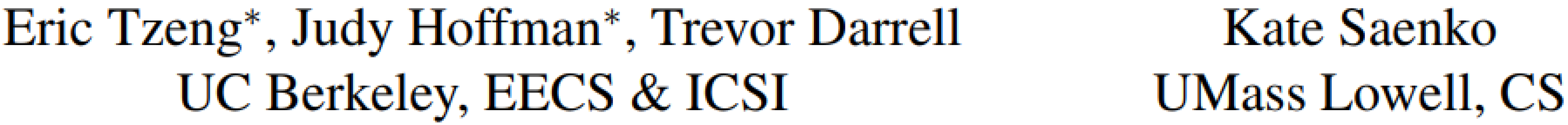
Simultaneous Deep Transfer Across Domains and Tasks



Eric Tzeng*, Judy Hoffman*, Trevor Darrell
UC Berkeley, EECS & ICSI

Kate Saenko
UMass Lowell, CS

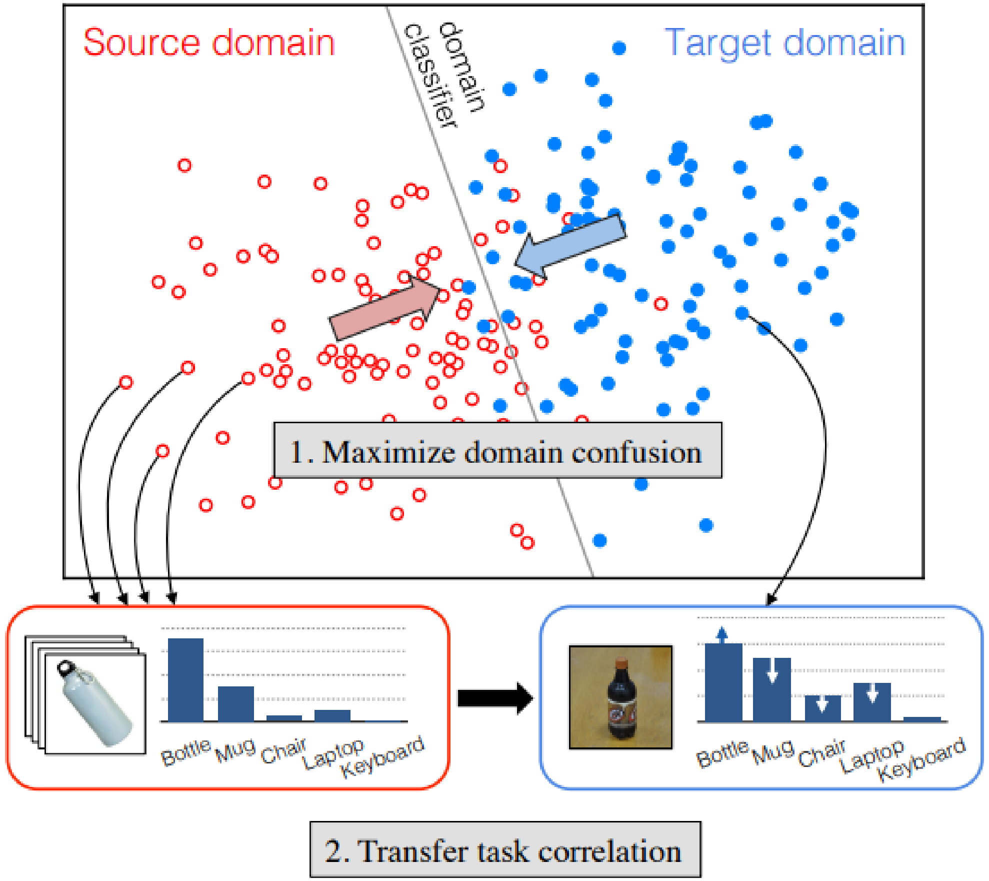
Presented by: Diego Machado

Motivation

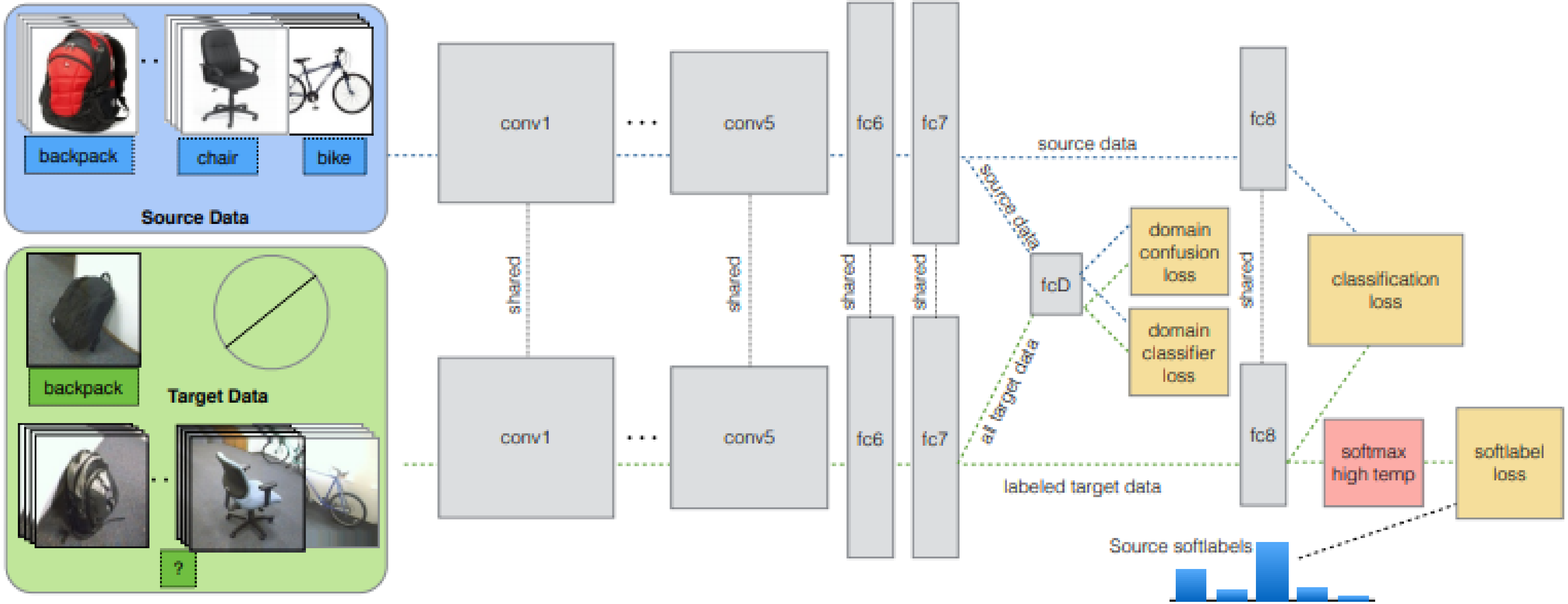
- Fine-tuning deep models in a new domain can require a significant amount of labeled data, which for many applications is simply not available.

Proposed

- Authors propose an algorithm that adapts between the source and target domains by utilizing generic statistics from target domain, and a few human labeled examples from a subset of the categories of interest.
- Transfer learning across domains and tasks.



Overview of CNN architecture



Joint loss function

$$\begin{aligned}\mathcal{L}(x_S, y_S, x_T, y_T, \theta_D; \theta_{\text{repr}}, \theta_C) = \\ &\mathcal{L}_C(x_S, y_S, x_T, y_T; \theta_{\text{repr}}, \theta_C) \\ &+ \lambda \mathcal{L}_{\text{conf}}(x_S, x_T, \theta_D; \theta_{\text{repr}}) \\ &+ \nu \mathcal{L}_{\text{soft}}(x_T, y_T; \theta_{\text{repr}}, \theta_C).\end{aligned} \quad (2)$$

$$\mathcal{L}_C(x, y; \theta_{\text{repr}}, \theta_C) = -\sum_k \mathbb{1}[y = k] \log p_k \quad (1)$$

$$\mathcal{L}_{\text{conf}}(x_S, x_T, \theta_D; \theta_{\text{repr}}) = -\sum_d \frac{1}{D} \log q_d. \quad (4)$$

$$\mathcal{L}_{\text{soft}}(x_T, y_T; \theta_{\text{repr}}, \theta_C) = -\sum_i l_i^{(y_T)} \log p_i \quad (7)$$

