# Semi-Supervised Semantic Segmentation with Cross-Consistency Training

Yassine Ouali, Céline Hudelot, Myriam Tami

Université Paris-Saclay, CentraleSupélec, MICS, 91190, Gif-sur-Yvette, France.

# Objectives

Semantic segmentation methods rely heavily on large annotated datasets.

- Object Segmentation is an extremely time consuming task
- COCO dataset, it requires over 22 worker hours per 1,000  segmentations.

Objectives:

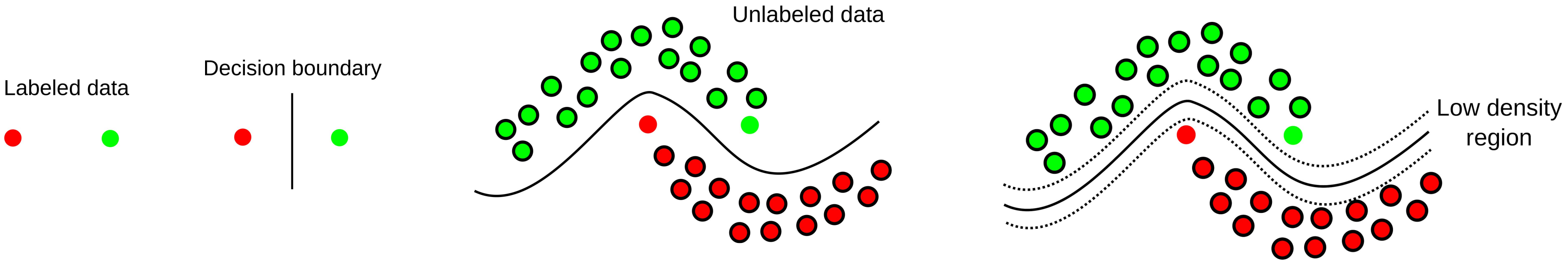
- Proposing data efficient & performing semantic segmentation.
- Leveraging the large amount of easily available unlabeled data.

**➙ We propose a novel semi-supervised method for semantic segmentation based on consistency training.**

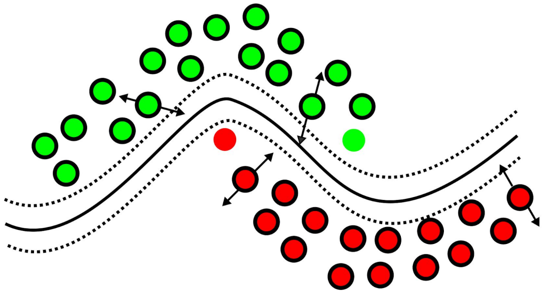
# Cluster Assumption

Applying Consistency training (CT) in semantic segmentation is not straightforward: even when impressive results were obtained with CT on semi-supervised image classification, the adoption of such methods in semantic segmentation is not as straight forward.

The Cluster Assumption: « **If points are in the same cluster, they are likely to be of the same class. »**




Consistency Training: « **if a realistic perturbation was applied to the unlabeled data point, the prediction should not change.** »



