanced
Solution: Sample images with rare classes more often

Sample images containing class c with probability $P(c) = \frac{e^{(1-f_c)/T}}{\sum_{c'=1}^{C} e^{(1-f_{c'})/T}}$, where

$f_c = \frac{\sum_{i=1}^{N_S}\sum_{j=1}^{H\times W}[y_S^{(i,j,c)}]}{N_S \cdot H \cdot W}$ is the class frequency in the source dataset.



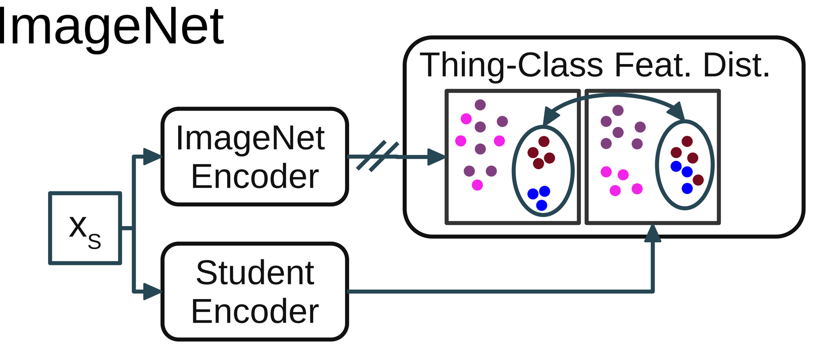
➔ RCS samples more pixels with rare classes



➔ Rare classes are learned better and more stably with RCS

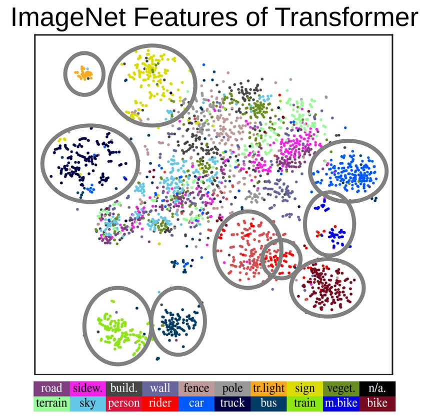
Thing-Class ImageNet Feature Distance (FD)

Problem: Distinctive features from ImageNet pre-training are forgotten
Solution: Reduce distance to ImageNet features of thing-classes, which are part of ImageNet

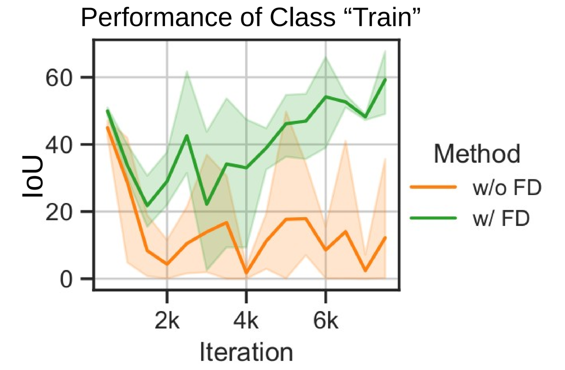


$$\mathcal{L}_{FD}^{(i)} = \frac{\sum_{j=1}^{H_F \times W_F} d^{(i,j)} \cdot M_{things}^{(i,j)}}{\sum_j M_{things}^{(i,j)}}$$

$$d^{(i,j)} = ||F_{ImageNet}(x_S^{(i)})^{(j)} - F_\theta(x_S^{(i)})^{(j)}||_2$$



➔ ImageNet features distinguish thing-classes



➔ w/o FD the class "train" is forgotten
➔ w/ FD "train" is successfully adapted

Learning Rate Warm-Up

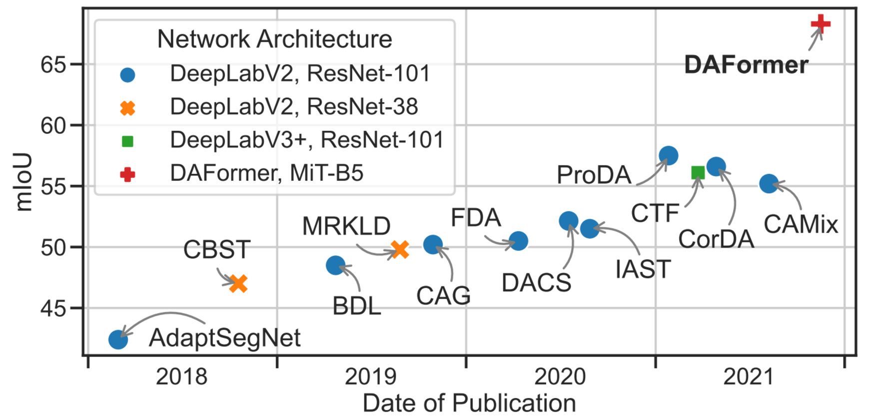
Problem: A high learning rate distorts features from pre-training in the early training
Solution: Start training with small learning rate