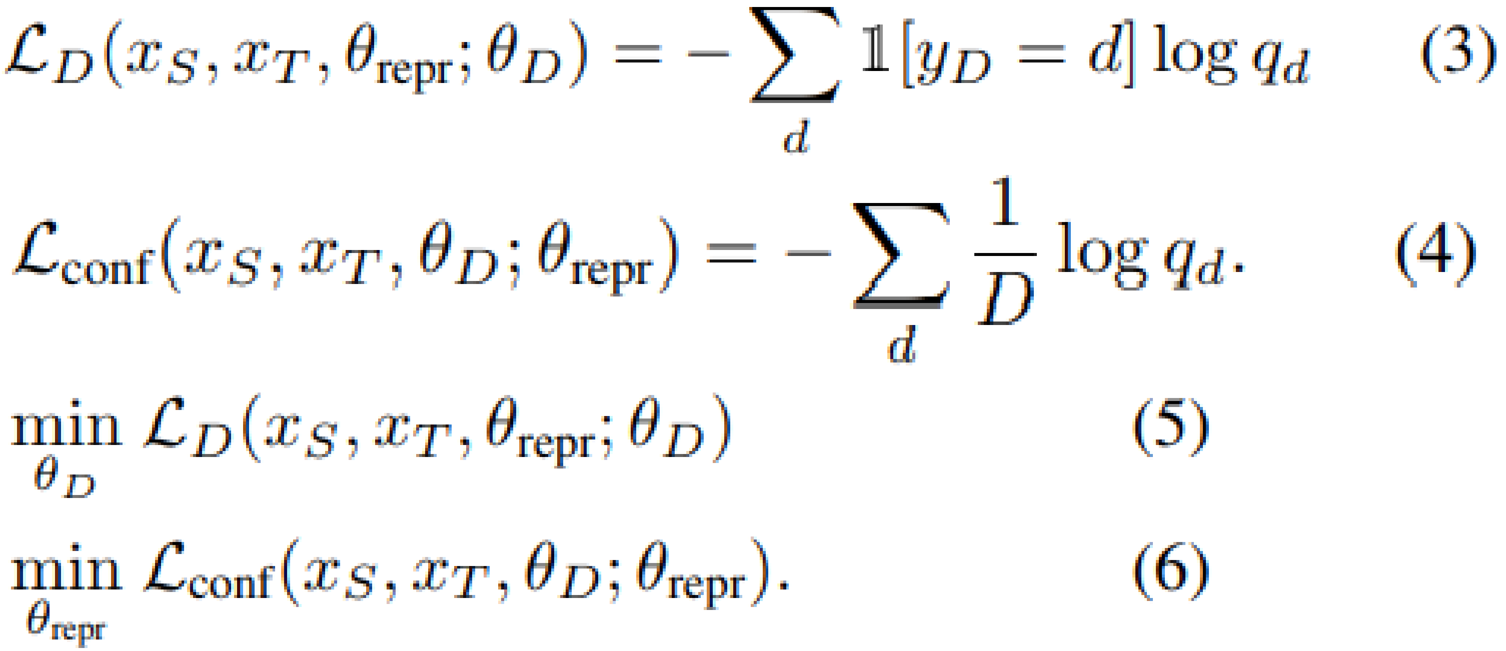
Classifier loss function

- category classifier θ_C
- image feature representation $f(x; \theta_{repr})$
- representation parameters θ_{repr}
- $p = \text{softmax}(\theta^T_C f(x; \theta_{repr}))$.

$$\mathcal{L}_C(x, y; \theta_{\text{repr}}, \theta_C) = -\sum_k \mathbb{1}[y = k] \log p_k \qquad (1)$$

Aligning domains

y_D denotes the domain that the example is drawn from.
q = softmax(θ^T_D f(x; θrepr))



$$\mathcal{L}_D(x_S, x_T, \theta_{\text{repr}}; \theta_D) = -\sum_d \mathbb{1}[y_D = d] \log q_d \qquad (3)$$

$$\mathcal{L}_{\text{conf}}(x_S, x_T, \theta_D; \theta_{\text{repr}}) = -\sum_d \frac{1}{D} \log q_d. \qquad (4)$$

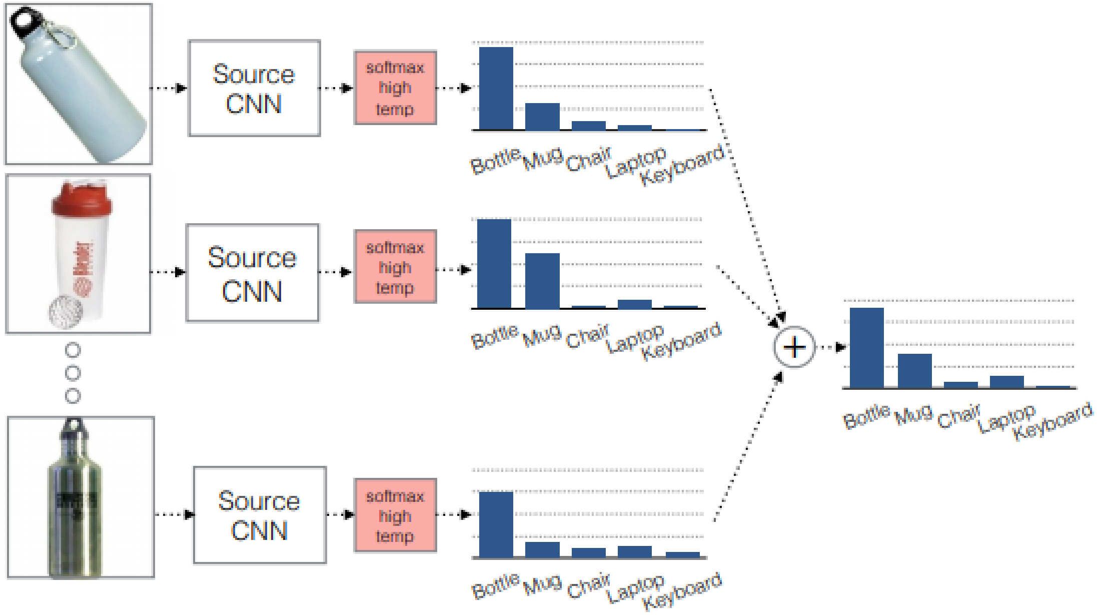
$$\min_{\theta_D} \mathcal{L}_D(x_S, x_T, \theta_{\text{repr}}; \theta_D) \qquad (5)$$

$$\min_{\theta_{\text{repr}}} \mathcal{L}_{\text{conf}}(x_S, x_T, \theta_D; \theta_{\text{repr}}). \qquad (6)$$