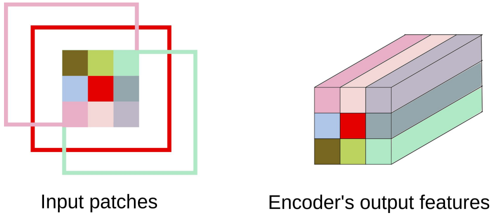
$$f(x_u) = f(x_u + \epsilon)$$

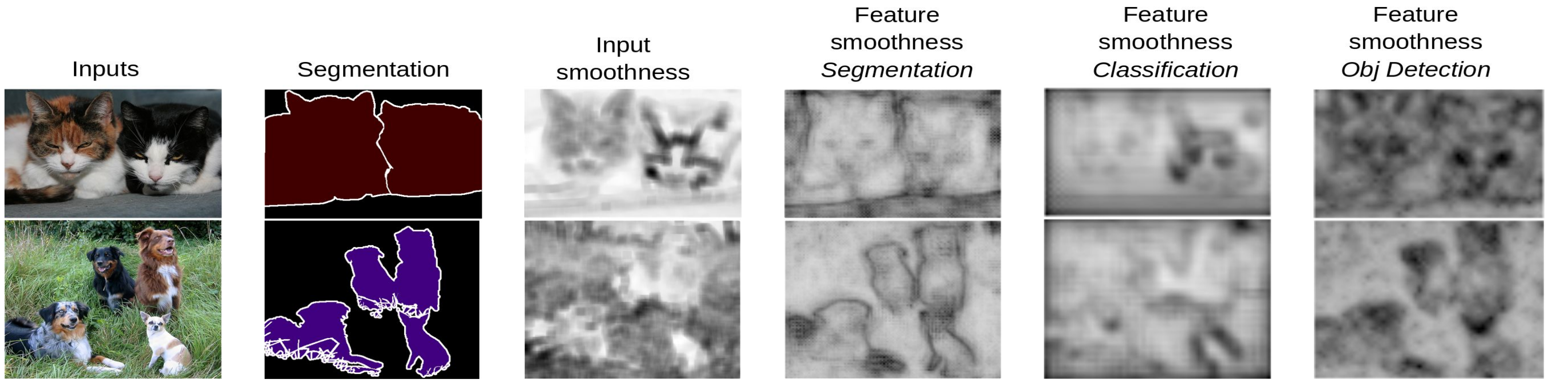
**Do we have the same behaviour at the pixel level for semantic segmentation ?**

# Cluster assumption in Semantic Segmentation

At the pixel level, the value of the neighboring patches varies smoothly even when the class of the pixel changes.

To illustrate this, we compute the local smoothness:



