Cross-Dataset Collaborative Learning for Semantic Segmentation

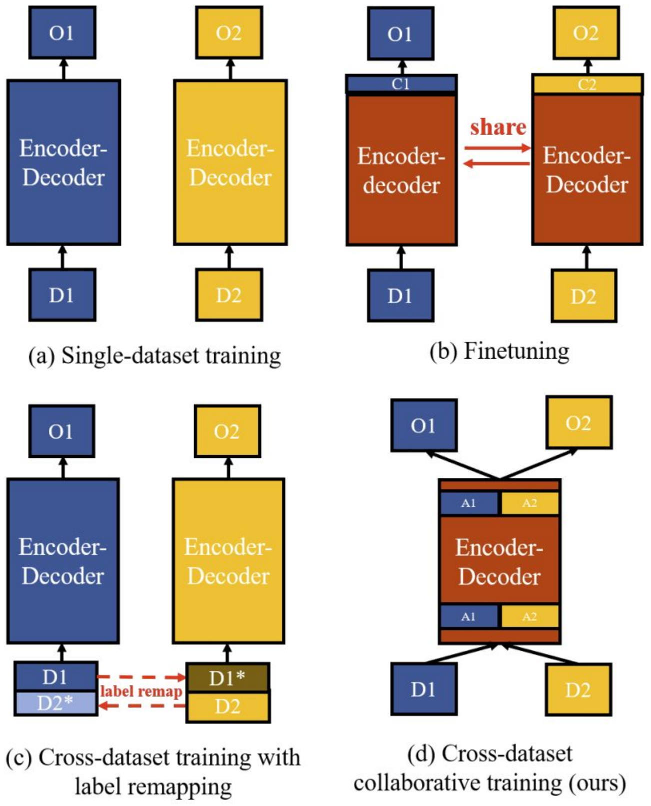
CVPR 2021

Sanghun Jung

DAVIAN
Data and Visual Analytics Lab

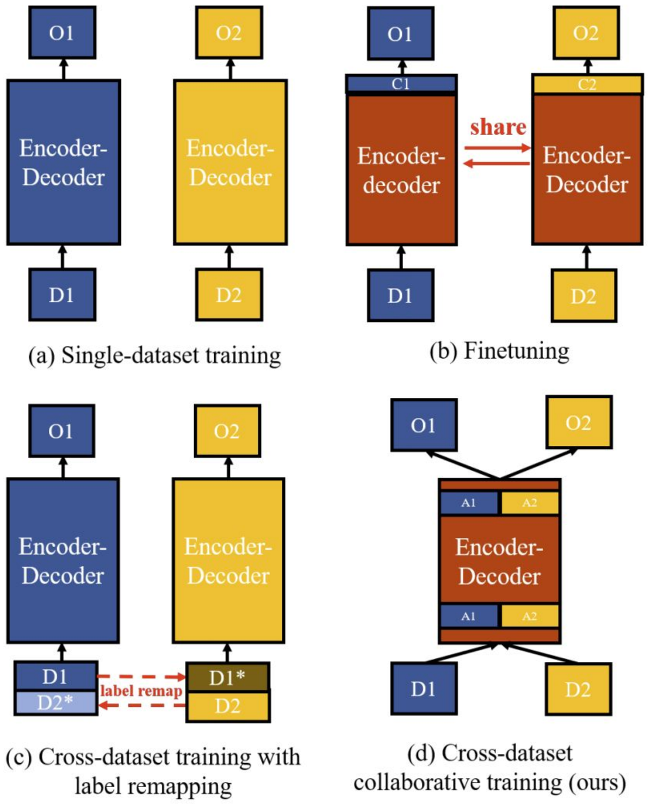
What is collaborative learning?

- Train multiple classifier heads of the same network on the same training dataset. → Better Generalization Capability!
  - Collaborative Learning for Deep Neural Networks (NeurIPS 2018)



- (a) Usual single-dataset training.
- (b) Train a model on auxiliary dataset and finetune the target dataset.
- (c) Concatenate the datasets by remapping the label space of the source to that of the target dataset.
- (d) Proposed approach using collaborative learning.
  - An unified network for multiple datasets!

Limitations of Previous Approaches - From data utilization perspective



- (a) Single-dataset Training:
  - Need to label more dataset for training
- (b) Finetuning:
  - Time consuming since it requires to repeat the training process for different datasets.
  - Not applicable for joint optimization
- (c) Label remapping:
  - Needs assumption of class overlap across different datasets.