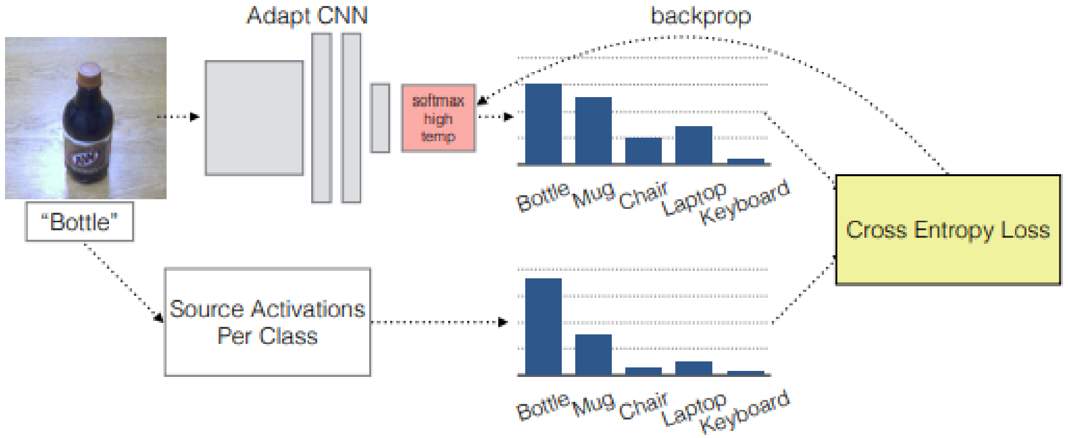
Aligning source and target classes

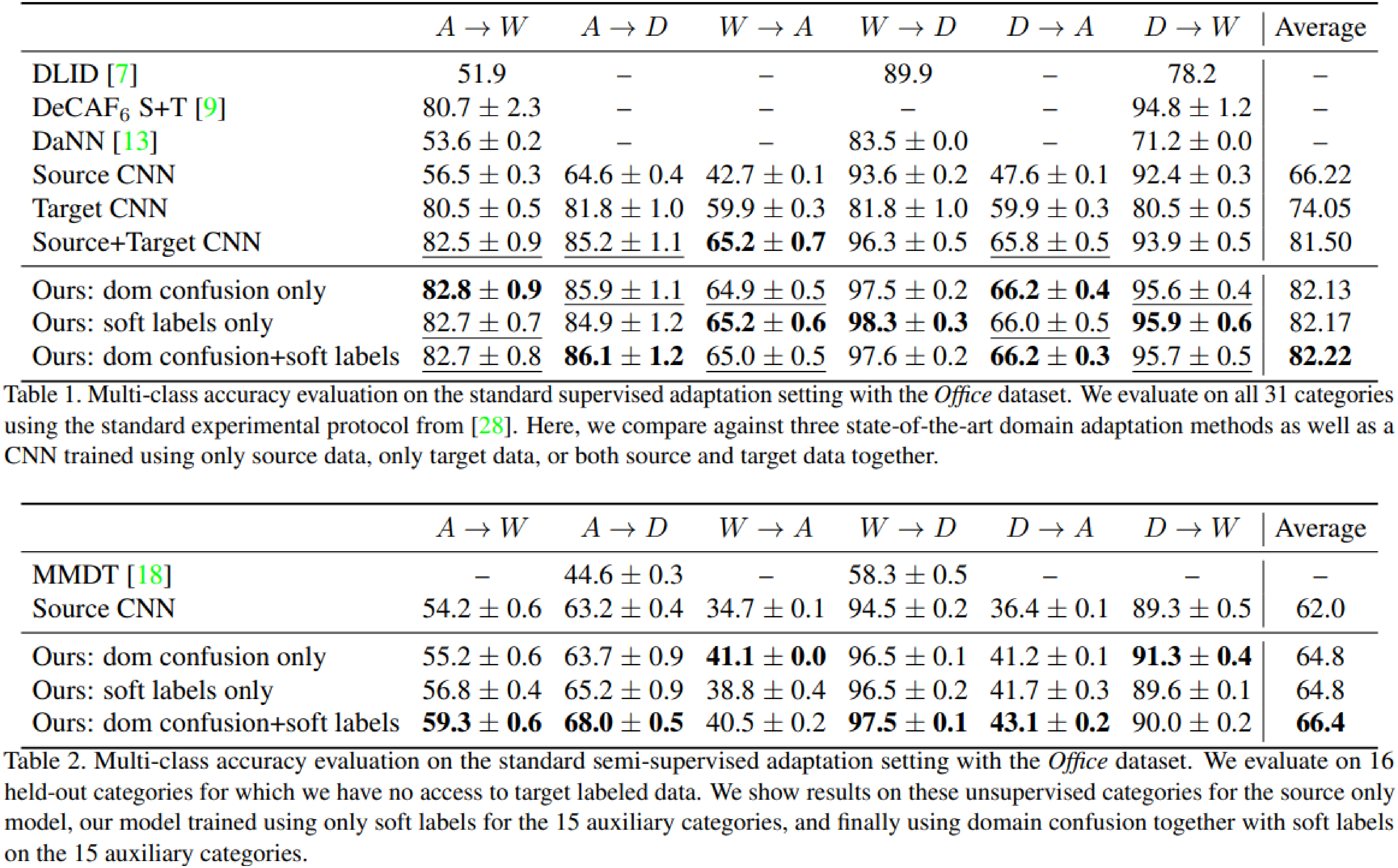
$l^{(k)}$: average over the softmax of all activations of source examples in category k.

p = softmax(θ^{T}_{C} f(xT ; θ_{repr})/τ).



$$\mathcal{L}_{\text{soft}}(x_T, y_T; \theta_{\text{repr}}, \theta_C) = -\sum_i l_i^{(y_T)} \log p_i \qquad (7)$$