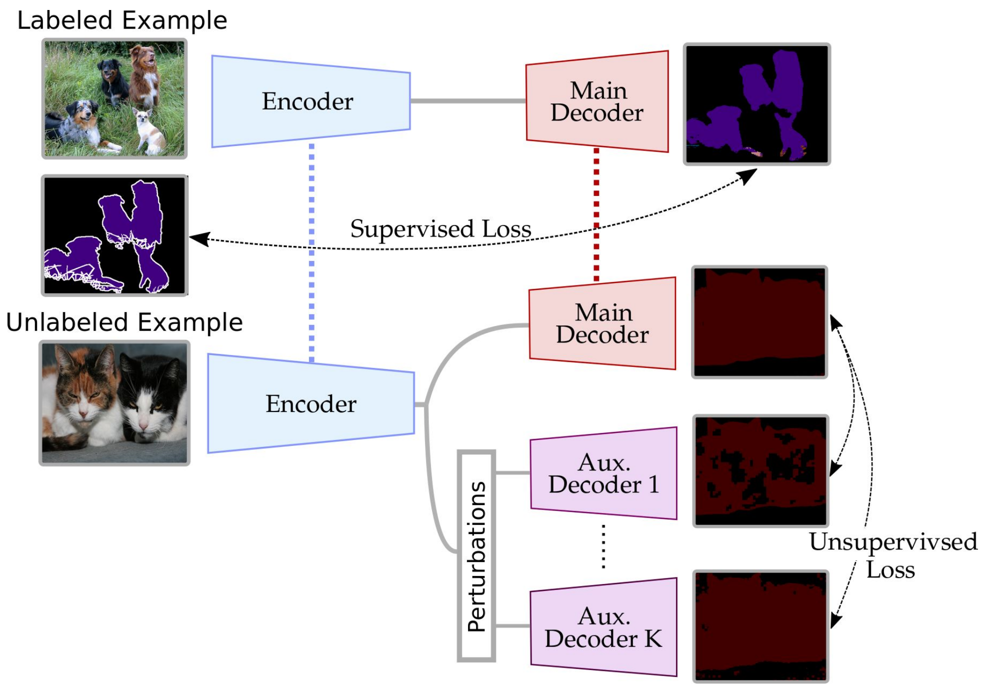

**The cluster assumption is violated at the input level but is maintained at the feature level.**

**→ Enforce the consistency over the encoder's outputs.**

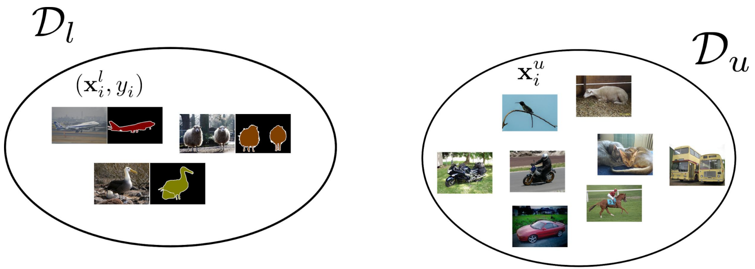
# Cross-Consistency Training (CCT)

Proposed method:

**➞ Cross-consistency training: enforce consistency of predictions on the unlabeled data over the features rather than the inputs.**

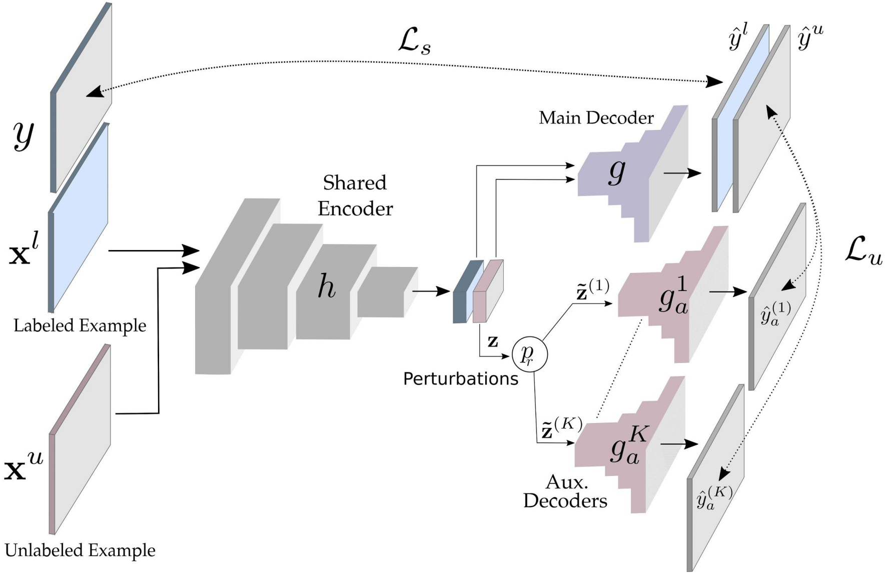


# Cross-Consistency Training (CCT)



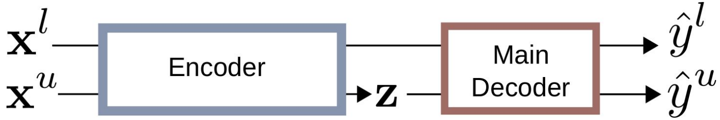

$|\mathcal{D}_u| = m \quad |\mathcal{D}_l| = n \qquad m >> n$

**Illustration of CCT**

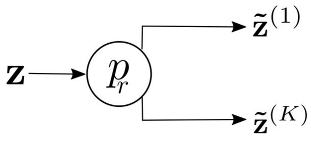



**Training:**

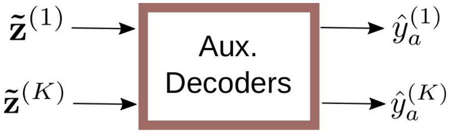
1- Forward both labeled and unlabeled images through the encoder & main decoder:




2- Apply *K* perturbations to the encoder's output:




3- Compute the aux. predictions:




4- Compute the supervised and unsupervised losses:

$$\boxed{\mathcal{L}_s = \frac{1}{|\mathcal{D}_l|} \sum_{\mathbf{x}_i^l, y_i \in \mathcal{D}_l} \mathbf{H}(y_i, f(\mathbf{x}_i^l))}$$

$$\boxed{\mathcal{L}_u = \frac{1}{|\mathcal{D}_u|} \frac{1}{K} \sum_{\mathbf{x}_i^u \in \mathcal{D}_u} \sum_{k=1}^{K} \mathbf{d}(g(\mathbf{z}_i), g_a^k(\mathbf{z}_i))}$$