Datasets

(a) Visualization for samples of different datasets

1. Urban-scene datasets
   a. Cityscapes (19 classes)
   b. BDD100K (19 classes)
   c. Camvid (11 classes)

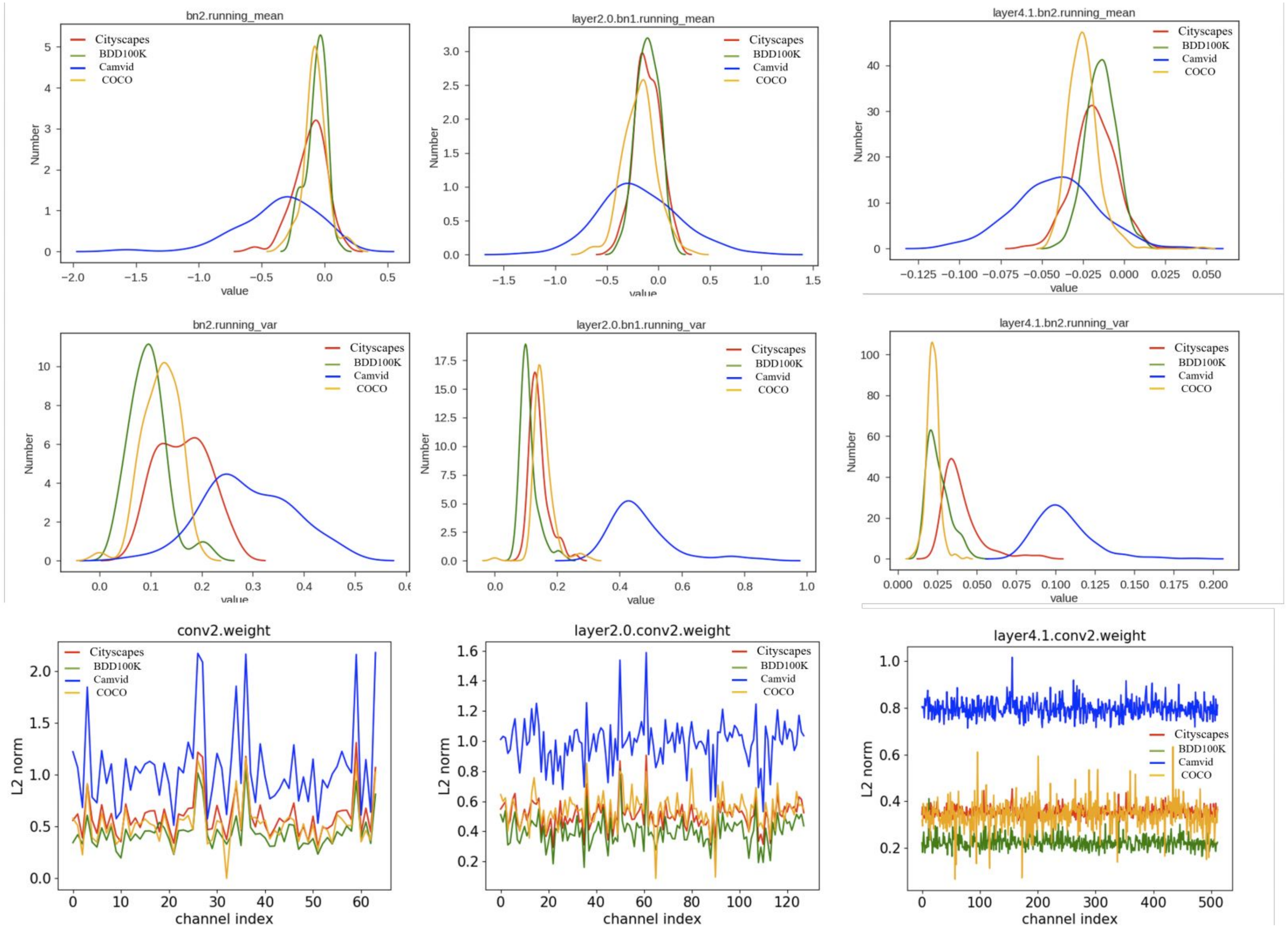
2. General-scene datasets
   a. COCO (182 categories, 91 from things and 91 from stuffs)

Key Observations (1)

