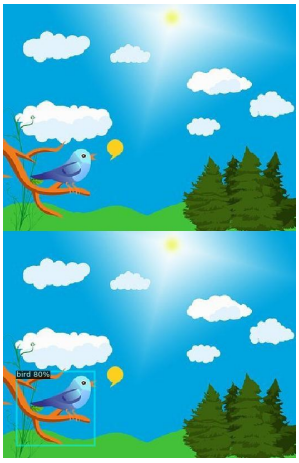
# Contrastive Mean Teacher for Domain Adaptive Object Detectors Supplementary Material

Shengcao Cao[1] Dhiraj Joshi[2] Liang-Yan Gui[1] Yu-Xiong Wang[1]
[1]University of Illinois at Urbana-Champaign [2]IBM Research

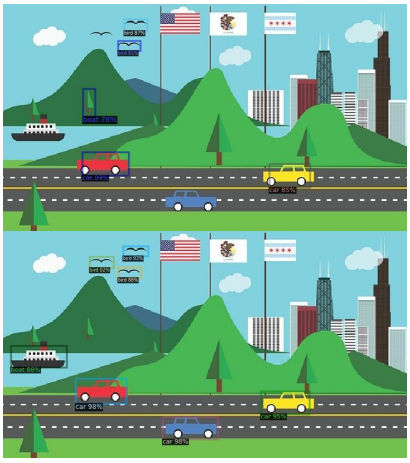
## A. Additional Visualization Results

We present more high-resolution visualization to compare the baseline Adaptive Teacher (AT) [1] and our AT + CMT qualitatively on the Pascal VOC → Clipart1k benchmark in Figure 1, and on the Cityscapes → Foggy Cityscapes benchmark in Figure 2. Each pair of images show results by AT (**top**) and AT + CMT (**bottom**).

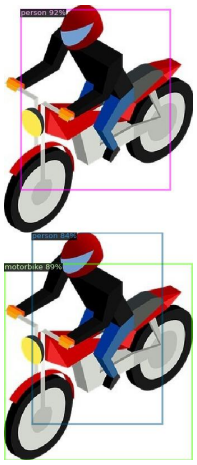


(a) CMT detects the bird missed by the baseline.

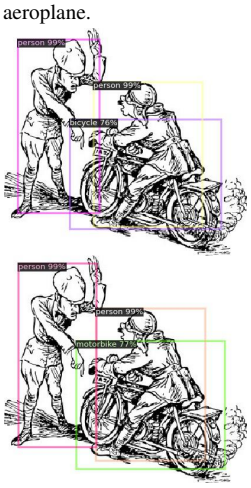
(b) CMT avoids detecting the moon and star as bird and aeroplane.



(c) CMT detects the boat and birds missed by the baseline.



(d) CMT detects the motorbike missed by the baseline.



(e) CMT fixes the mis-classification of the motorbike.

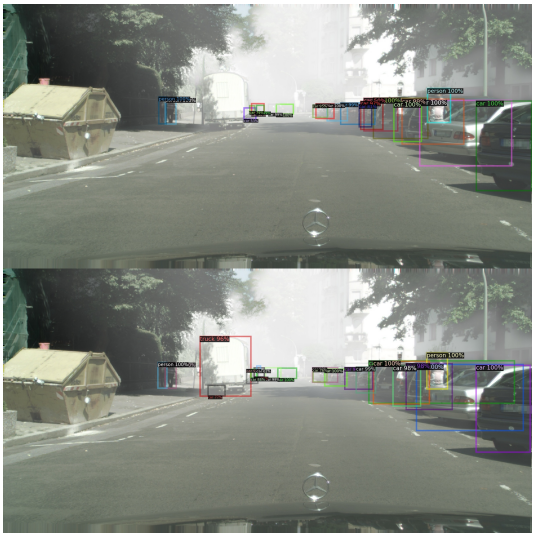
(f) CMT detects the persons missed by the baseline.

Figure 1. **Additional qualitative results from Clipart1k.** In the visualized images, AT (**top** in each sub-figure) makes some incorrect predictions, while AT + CMT (**bottom** in each sub-figure) can correct them.

(a) CMT detects the car in the mirror and distant persons.