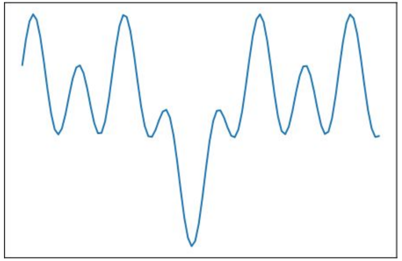
$$\boxed{\mathcal{L} = \mathcal{L}_\mathrm{s} + \omega_u \mathcal{L}_\mathrm{u}}$$

# Perturbations

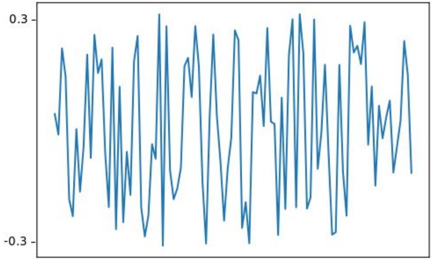
We define 3 types of perturbations: **feature based**, **prediction based** and **random perturbations**.

- ***Feature noise*** (F-noise)

Activations $\mathbf{z}$

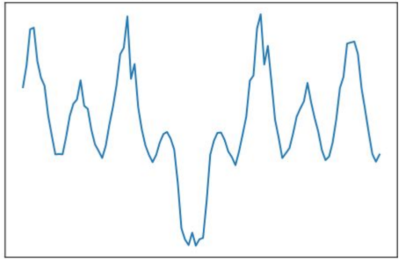


Uniform noise
$\mathbf{N} \sim \mathcal{U}(-0.3, 0.3)$



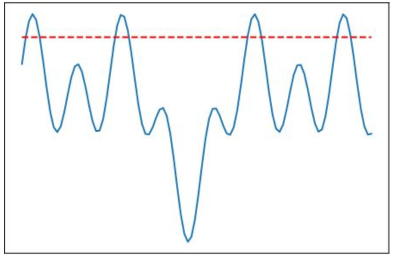

Adjs the amplitude
$\mathbf{z} \odot \mathbf{N}$


Perturbed activations
$\tilde{\mathbf{z}} = (\mathbf{z} \odot \mathbf{N}) + \mathbf{z}$

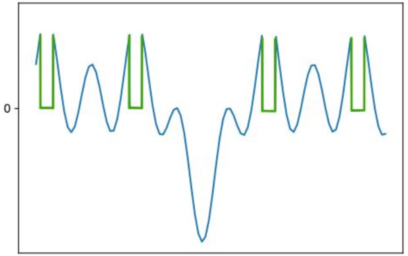


- ***Feature drop*** (F-drop)