1. The convolution filters can be shared for multiple datasets to maintain the network efficiency.

2. The batch normalization layers are not appropriate to share across different datasets.

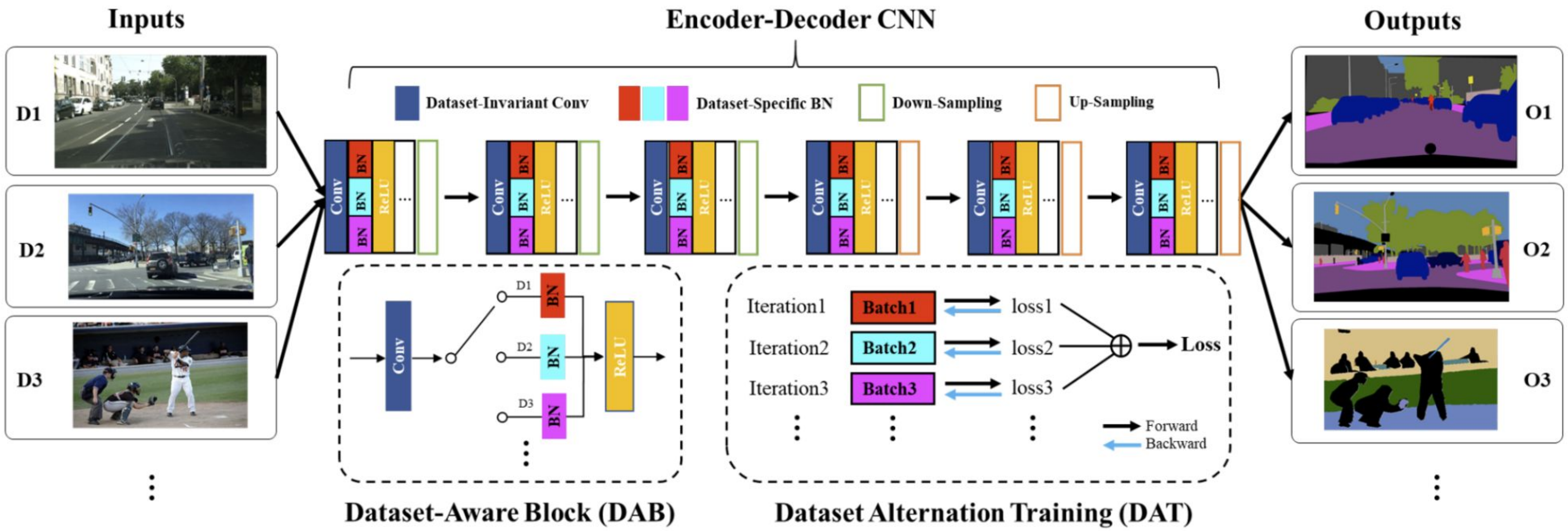
Key Observations (2)



1. Running mean and variance of BN have different distribution for different datasets.
   a. Need dataset-specific BN!

2. Convolution filters hold the same distribution across different datasets.
   a. Conv layers can be shared!

Method Overview



Dataset-Aware Block (DAB)

- DAB consists of
  a. A dataset-invariant convolution layer
  b. Multiple dataset-specific batch normalization layers
     - Dataset-specific batch normalization (DSBN - Domain-Specific Batch Normalization for Unsupervised Domain Adaptation)
  c. An activation layer

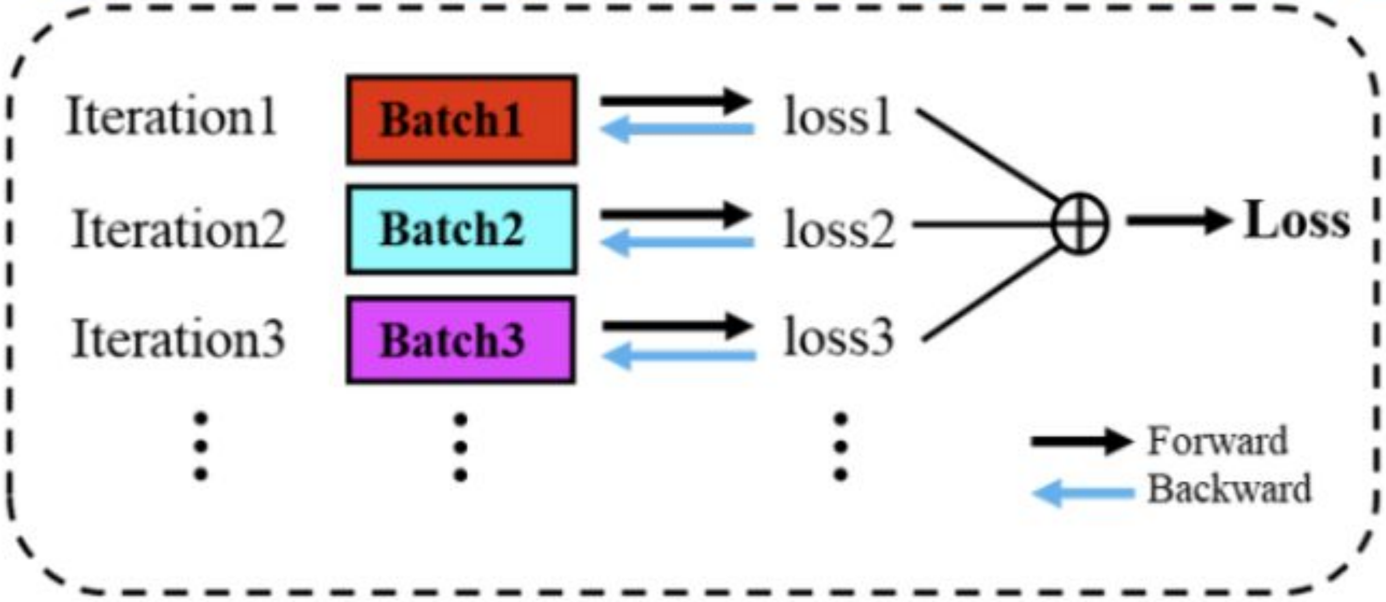
$$\mathrm{DSBN}\{D_i\}(X_i;\gamma_i,\beta_i) = \gamma_i \hat{X}_i + \beta_i \quad (1)$$