	$A \rightarrow W$	$A \rightarrow D$	$W \rightarrow A$	$W \rightarrow D$	$D \rightarrow A$	$D \rightarrow W$	Average
DLID [7]	51.9	–	–	89.9	–	78.2	–
DeCAF$_6$ S+T [9]	80.7 ± 2.3	–	–	–	–	94.8 ± 1.2	–
DaNN [13]	53.6 ± 0.2	–	–	83.5 ± 0.0	–	71.2 ± 0.0	–
Source CNN	56.5 ± 0.3	64.6 ± 0.4	42.7 ± 0.1	93.6 ± 0.2	47.6 ± 0.1	92.4 ± 0.3	66.22
Target CNN	80.5 ± 0.5	81.8 ± 1.0	59.9 ± 0.3	81.8 ± 1.0	59.9 ± 0.3	80.5 ± 0.5	74.05
Source+Target CNN	82.5 ± 0.9	85.2 ± 1.1	**65.2 ± 0.7**	96.3 ± 0.5	65.8 ± 0.5	93.9 ± 0.5	81.50
Ours: dom confusion only	**82.8 ± 0.9**	85.9 ± 1.1	64.9 ± 0.5	97.5 ± 0.2	**66.2 ± 0.4**	95.6 ± 0.4	82.13
Ours: soft labels only	82.7 ± 0.7	84.9 ± 1.2	**65.2 ± 0.6**	**98.3 ± 0.3**	66.0 ± 0.5	**95.9 ± 0.6**	82.17
Ours: dom confusion+soft labels	82.7 ± 0.8	**86.1 ± 1.2**	65.0 ± 0.5	97.6 ± 0.2	**66.2 ± 0.3**	95.7 ± 0.5	**82.22**

Table 1. Multi-class accuracy evaluation on the standard supervised adaptation setting with the *Office* dataset. We evaluate on all 31 categories using the standard experimental protocol from [28]. Here, we compare against three state-of-the-art domain adaptation methods as well as a CNN trained using only source data, only target data, or both source and target data together.

	$A \rightarrow W$	$A \rightarrow D$	$W \rightarrow A$	$W \rightarrow D$	$D \rightarrow A$	$D \rightarrow W$	Average
MMDT [18]	–	44.6 ± 0.3	–	58.3 ± 0.5	–	–	–
Source CNN	54.2 ± 0.6	63.2 ± 0.4	34.7 ± 0.1	94.5 ± 0.2	36.4 ± 0.1	89.3 ± 0.5	62.0
Ours: dom confusion only	55.2 ± 0.6	63.7 ± 0.9	**41.1 ± 0.0**	96.5 ± 0.1	41.2 ± 0.1	**91.3 ± 0.4**	64.8
Ours: soft labels only	56.8 ± 0.4	65.2 ± 0.9	38.8 ± 0.4	96.5 ± 0.2	41.7 ± 0.3	89.6 ± 0.1	64.8
Ours: dom confusion+soft labels	**59.3 ± 0.6**	**68.0 ± 0.5**	40.5 ± 0.2	**97.5 ± 0.1**	**43.1 ± 0.2**	90.0 ± 0.2	**66.4**

Table 2. Multi-class accuracy evaluation on the standard semi-supervised adaptation setting with the *Office* dataset. We evaluate on 16 held-out categories for which we have no access to target labeled data. We show results on these unsupervised categories for the source only model, our model trained using only soft labels for the 15 auxiliary categories, and finally using domain confusion together with soft labels on the 15 auxiliary categories.