Activations $\mathbf{z}$
& Threshold $\gamma \sim \mathcal{U}(0.6, 0.9)$

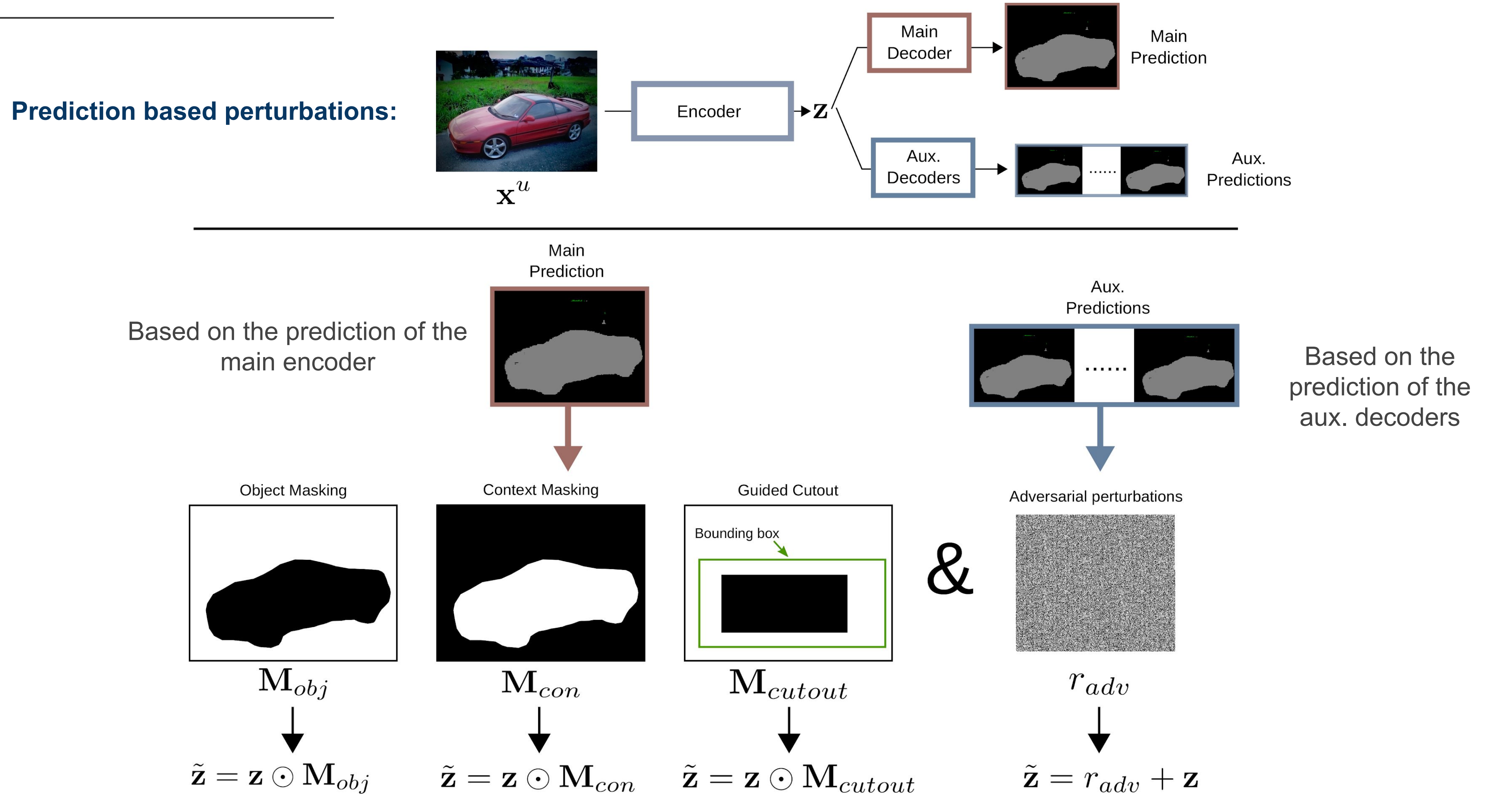


Perturbed activations
$\tilde{\mathbf{z}} = \mathbf{z} \odot \mathbf{M}_{\text{drop}}$



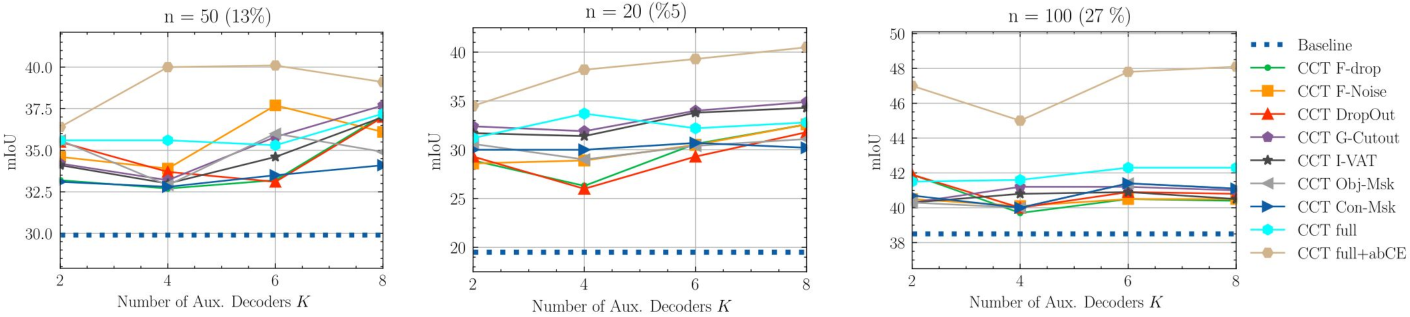

- ***Random perturbations*** (DropOut): simple spatial dropout.

