Supplementary material for: A Novel Unsupervised Camera-aware Domain Adaptation Framework for Person Re-identification

Lei Qi[1], Lei Wang[2*], Jing Huo[1], Luping Zhou[3], Yinghuan Shi[1], and Yang Gao[1*]

[1]State Key Laboratory for Novel Software Technology, Nanjing University
[2]School of Computing and Information Technology, University of Wollongong
[3]School of Electrical and Information Engineering, The University of Sydney

A1. Theoretical analysis of the CCE-based scheme

The following conducts theoretical analysis for the proposed CCE- (i.e., cross-domain camera equiprobability) based scheme.

Proposition. *Let $\mathcal{S}$ and $\mathcal{T}$ denote source and target domains. x^s and x^t are the images from the two domains; $p_s(x)$ and $p_t(x)$ are their probability density functions; and C_s and C_t are the number of camera classes in these two domains. Let $p(x|\mathcal{C}_i^s)$ and $p(x|\mathcal{C}_i^t)$ be the class-conditional density functions of the ith camera class in source and target domains, respectively. It can be proved that ideally, minimizing the CCE loss will lead to*

$$\begin{aligned} p(x^s|\mathcal{C}_i^t) &= p_s(x^s), \quad \forall x^s \in \mathcal{S};\ i = 1, \cdots, C_t. \\ p(x^t|\mathcal{C}_i^s) &= p_t(x^t), \quad \forall x^t \in \mathcal{T};\ i = 1, \cdots, C_s. \\ p_s(x) &= p_t(x), \quad \forall x \in \mathcal{S} \cup \mathcal{T}. \end{aligned} \tag{8}$$

Proof. All the following analysis is conducted in the context of the learned feature representation (or equally, the learned shared subspace). Given an image x^s from source domain, its posteriori probability with respect to the ith camera class in target domain (denoted by $\mathcal{C}_i^t (i = 1, \cdots, C_t)$) can be expressed via the Bayes' rule as

$$P(\mathcal{C}_i^t|x^s) = \frac{p(x^s|\mathcal{C}_i^t)P(\mathcal{C}_i^t)}{p_s(x^s)}, \quad \forall x^s \in \mathcal{S};\ i = 1, \cdots, C_t, \tag{9}$$

where $p(x^s|\mathcal{C}_i^t)$ is the class-conditional probability density function of the ith camera class in target domain, $p_s(x^s)$ denotes the probability density function of the images in source domain, and $P(\mathcal{C}_i^t)$ is the priori probability of the ith camera class in target domain. Referring to Eq.(3) in our paper, $P(\mathcal{C}_i^t|x^s)$ is just $D(B(x), j)$ in the CCE loss.

Let us investigate the CCE loss in Eq.(3) of our paper to gain understanding on the optimal value of $D(B(x), j)$ when this loss is minimized. Since the CCE loss is defined for each individual image x independently, it will be sufficient to investigate the minimization of the loss for any given image x. Without loss of generality, it is assumed that x is from source domain. For clarity, $D(B(x), j)$ is compactly denoted by D_j. With respect to $D_j (j = 1, \cdots, C_t)$, the minimization of the CCE loss can be expressed as a constrained optimization

$$\min_{\{D_1, \cdots, D_{C_t}\}} \left(-\frac{1}{C_t} \sum_{j=1}^{C_t} \log(D_j) \right) \tag{10}$$

with the constraints of $D_j \geq 0$ and $\sum_{j=1}^{C_t} D_j = 1$, considering that D_j represents the posteriori probability. Due to the symmetry of the objective function with respect to the variables $D_1, \cdots, D_{C_t}$, it is not difficult to see that the optimal value of D_j is $1/C_t$ for $j = 1, \cdots, C_t$. A rigorous proof can be readily obtained by applying the Karush-Kuhn-Tucker conditions to this optimization, which is omitted here. This indicates that $P(\mathcal{C}_i^t|x^s)$ will equal $1/C_t$ when the CCE loss is minimized for this given image x. Now we assume the ideal case that this CCE loss is minimized for any given image x in source domain[1].

*Corresponding authors: Yang Gao; Lei Wang. The work of J. Huo was supported by NSFC (61806092) and Jiangsu Natural Science Foundation (BK20180326). The work of Y. Shi was supported by the Fundamental Research Funds for the Central Universities (020214380056), NSFC (61673203), CCF-Tencent Open Research Fund (RAGR20180114). The work of Y. Gao was supported by NSFC (61432008).

[1]Note that such an ideal case may not be really achieved in practice. Nevertheless, it helps to clearly reveal the effect of minimizing the CCE loss in the theoretical sense.

