Supplementary Material

A. Negative transfer decision boundaries visualization

To demonstrate the effectiveness of MUST to reduce negative transfer, we generate an example such that the source data can be perfectly classified by either of the two features. The target data is agnostic to one of the features and can be perfectly classified using the other. Figure 5 shows the decision boundaries of source-only model and MUST for 20 initializations. Source-only model is trained using only the source data. Its decision boundaries are a linear combination of the two features, which heart the performance on the target. MUST teacher successfully learns to classify using the feature that is relevant to both source and target domains.

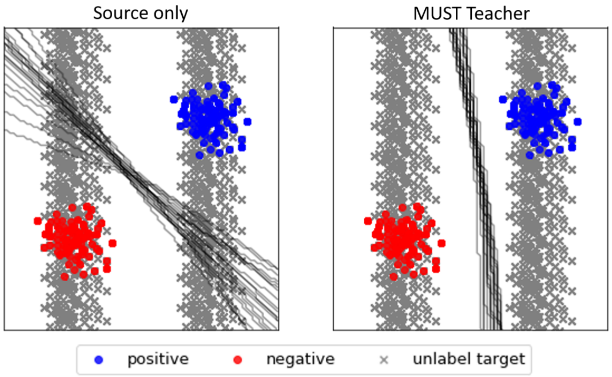

Figure 5. Decision boundaries of source-only models and MUST for 20 initializations. Blue: positive source samples. Red: negative source samples. Gray: unlabeled target samples. The source-only model classifies perfectly the source data but uses the horizontal feature that is not relevant to the target domain. The MUST teacher learns to ignore the horizontal feature and avoiding negative transfer.

B. Regression experiment

The common MSDA banchmarks are for classification task. We evaluated MUST also for regression, using WebCamT dataset for vehicle counting task. The dataset contain cameras from different geographic locations and each camera see an intersection or a road from different perspectives. Each camera is consider as a domain.

The data: WebCamT is a public dataset for vehicle counting from large-scale city camera videos, which has low resolution (352 × 240), low frame rate (1 frame/second), and high occlusion. It has 60,000 frames annotated with vehicle bounding box and count, divided into training set (42,200 frames) and testing set (17,800 frames).

Target camera	FCN (Zhang, 2017)	MUST (ours)
410	2.58	**2.42**
551	5.70	**4.85**
173	4.07	**4.04**
403	2.30	**2.28**
495	3.20	**3.06**
170	6.04	**5.76**
511	2.74	**2.73**
398	5.04	**4.99**
Avg. MAE	3.96	**3.77**

Table 9. MAE for vehicle counting

Experimental setup: We demonstrate the effectiveness of MUST to count vehicles from an unlabeled target camera by adapting from multiple labeled source cameras: we select 8 cameras located in different intersections of the city with different scenes, and each has more than 2,000 labeled images for our evaluations. Among these 8 cameras, we take one camera as the target camera, and use the other 7 cameras as sources. We calculated mean absolute error (MAE) between true count and estimated count.

Compared approaches: We compare our method with FCN [43], a basic network without domain adaptation.

Results: As shown in Table 9, MUST out-performs FCN, in all experiments.

C. Analysis of optimization dynamics

A closer look into the training process can help us understand the dynamic interplay of the two networks. A detailed analysis

Figure 6 traces the target accuracy of the student and teacher together with $L_{teacher}$ and $L_{student}$ during training. In addition, we plot the number of samples that cross the confidence threshold and the reverse validation score. Interestingly, learning follows through four phases.

A closer look into the training process can help us understand the dynamic interplay of the two networks. Figure 6 traces the target accuracy of the student and teacher together with $L_{teacher}$ and $L_{student}$ during training. In addition, we plot the number of samples that cross the confidence threshold and the reverse validation score. Interestingly, learning follows through four phases.

(1) Teacher learns, Iterations 1 – 1000: The teacher train on the source domains. The predictions of the teacher on the target domain are under the confidence threshold (C_{th}), so the student does not train ($\mathcal{L}_{student} = 0$), and the teacher optimize only $\mathcal{L}_{source}$.

(2) Sync, Iterations 1000 – 2000: The teacher confidence on target samples grows and the student starts receiving labels to train on ($\mathcal{L}_{student} > 0$). Surprisingly, there is no change in the student target accuracy, even though the teacher provides the student with good quality pseudo-labels to train on. There is a small drop in the teacher accu-

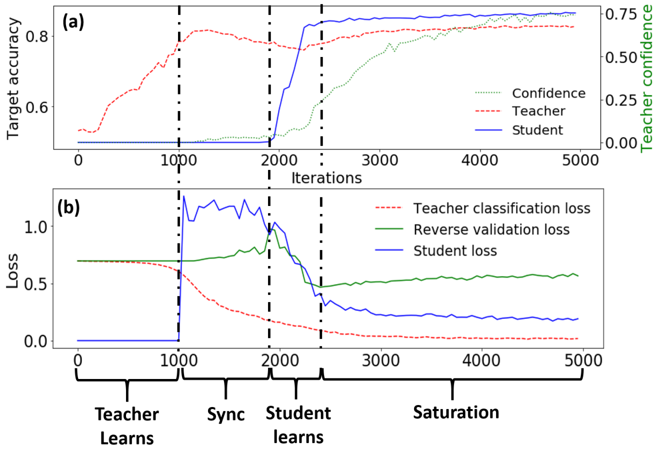

Figure 6. **The dynamics of teacher and student joint-learning**. (a) Target accuracy of the teacher (red) and the student (blue) during training and the percent of samples that passed the confidence threshold (green). (b) Loss functions values during training $L_{teacher}$ (red), $L_{student}$ (blue) and reverse validation loss (green).

racy. Since $L_{student}$ also regularize the teacher, the teacher now optimizes $\mathcal{L}_{teacher} = \mathcal{L}_{source} + \lambda \cdot \mathcal{L}_{student}$. Optimize $\mathcal{L}_{student}$ by the teacher create more consistent predictions, so the student can fit them better.

(3) Student learns, Iterations 2000 – 2500: Student accuracy improves quickly on the target data. The reverse validation loss decline, indicating that the student focuses on features that are relevant to both the source and target data.

(4) Saturation, Iterations 2500 – 5000: Networks reach saturation. The student accuracy is higher than the teacher, indicating that the effect of negative transfer is reduced.