# Perturbations

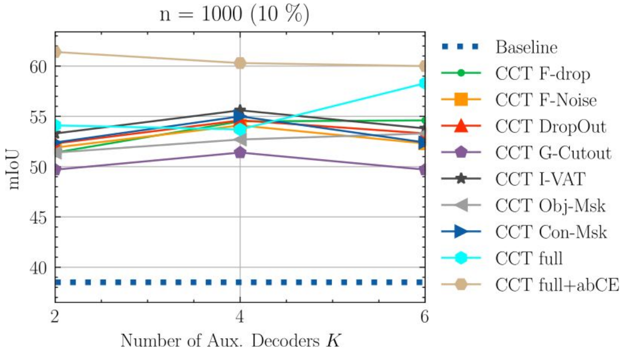


# Results

## Cam-Vid




## Pascal-Voc



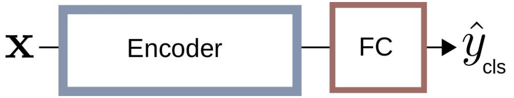

| Method | Pixel-level Labeled Examples | Image-level Labeled Examples | Val |
|---|---|---|---|
| WSSL [37] | 1.5k | 9k | 64.6 |
| GAIN [31] | 1.5k | 9k | 60.5 |
| MDC [51] | 1.5k | 9k | 65.7 |
| DSRG [22] | 1.5k | 9k | 64.3 |
| Souly *et al.* [47] | 1.5k | 9k | 65.8 |
| FickleNet [30] | 1.5k | 9k | 65.8 |
| Souly *et al.* [47] | 1.5k | - | 64.1 |
| Hung *et al.* [23] | 1.5k | - | 68.4 |
| CCT | 1k | - | 64.0 |
| CCT | 1.5k | - | 69.4 |

The results confirm the effectiveness of enforcing the consistency over the hidden representations for semantic segmentation and highlight the versatility of CCT and its success with numerous perturbations.

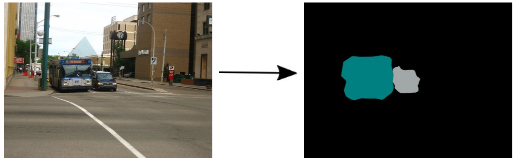
# Using image-level labels

**Generate pseudo pixel-level labels from image level labels.**

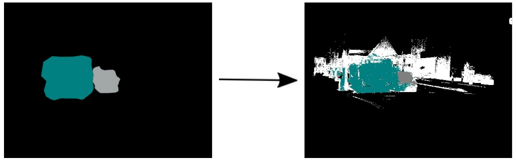
1- Train the encoder for image classification:




2- Use the trained classifier to generate class activation maps M:



3- Considering only positions with high confidence & Applying a CRF preprocessing:




**Train the aux. decoders using the generated pseudo labels**

$$\mathcal{L}_w = \frac{1}{|\mathcal{D}_w|}\frac{1}{K}\sum_{\mathbf{x}_i^w \in \mathcal{D}_w}\sum_{k=1}^{K}\mathbf{H}(y_p, g_a^k(\mathbf{z}_i))$$