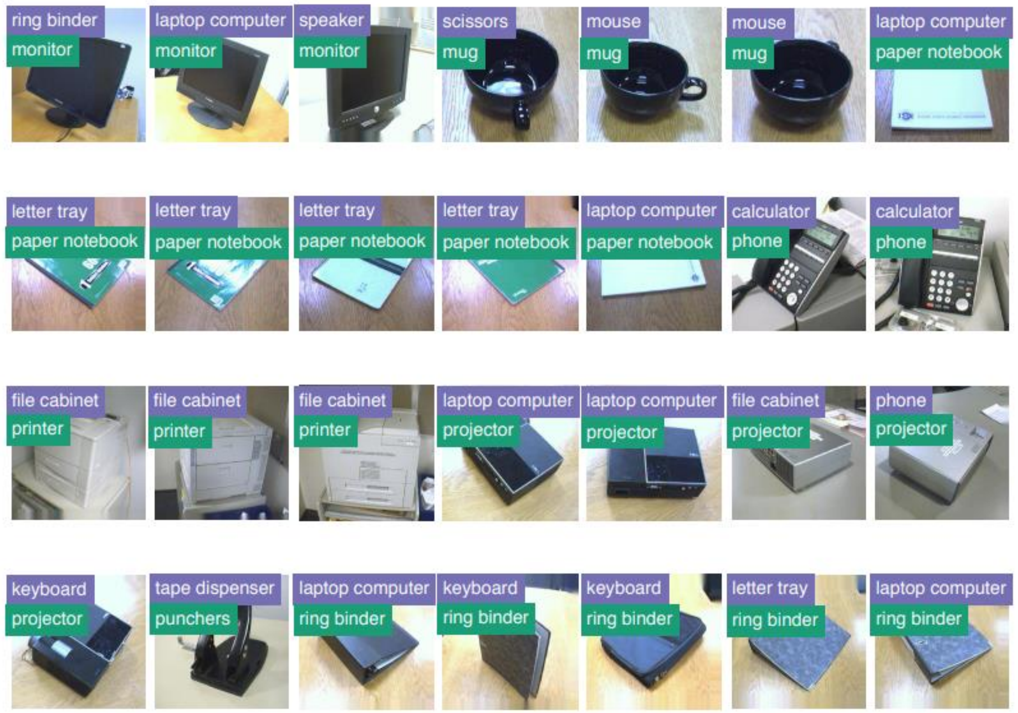
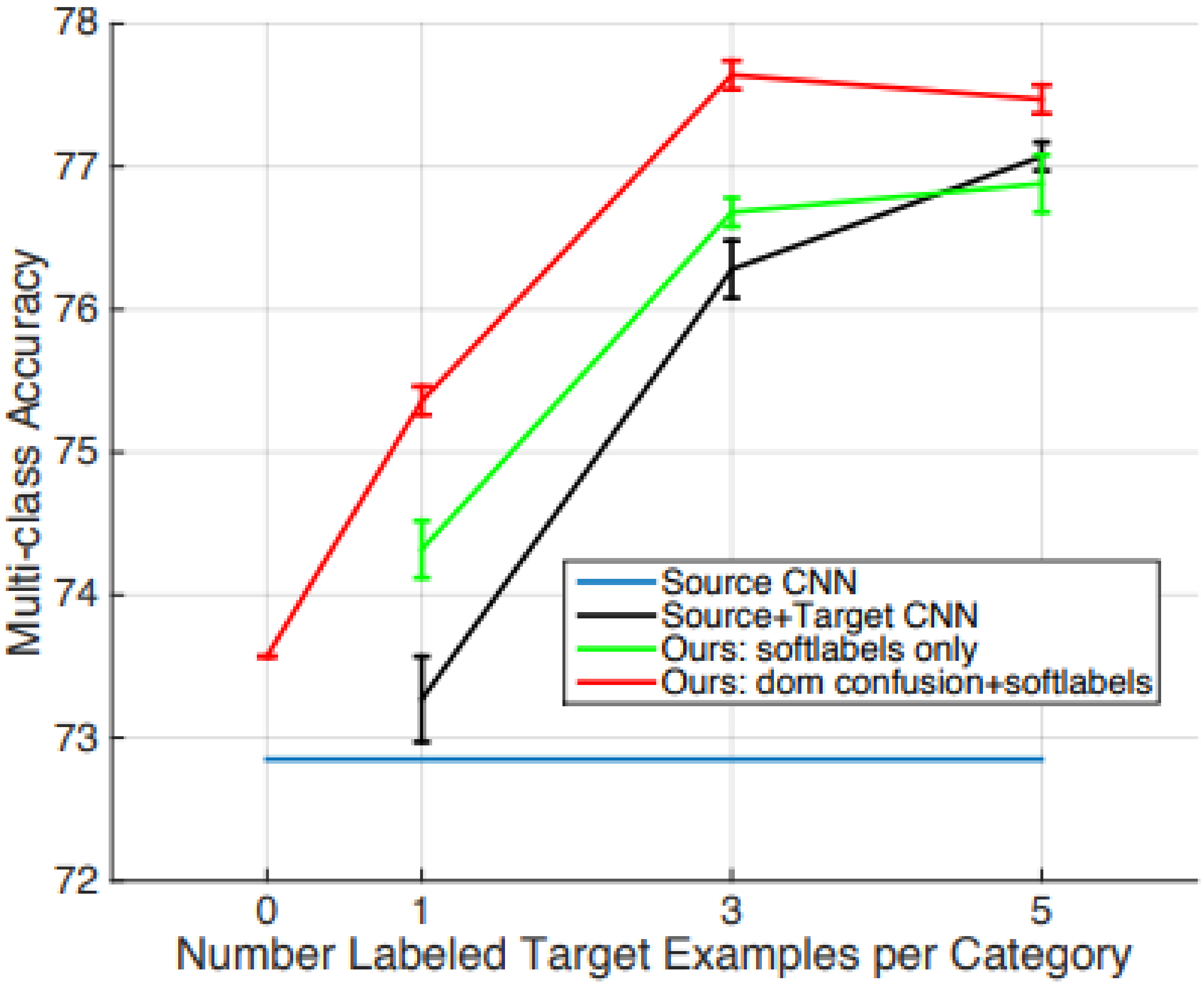