where,

$$\hat{X}_i = \frac{X_i - \mu_i}{\sqrt{\sigma_i^2 + \epsilon}}$$
$$\mu_i = \frac{1}{B}\sum_{j=1}^{B} X_i^j,\ \sigma_i^2 = \frac{1}{B}\sum_{j=1}^{B}(X_i^j - \mu_i)^2 \quad (2)$$

Dataset Alteration Training (DAT)

- Build each batch using only one of the training datasets.
    a. Accumulate the losses all datasets and update them all at the same time.



Dataset Alternation Training (DAT)

Implementation Details

- PSPNet with ResNet-18 and ResNet-101
- SGD
- Polynomial learning rate decay
- Random scale and horizontal flip
- Crop size of 512 x 512

Experimental Results

Method	Cityscapes (%)		BDD100K(%)
	Val.	Test	Val.
Single-dataset	67.52	67.75	53.88
Finetuning	67.79	66.52	58.30
Label remapping	66.23	66.39	58.74
CDCL (Ours)	**72.63**	**71.55**	**60.47**

(a) Cityscapes + BDD100K

Method	Cityscapes (%)		CamVid (%)	
	Val.	Test	Val.	Test
Single-dataset	67.52	67.75	73.05	70.41
Finetuning	67.35	67.87	74.83	71.16
Label remapping	67.13	68.22	78.03	76.86
CDCL (Ours)	**69.77**	**68.56**	**78.45**	**77.34**

(b) Cityscapes + CamVid

Method	Cityscapes (%)		COCO (%)
	Val.	Test	Val.
Single-dataset	67.52	67.75	32.86
Finetuning	70.44	70.23	32.10
Label remapping	50.35	52.39	32.67
CDCL (Ours)	**72.63**	**72.52**	**32.87**

(c) Cityscapes + COCO

Method	GFLOPs	Cityscapes (%)
Current state-of-the-art results		
SegNet [2]	286.0	56.10
ENet [27]	7.6	58.30
ESPNetv2 [26]	5.4	65.10
ESPNet [25]	8.9	60.30
ERFNet [29]	25.6	68.00
FCN-8s [24]	1335.6	65.30
RefineNet [22]	2102.8	73.60
Results w/o and w/ our scheme		
PSPNet (ResNet-18) [23]	512.8	67.60
PSPNet (ResNet-18)† [23]	512.8	71.40
PSPNet (ResNet-18) (Ours)	512.8	71.00
PSPNet (ResNet-18) (Ours)‡	1730.7	72.52
PSPNet (ResNet-101) [23]	2299.8	77.60
PSPNet (ResNet-101) (Ours)*	2299.8	78.74
PSPNet (ResNet-101)*‡	7762.0	78.40
PSPNet (ResNet-101) (Ours)*‡	7762.0	**79.73**

Table 6: Comparisons with the state-of-the-art methods on the Cityscapes test set. $*$: refers training with random crop of 769×769. †: refers using knowledge distillation method in [23]. ‡: refers testing with multiple scales.