Let us turn to Eq.(9) and rearrange it as

$$p(x^s|\mathcal{C}_i^t) = \frac{P(\mathcal{C}_i^t|x^s)}{P(\mathcal{C}_i^t)} p_s(x^s), \quad \forall x^s \in \mathcal{S};\ i = 1, \cdots, C_t. \tag{11}$$

Without loss of generality, equal priori probability can be set for the C_t camera classes in target domain, that is, $P(\mathcal{C}_i^t)$ is constant $1/C_t$. Further, note that by optimizing D_j in Eq.(10) above, it can be known that

$$P(\mathcal{C}_i^t|x^s) = \frac{1}{C_t}, \quad \forall x^s \in \mathcal{S};\ i = 1, \cdots, C_t. \tag{12}$$

Combining the above results, Eq.(11) becomes

$$p(x^s|\mathcal{C}_i^t) = \frac{1/C_t}{1/C_t} p_s(x^s) = p_s(x^s), \quad \forall x^s \in \mathcal{S};\ i = 1, \cdots, C_t. \tag{13}$$

Note that the right hand side of this equation does not depend on the index i. This indicates that *in the learned shared subspace, the class-conditional density function for each camera class in target domain becomes same for any given* $x^s \in \mathcal{S}$. Applying the same argument to the images in target domain can similarly obtain

$$p(x^t|\mathcal{C}_i^s) = p_t(x^t), \quad \forall x^t \in \mathcal{T};\ i = 1, \cdots, C_s, \tag{14}$$

where $p(x^t|\mathcal{C}_i^s)$ and $p_t(x^t)$ are defined in the similar way as the above. This result indicates that *in the learned shared subspace, the class-conditional density function for each camera class in source domain becomes same for any given* $x^t \in \mathcal{T}$.

The above results indicate that for an image in source domain, it will not feel the distribution discrepancy among the camera classes in target domain. Furthermore, its class-conditional density function value (e.g., $p(x^s|\mathcal{C}_i^t)$) for those camera classes just equals its density function value in source domain (e.g., $p_s(x^s)$). The similar remark can be made for an image in target domain.

Upon the above results, the following further proves that in the learned shared subspace, the data distributions of source and target domains will become identical and this removes the domain-level distribution discrepancy.

For any image x^s from source domain, its value evaluated by the probability density function of target domain can be obtained as

$$p_t(x^s) = \sum_{i=1}^{C_t} p(x^s|C_i^t)P(C_i^t) = \sum_{i=1}^{C_t} p_s(x^s)P(C_i^t) = p_s(x^s), \tag{15}$$

where the first equality is due to Eq.(13) and the second one is because $\sum_{i=1}^{C_t} P(C_i^t) = 1$. Similarly, the result for any given image x^t from target domain can be obtained as

$$p_s(x^t) = \sum_{i=1}^{C_s} p(x^t|C_i^s)P(C_i^s) = \sum_{i=1}^{C_s} p_t(x^t)P(C_i^s) = p_t(x^t), \tag{16}$$

where the first equality is due to Eq.(14) and the second one is because $\sum_{i=1}^{C_s} P(C_i^s) = 1$.

Collectively, the above two results indicate that for any image x from either source or target domain, the following result can be obtained.

$$p_s(x) = p_t(x), \quad \forall x \in \mathcal{S} \cup \mathcal{T}. \tag{17}$$

This means that the two distributions, $p_s(x)$ and $p_t(x)$, are identical on the set $\mathcal{S} \cup \mathcal{T}$. With respect to the definitions of the two distributions, this indicates that *upon the learned feature representation, the data distributions of source and target domains become identical on the set* $\mathcal{S} \cup \mathcal{T}$ *and that the distribution discrepancy is therefore removed.* ■