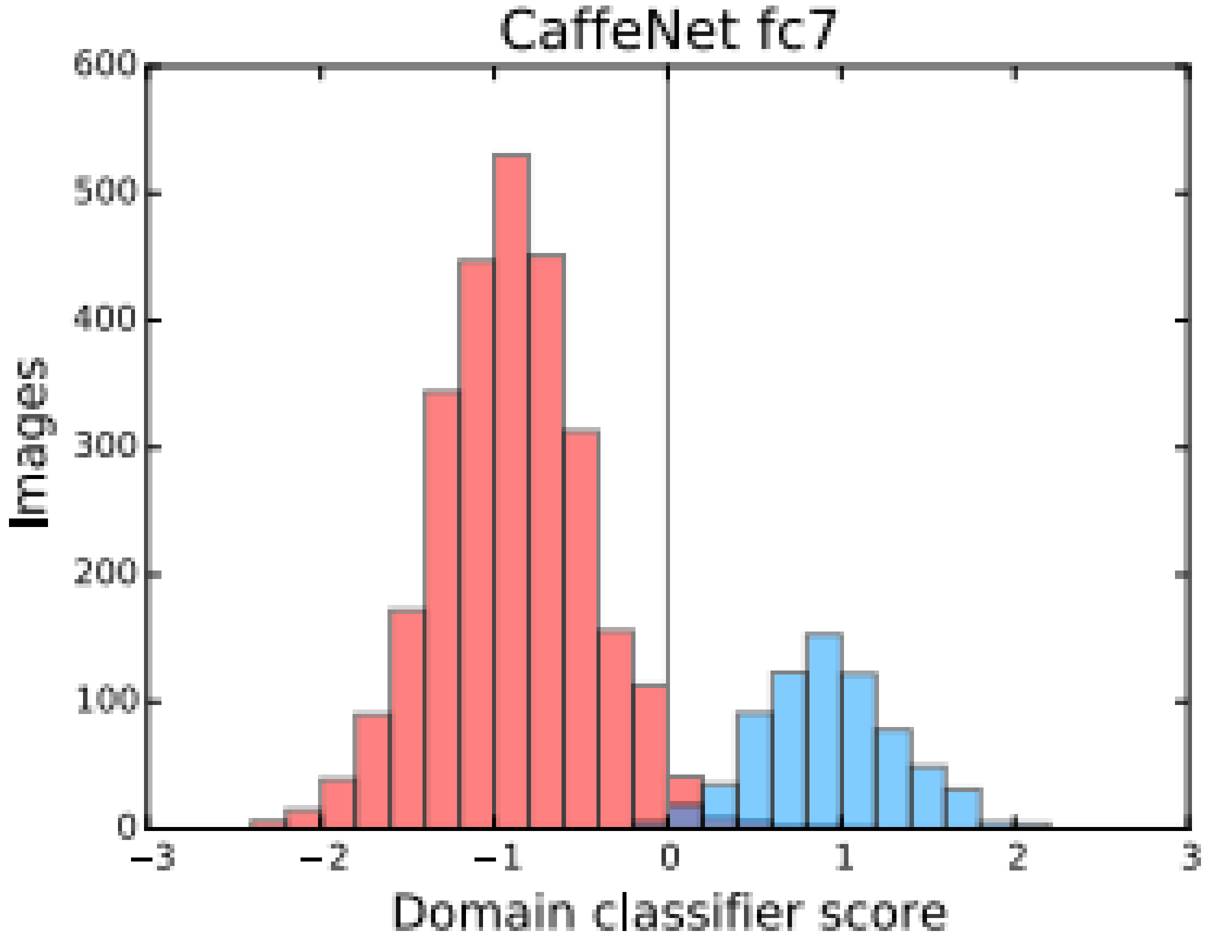