Method	Cityscapes (%)		CamVid (%)		BDD100K (%)
	Validation	Test	Validation	Test	Validation
Single-dataset	67.52	67.75	73.05	70.41	53.88
CDCL (Ours)	73.17 **(+5.65)**	70.98 **(+3.23)**	78.84 **(+5.79)**	75.52 **(+5.11)**	60.45 **(+6.57)**

Table 2: Performance comparisons between the single-dataset baseline and our method with the same ResNet-18 backbone in the three-dataset setting.

Ablation Studies

Method	Norm	DAT	Cityscapes	BDD100K
Single-dataset	BN		67.75%	53.88%
Cross-dataset	BN		62.10%	53.34%
Cross-dataset	BN	✓	64.69%	56.33%
Cross-dataset	DSBN		68.55%	58.93%
CDCL (Ours)	DSBN	✓	**72.63%**	**60.47%**

Table 3: Ablation studies on DSBN and DAT with the ResNet-18 backbone on the Cityscapes and BDD100K validation sets.

Iteration interval	Cityscapes (%)	BDD100K (%)
$t = 1$	72.63	60.47
$t = 2$	68.88	59.17
$t = 3$	65.11	57.79
$t = 4$	62.86	57.21
$t = 5$	61.12	56.06

Table 4: Effect of iteration interval t in DAT on the Cityscapes and BDD100K validation sets.

Method	Cityscapes (%)		BDD100K (%)
	Val.	Test	Val.
Single-dataset	73.51	74.45	57.47
CDCL (Ours)	**75.83**	**75.95**	**63.84**

(a) Cityscapes + BDD100K

Method	Cityscapes (%)		CamVid (%)	
	Val.	Test	Val.	Test
Single-dataset	73.51	74.45	75.86	74.72
CDCL (Ours)	**75.00**	**74.77**	**81.15**	**79.33**

(b) Cityscapes + CamVid

Table 5: Performance comparisons between the single-dataset baseline and our method with the same ResNet-101 backbone in the two-dataset setting.

Qualitative Results

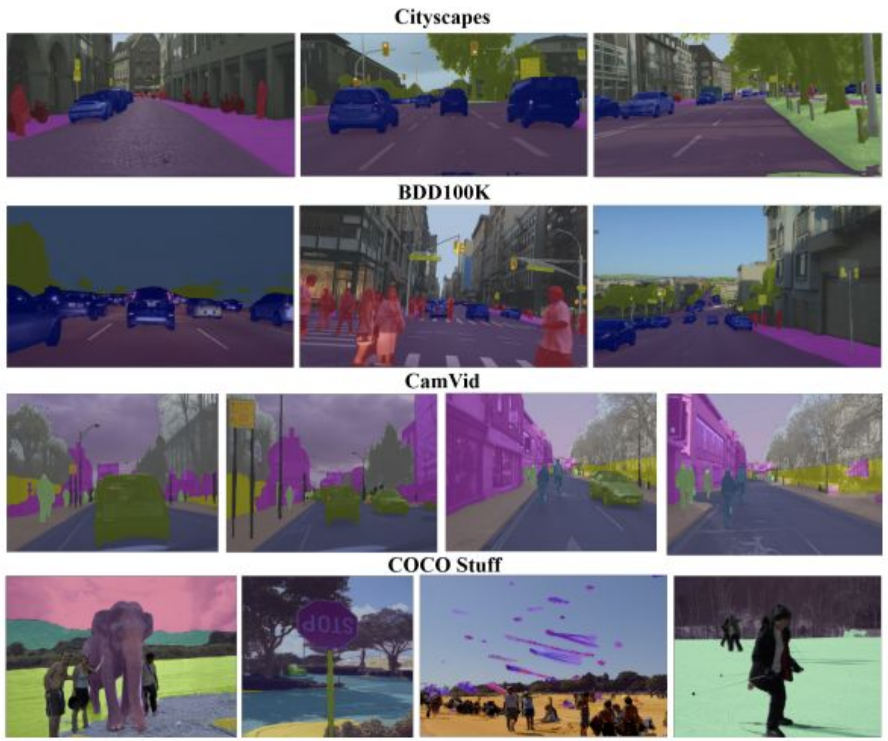


Figure 5: Qualitative results on four datasets, produced from PSPNet (ResNet-18) with our provided method.

Limitations we can think of...

- Having similar distribution of L2-norm of convolution filters over channel index doesn't guarantee the similar convolution weights.

- How can we deal with the totally unseen datasets?
    a. Which batch normalization layer should be used?

- Usually, datasets have different number of samples.
    a. Better way to balance the number of samples of each dataset?

Thank you!