References

[1] Chen et al. "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs." PAMI. 2017
[2] Tranheden et al. "DACS: Domain Adaptation via Cross-domain Mixed Sampling." WACV. 2021.
[3] Xie et al. "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers." NeurIPS. 2021
[4] Zhang et al. "Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation." CVPR. 2021

4 Evaluation on GTA→Cityscapes

Comparison with State-of-the-Art



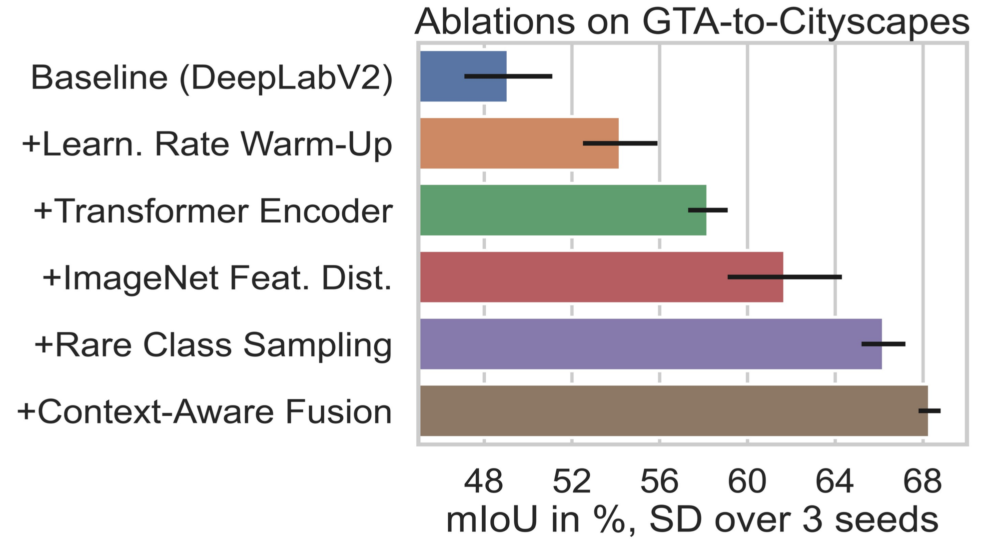
➔ DAFormer improves the State-of-the-Art by 10.8 mIoU

Class-Wise IoU Improvement over SOTA

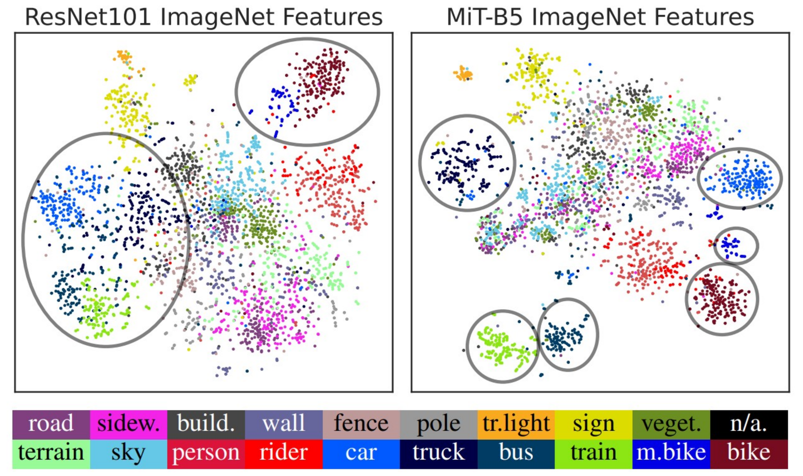
	Road	S.walk	Build.	Wall	Fence	Pole	T.Light	T.Sign	Veget.	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	M.bike	Bike
ProDA [4]	88	56	80	46	45	46	54	54	89	45	82	71	39	89	46	59	1	49	56
DAFormer	96	70	89	53	48	50	56	59	90	48	93	72	45	92	75	78	65	56	62

➔ DAFormer particularly improves difficult classes

Ablation Study



➔ Each component contributes a significant performance improvement



➔ The pre-trained Transformer encoder better separates vehicle classes

Example Predictions