1. Proof of Theorem

We resort to the pseudo-labels to bound the combined error of the ideal hypothesis, $C$. Then, the following inequality holds:

**Theorem 1.** Let $f_T$ be the pseudo-labeling function. Given $R_{T'}(f_S, f_T)$ and $R_{T'}(f_T, f_T)$ as the minimum shared error and the degree to which the target samples are falsely labeled on $D_t$, respectively. We have

$$C \leq \min_{h \in H} R_S(h, f_S) + R_{T'}(h, f_T) + 2R_{T'}(f_S, f_T) + R_{T'}(f_T, f_T).$$

(1)

**Proof.** This proof relies on the triangle inequality for classification error [1, 2], which implies that for any labeling functions $f_1$, $f_2$, and $f_3$, we have $R(f_1, f_2) \leq R(f_1, f_3) + R(f_2, f_3)$. Then

$$C = \min_{h \in H} R_S(h, f_S) + R_{T'}(h, f_T)$$

$$\leq \min_{h \in H} R_S(h, f_S) + R_{T'}(h, f_S) + R_{T'}(f_S, f_T)$$

$$\leq \min_{h \in H} R_S(h, f_S) + R_{T'}(h, f_T) + R_{T'}(f_S, f_T)$$

$$+ R_{T'}(f_S, f_T) + R_{T'}(f_T, f_T)$$

$$= \min_{h \in H} R_S(h, f_S) + R_{T'}(h, f_T)$$

$$+ 2R_{T'}(f_S, f_T) + R_{T'}(f_T, f_T)$$

Therefore, the main inequality (1) holds.

2. Experiments

**More Training Details.** For Office-31 and ImageCLEF-DA datasets (AlexNet-based): (1) we augment the input images by scaling all images to $256 \times 256$, randomly cropping

Table 5: Comparing Different Temperature $T$ on A→W.

<table>
<thead>
<tr>
<th>$T$</th>
<th>1.5</th>
<th>1.6</th>
<th>1.7</th>
<th>1.8</th>
<th>1.9</th>
<th>2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy (%)</td>
<td>80.9</td>
<td>81.8</td>
<td>82.1</td>
<td><strong>83.0</strong></td>
<td>80.9</td>
<td>80.7</td>
</tr>
</tbody>
</table>

227×227 patches and executing random flips; (2) we set the batch size to 100 for each domain; (3) We add a bottleneck layer fc8 with 256 units after the fc7 layer for safer transfer representation learning; (4) the discriminator consists of 3 fully connected layers: 1024 → 1024 → 1. For digit recognition datasets: (1) all images are cast to $32 \times 32 \times 1$ in all experiments for fair comparison; (2) we set the batch size to 128 for each domain; (3) the discriminator consists of 3 fully connected layers: 500 → 500 → 1; (4) We use the Adam optimizer with learning rates set to 0.01 instead of the learning rate annealing strategy.

**Comparing Different Temperature $T$.** We perform sensitivity analysis of the temperature $T$ on transfer task A → W. We provide the classification accuracy as the changing of $T$ in [1.5, 0.1, 2.0] and the results are reported in Table 5. The accuracy first increases and then decreases as $T$ varies and shows a best result when $T = 1.8$. The results implicitly confirm our assumption that a good UDA model needs a non-saturated source classifier.

**References**