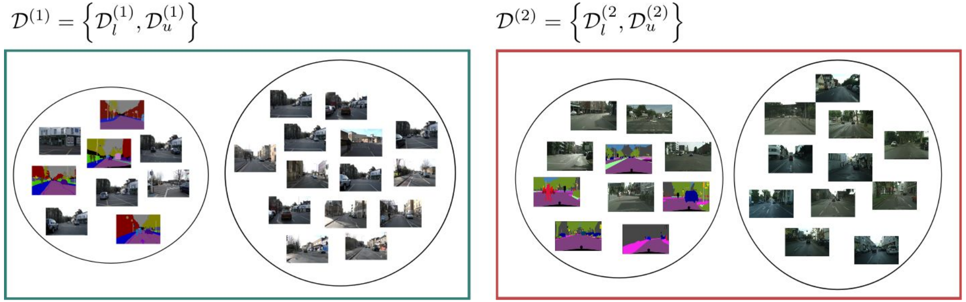
$$\mathcal{L} = \mathcal{L}_s + \omega_u \mathcal{L}_u + \omega_w \mathcal{L}_w$$

**Results**

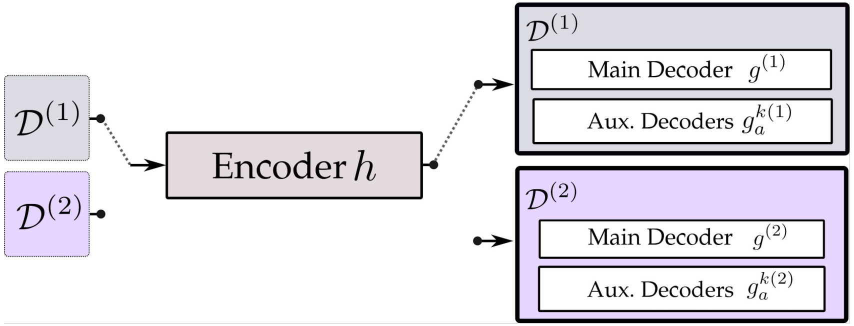
| Method | Pixel-level Labeled Examples | Image-level Labeled Examples | Val |
|---|---|---|---|
| CCT | 1k | - | 64.0 |
| CCT | 1.5k | - | 69.4 |
| CCT | 1.5k | 9k | **73.2** |

# CCT on multiple domains

CCT can be easily extended to multiple domains with partially or fully non-overlapping label spaces.




Train a shared encoder on both domains & enforce consistency of predictions on both domains.




**Cityscapes + SUN RGB-D**

| Method | Labeled Examples | CS | SUN | Avg. |
|---|---|---|---|---|
| SceneNet [34] | Full (5.3k) | - | 49.8 | - |
| Kalluri, *et al.* [24] | 1.5k | 58.0 | 31.5 | 44.8 |
| Baseline | 1.5k | 54.3 | 38.1 | 46.2 |
| CCT | 1.5k | 58.8 | 45.5 | **52.1** |

**Cityscapes + CamVid**

| Method | n=50 | | | n=100 | | |
|---|---|---|---|---|---|---|
| | CS | CVD | Avg. | CS | CVD | Avg. |
| Kalluri, *et al.* [24] | 34.0 | 53.2 | 43.6 | 41.0 | 54.6 | 47.8 |
| Baseline | 31.2 | 40.0 | 35.6 | 37.3 | 34.4 | 35.9 |
| CCT | 35.0 | 53.7 | **44.4** | 40.1 | 55.7 | **47.9** |

# Conclusion

**We presented the following main-contributions:**

**(1) Consistency Training for semantic segmentation.**

We observed that for semantic segmentation, due to the dense nature of the task, the cluster assumption is more easily enforced over the hidden representations rather than the inputs.

**(2) Cross-Consistency Training.**

We proposed CCT (Cross-Consistency Training) for semi-supervised semantic segmentation, where we define several novel perturbations, and show the effectiveness of enforcing consistency over the encoder outputs rather than the inputs.

**(3) Using weak-labels and pixel-level labels from multiple domains.**

The proposed method is quite simple and flexible, and can easily be extended to use image-level labels and pixel-level labels from multiple-domains.

**(4) Competitive results.**

We showed competitive results on several  semantic segmentation benchmarks.

Yassine Ouali, Céline Hudelot, Myriam Tami

# Thank you

For more details, please visit the project's webpage