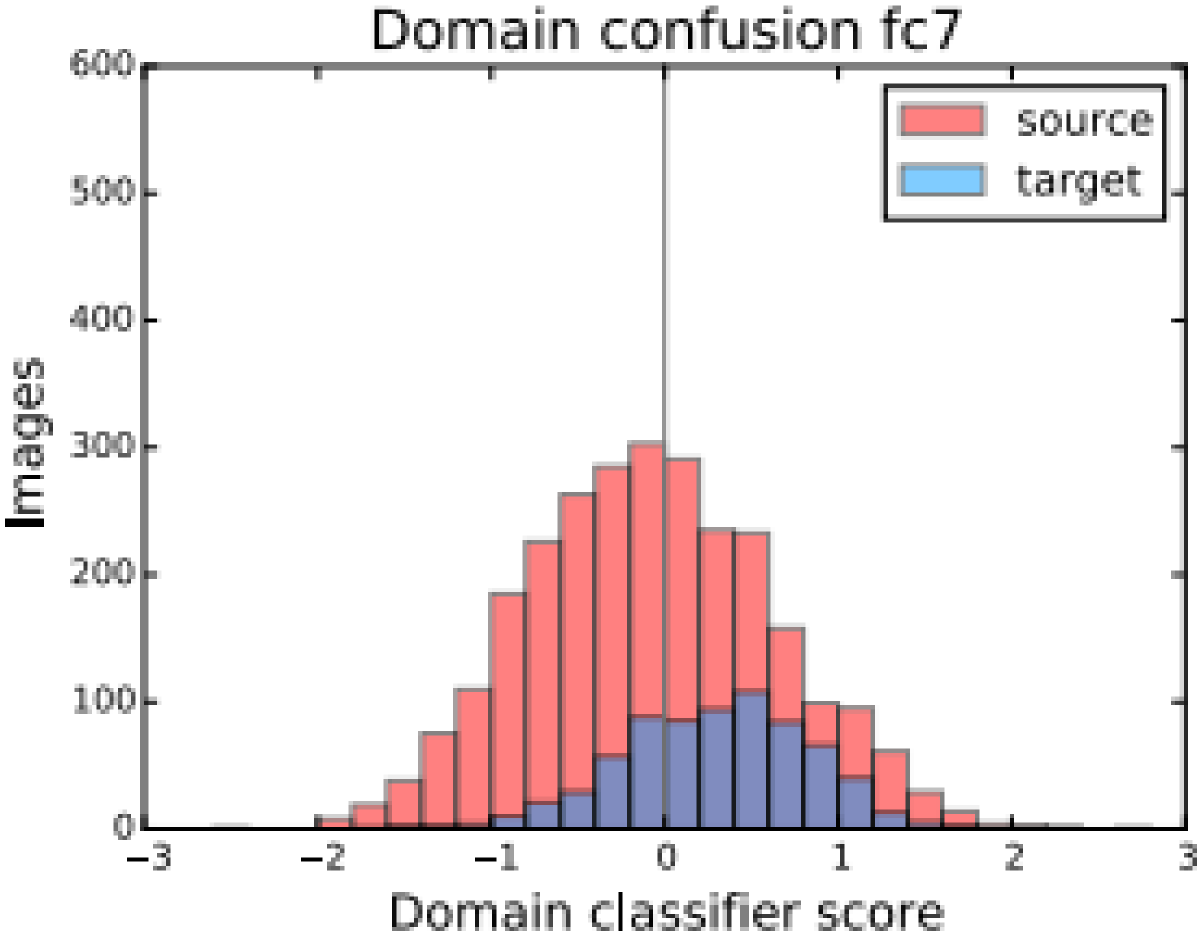
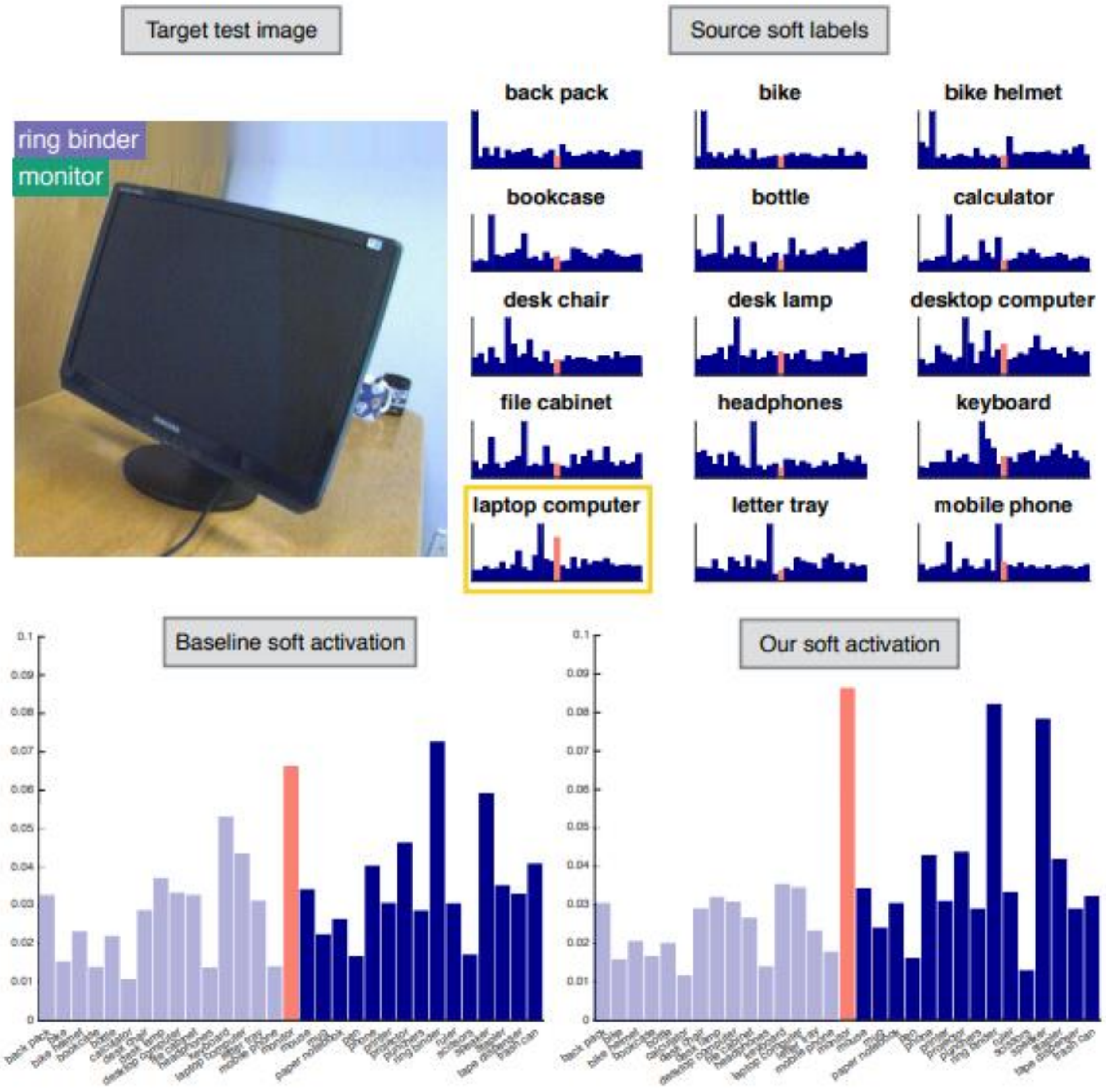


