Semi-Supervised Semantic Segmentation with Cross-Consistency Training

Yassine Ouali, Céline Hudelot, Myriam Tami
Université Paris-Saclay, CentraleSupélec, MICS, 91190, Gif-sur-Yvette, France

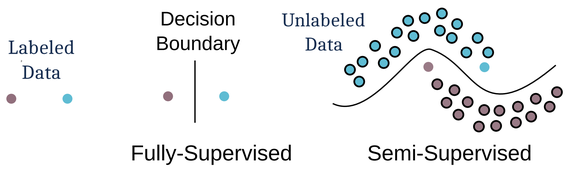
Project's Webpage

Motivation

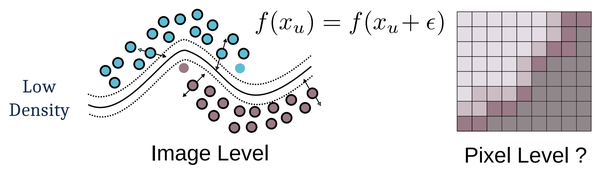
- The success of **semantic segmentation** relies heavily on the availability of large annotated datasets.
- **Semi-Supervised Learning** (SSL) with consistency training is appealing but still confined to classification.

Goal: Adapting Consistency Training for semantic segmentation & leveraging unlabeled data.

Semi-Supervised Learning (SSL)

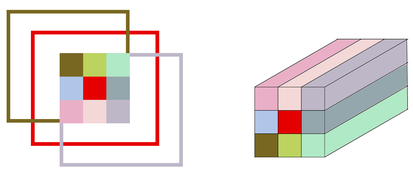


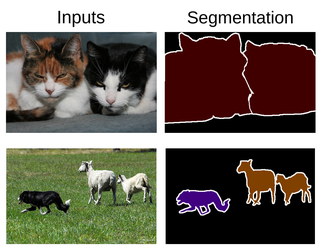
Cluster Assumption



Is the cluster assumption maintained at the pixel-level ? Pixel values may vary smoothly even when the class changes.

Cluster Assumption at the pixel level: Can be investigated by computing 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.

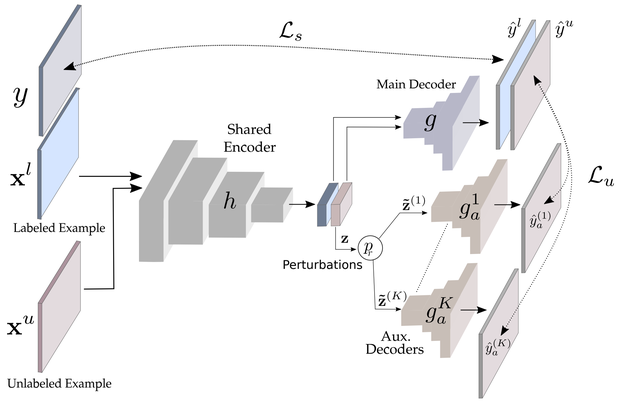
Cross-Consistency Training (CCT)

Problem Formulation:

Objective: exploit the unlabeled set $\mathcal{D}_u$ together with a smaller labeled set $\mathcal{D}_l$ to train a segmentation network f to perform well on test data. In CCT, we define:

- A segmentation network $f = g \circ h$ with an encoder h and a main decoder g and K auxiliary decoders g_a^k, $k \in [1, K]$.
- r perturbation functions p_r to be applied to the encoder's output $\mathbf{z}$ corresp. to an unlabeled example x^u.

Main idea: Enforce a consistency of predictions on $\mathcal{D}_u$ between the output of the main decoder $f(\mathbf{x}_i^u)$, considered as target, and that of the aux. decoders $g_a^k(\tilde{\mathbf{z}}_i)$ over various perturbations applied to $\mathbf{z}$ to get $\tilde{\mathbf{z}}$.



Training:

- Forward both labeled $\mathbf{x}_l$ and unlabeled $\mathbf{x}_u$ images through the encoder & main decoder.
- Apply K perturbations to the encoder's output $\mathbf{z}$.
- Compute the aux. predictions $g_a^k(\tilde{\mathbf{z}})$.
- Compute the loss $\mathcal{L}$ using the supervised $\mathcal{L}_s$ and unsupervised $\mathcal{L}_u$ losses: $\mathcal{L} = \mathcal{L}_s + \omega_u \mathcal{L}_u$ with:

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

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

Inference: Only the seg. network f is used.