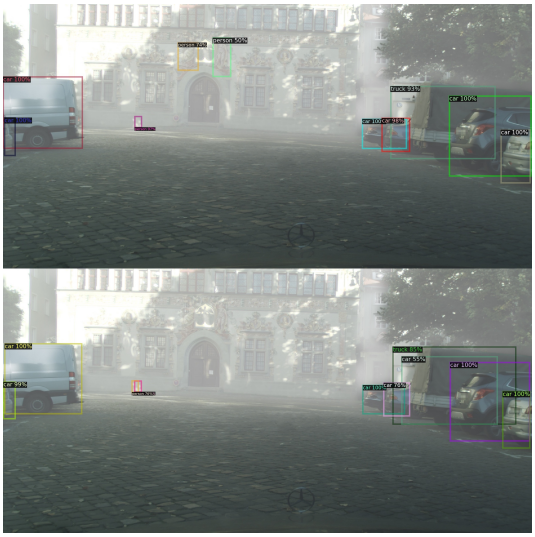
(b) CMT fixes the mis-classification of the train.

(c) CMT provides a more precise bounding box of the train.



(d) CMT detects the truck missed by the baseline.

(e) CMT fixes the mis-classification of the bus.



(f) CMT avoids detecting the sculptures as real persons.

Figure 2. **Additional qualitative results from Foggy Cityscapes.** In the visualized images, AT (**top** in each sub-figure) makes some incorrect predictions, while AT + CMT (**bottom** in each sub-figure) can correct them.

## B. Pseudo-code for Contrastive Mean Teacher

In Algorithm 1, we present a pseudo-code outline of our Contrastive Mean Teacher (CMT) framework. The key difference between CMT and a traditional Mean Teacher (MT) [2] framework is highlighted.

**Algorithm 1:** Contrastive Mean Teacher

**Input:** Object detectors: Student $D(\cdot;\theta^{\mathcal{Q}})$ and Teacher $D(\cdot;\theta^{\mathcal{K}})$. For consistency with prior work, we use letter $\mathcal{Q}$ for student-related variables, and $\mathcal{K}$ for teacher-related ones. We denote their feature extraction modules as $f(\cdot;\theta^{\mathcal{Q}})$ and $f(\cdot;\theta^{\mathcal{K}})$. Hyper-parameters: Momentum $\alpha$ in exponential moving average (EMA), pseudo-label score threshold $\gamma$, temperature $\tau$ in contrastive loss, loss weights $\lambda_{\text{contrast}}, \lambda_{\text{unsup\_det}}, \lambda_{\text{sup\_det}}$, and learning rate $\eta$.

**Output:** Student $D(\cdot;\theta^{\mathcal{Q}})$ and Teacher $D(\cdot;\theta^{\mathcal{K}})$ after unsupervised domain adaptation.

**for** $iteration \leftarrow 1$ **to** $T_{max\_iterations}$ **do**

`// 1. Load data mini-batch`

Get batch of source-domain labeled images $\mathcal{I}^{\text{labeled}}$, corresponding bounding boxes $\mathcal{B}^{\text{labeled}}$, and classes $\mathcal{C}^{\text{labeled}}$

Get batch of target-domain unlabeled images $\mathcal{I}^{\text{unlabeled}}$

Student's strong augmentation: $\mathcal{I}^{\text{labeled},\mathcal{Q}} = t^{\mathcal{Q}}(\mathcal{I}^{\text{labeled}}), \mathcal{I}^{\text{unlabeled},\mathcal{Q}} = t^{\mathcal{Q}}(\mathcal{I}^{\text{unlabeled}})$

Teacher's weak augmentation: $\mathcal{I}^{\text{labeled},\mathcal{K}} = t^{\mathcal{K}}(\mathcal{I}^{\text{labeled}}), \mathcal{I}^{\text{unlabeled},\mathcal{K}} = t^{\mathcal{K}}(\mathcal{I}^{\text{unlabeled}})$

`// 2. Update Teacher`

Update Teacher by EMA: $\theta^{\mathcal{K}} = \alpha\theta^{\mathcal{K}} + (1-\alpha)\theta^{\mathcal{Q}}$