Figure 5. Examples from the Amazon→Webcam shift in the semi-supervised adaptation setting, where our method (the bottom turquoise label) correctly classifies images while the baseline (the top purple label) does not.


78
77
76
75
74
73
72
Multi-class Accuracy
0
1
3
5
Number Labeled Target Examples per Category
Source CNN
Source+Target CNN
Ours: softlabels only
Ours: dom confusion+softlabels


CaffeNet fc7
Images
600
500
400
300
200
100
0
−3
−2
−1
0
1
2
3
Domain classifier score


Domain confusion fc7
source
target
Images
600
500
400
300
200
100
0
−3
−2
−1
0
1
2
3
Domain classifier score


Target test image
ring binder
monitor
Source soft labels
back pack
bike
bike helmet
bookcase
bottle
calculator
desk chair
desk lamp
desktop computer
file cabinet
headphones
keyboard
laptop computer
letter tray
mobile phone
Baseline soft activation
Our soft activation
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
back pack
bike
bike helmet
bookcase
bottle
calculator
desk chair
desk lamp
desktop computer
file cabinet
headphones
keyboard
laptop computer
letter tray
mobile phone
monitor
mouse
mug
paper notebook
pen
phone
printer
projector
punchers
ring binder
ruler
scissors
speaker
stapler
tape dispenser
trash can