Perturbations

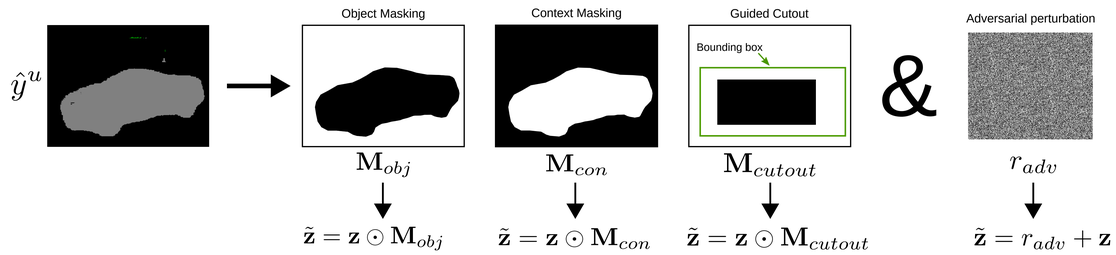
The following perturbations are to be applied to the encoder's output $\mathbf{z}$:

Feature Based. They consist of either injecting noise into or dropping some of the activations of $\mathbf{z}$.

- A uniform tensor noise $\mathbf{N} \sim \mathcal{U}(-0.3, 0.3)$ is injected into $\mathbf{z}$ after adjusting its amplitude: $\tilde{\mathbf{z}} = (\mathbf{z} \odot \mathbf{N}) + \mathbf{z}$.
- The maximally activated features are dropped using a threshold γ: $\tilde{\mathbf{z}} = (\mathbf{z} \odot \mathbf{M}_{\text{drop}})$ with $\mathbf{M}_{\text{drop}} = \{\mathbf{z} < \gamma\}_1$.

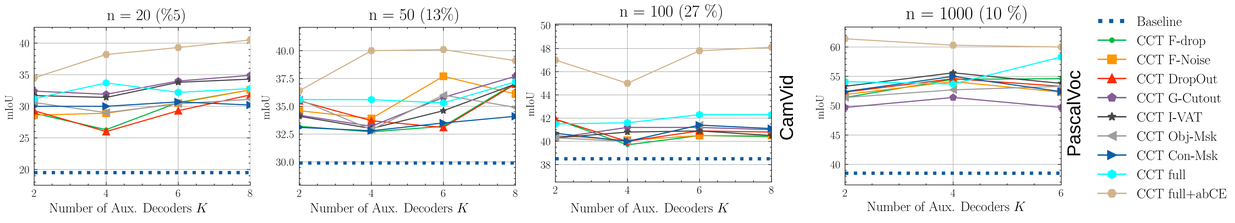
Random. A simple application of spatial dropout to the activations $\mathbf{z}$.

Prediction Based. Using the predictions of the main and aux. decoders, we generate masks $\mathbf{M}_*$ and adversarial noise to be applied to $\mathbf{z}$.



Experiments & Results

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



Results. CCT outperforms previous works relying on the same level of supervision and even methods which exploit image-level labels. We also obtain impressive results when using with image-level labels and when training on multiple domain confirming the flexibility of CCT.

PascalVoc

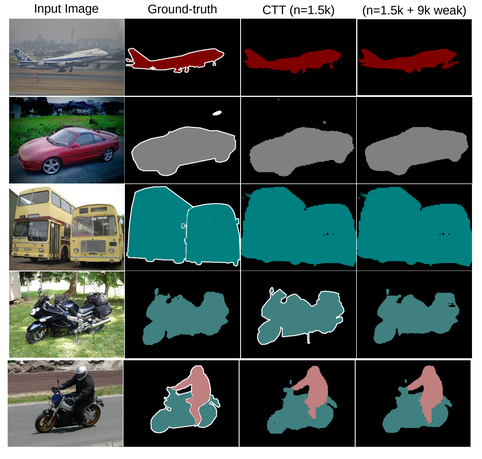
Method	Pixel-level Labeled Examples	Image-level Labeled Examples	Val
WSSL [37]	1.5k	9k	64.6
GAIN [31]	1.5k	9k	60.5
MDC [51]	1.5k	9k	65.7
DSRG [22]	1.5k	9k	64.3
Souly et al. [47]	1.5k	9k	65.8
FickleNet [30]	1.5k	9k	65.8
Souly et al. [47]	1.5k	-	64.1
Hung et al. [23]	1.5k	-	68.4
CCT	1k	-	64.0
CCT	1.5k	-	69.4
CCT	1.5k	9k	73.2

CityScapes + SUB-RGBD

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

In this work, we: **(1) investigate the cluster assumption in semantic segmentation; (2) propose CCT where we enforce the consistency over the encoder's outputs rather than the inputs; (3) extend CCT to use weak-labels and pixel-level labels from multiple domains.** *For more details, please see the paper & code.*