`// 3. Pseudo-label`

Generate pseudo-labels with Teacher detector: $\mathcal{B}^{\text{unlabeled}}, \mathcal{C}^{\text{unlabeled}} = \text{Filter}\left(D(\mathcal{I}^{\text{unlabeled},\mathcal{K}};\theta^{\mathcal{K}}), \gamma\right)$

`// 4. Compute multi-scale feature maps`

`// In practical implementation, feature maps are obtained from forward passes needed for pseudo-labels and unsupervised detection loss, so there is no computation overhead.`

Compute Student's features: $\mathcal{F}^{\mathcal{Q}} = f(\mathcal{I}^{\text{unlabeled},\mathcal{Q}};\theta^{\mathcal{Q}})$

Compute Teacher's features: $\mathcal{F}^{\mathcal{K}} = f(\mathcal{I}^{\text{unlabeled},\mathcal{K}};\theta^{\mathcal{K}})$

`// 5. Unsupervised branch: object-level contrastive loss`

(Optional) Post-processing pseudo-labels: $\mathcal{B}^{\text{unlabeled}}, \mathcal{C}^{\text{unlabeled}} = \text{PostProc}(\mathcal{B}^{\text{unlabeled}}, \mathcal{C}^{\text{unlabeled}})$

Get number of objects: $N = \text{len}(\mathcal{B}^{\text{unlabeled}})$

`// Each level of multi-scale features`

**for** $k \leftarrow 1$ **to** $K_{max\_levels}$ **do**

`// Each object`

**for** $i \leftarrow 1$ **to** $N$ **do**

Locate Student's object-level features: $z^{\mathcal{Q}}_{k,i} = \text{Normalize}(\text{ROIAlign}(F^{\mathcal{Q}}_k, B^{\text{unlabeled}}_i))$

Locate Teacher's object-level features: $z^{\mathcal{K}}_{k,i} = \text{Normalize}(\text{ROIAlign}(F^{\mathcal{K}}_k, B^{\text{unlabeled}}_i))$

Compute contrastive loss according to Equation 4:

$$L_{\text{contrast},k} = \mathcal{L}_{\text{contrast}}\left(\{z^{\mathcal{Q}}_{k,1},\ldots,z^{\mathcal{Q}}_{k,N}\}, \{z^{\mathcal{K}}_{k,1},\ldots,z^{\mathcal{K}}_{k,N}\}, \{C^{\text{unlabeled}}_1,\ldots,C^{\text{unlabeled}}_N\}, \tau\right)$$

Compute total contrastive loss: $L_{\text{contrast}} = \sum_{k=1}^{K_{\text{max\_levels}}} L_{\text{contrast},k}$

`// 6. Unsupervised branch: detection loss`

Compute unsupervised detection loss: $L_{\text{unsup\_det}} = \mathcal{L}_{\text{det}}(D(\mathcal{I}^{\text{unlabeled},\mathcal{Q}};\theta^{\mathcal{Q}}), \mathcal{B}^{\text{unlabeled}}, \mathcal{C}^{\text{unlabeled}})$

`// 7. Supervised branch: detection loss`

Compute supervised detection loss: $L_{\text{sup\_det}} = \mathcal{L}_{\text{det}}(D(\mathcal{I}^{\text{labeled},\mathcal{Q}};\theta^{\mathcal{Q}}), \mathcal{B}^{\text{labeled}}, \mathcal{C}^{\text{labeled}})$

`// 8. Optimize`

Compute total loss: $L = \lambda_{\text{contrast}}L_{\text{contrast}} + \lambda_{\text{unsup\_det}}L_{\text{unsup\_det}} + \lambda_{\text{sup\_det}}L_{\text{sup\_det}}$

Take SGD step: $\theta^{\mathcal{Q}} = \theta^{\mathcal{Q}} - \eta\nabla_{\theta^{\mathcal{Q}}}L$

## References


[1] Yu-Jhe Li, Xiaoliang Dai, Chih-Yao Ma, Yen-Cheng Liu, Kan Chen, Bichen Wu, Zijian He, Kris Kitani, and Peter Vajda. Cross-domain adaptive teacher for object detection. In *CVPR*, 2022.

[2] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017.