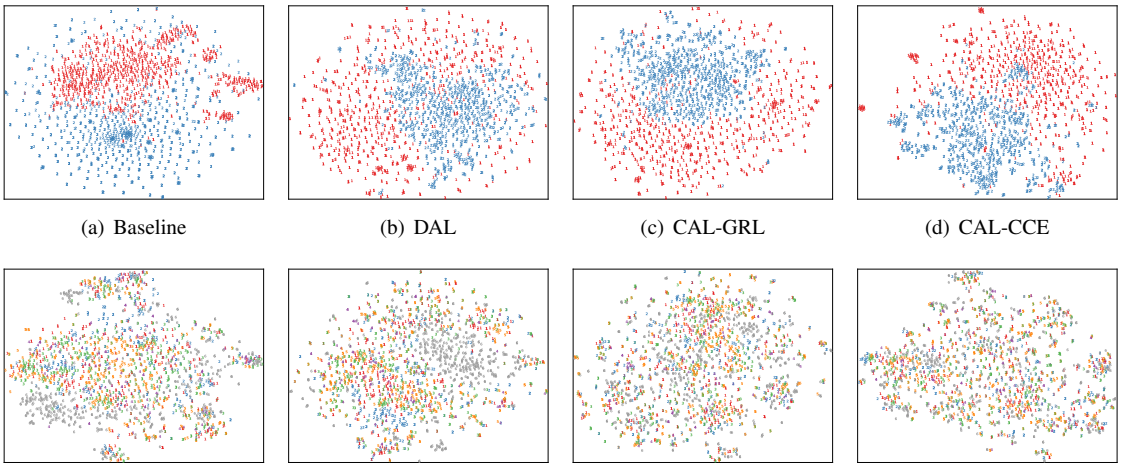
In addition, it is worth mentioning that the ideal minimization of the CCE loss does not theoretically guarantee that in the shared subspace, an image from either source or target domain will not feel the distribution discrepancy among the camera classes in *its own* domain. In other words, the results that $p(x^s|\mathcal{C}_1^s) = \cdots = p(x^s|\mathcal{C}_{C_s}^s)$ or $p(x^t|\mathcal{C}_1^t) = \cdots = p(x^t|\mathcal{C}_{C_t}^t)$ cannot directly be derived from the ideal minimization of the CCE loss.

Nevertheless, note that Eq.(17) implies that at any place x in $\mathcal{S} \cup \mathcal{T}$, the probability density of images from source domain is the same as the probability density of images form target domain. This means that the images from two domains have been adequately mixed up. In this case, considering that the result $p(x^s|\mathcal{C}_1^t) = \cdots = p(x^s|\mathcal{C}_{C_t}^t)$ is true (as proved in Eq.(13)), we can reasonably expect that this result shall be generalized from x^s to x^t, that is, $p(x^t|\mathcal{C}_1^t) = \cdots = p(x^t|\mathcal{C}_{C_t}^t)$ becomes true. Applying the same argument can obtain the result $p(x^s|\mathcal{C}_1^s) = \cdots = p(x^s|\mathcal{C}_{C_s}^s)$. Therefore, it can be reasonably expected that in practice in the shared subspace, an image from either source or target domain will not feel the distribution discrepancy among the camera classes in *its own* domain. Experimental study has been conducted to show that this tendency can indeed be observed in practice, as shown by Table 6 in our paper and the following Fig. A1.

A2. Visualization of data distributions

The data distributions are visualized at the domain-level and camera-level via t-SNE [1] in Fig. A1. We extract the features of each image by the baseline model (BL), DAL, CAL-GRL and CAL-CCE, respectively, in the task of "DukeMTMC-reID→Market1501". The top row shows the distributions of source and target domains (i.e., inter-domain), where blue and red colors indicate source and target domains, respectively. The bottom row illustrates the distribution of each camera class in target domain (i.e., inter-camera on Market1501), where different colors denote different camera classes.

First, from the inter-domain results shown in the top row of Fig. A1, it can be seen that DAL, CAL-GRL and CAL-



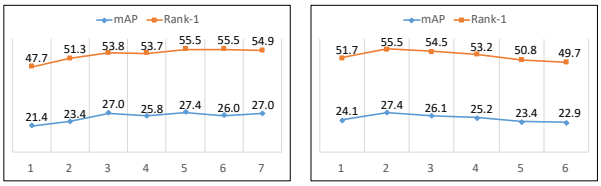
(a) Baseline (b) DAL (c) CAL-GRL (d) CAL-CCE

(e) Baseline (f) DAL (g) CAL-GRL (h) CAL-CCE

Figure A1. Visualization of data distributions at the domain-level and camera-level via t-SNE [1]. The features of each image are extracted by the baseline (BL), DAL, CAL-GRL and CAL-CCE in the task of "DukeMTMC-reID→Market1501", respectively. The top shows the distributions of source and target domains (i.e., inter-domain), where blue and red colors indicate source and target domains, respectively. The bottom illustrates the distribution of each camera class in target domain (i.e., inter-camera on Market1501), where different colors denote different camera classes. Note that all figures correspond to the experimental results in Table 6 of our paper.

CCE can effectively "mix" the two domains when compared with BL. This validates that they are all able to reduce the data distribution discrepancy between source and target domains.

Second, from the inter-camera result shown in the bottom row in Fig. A1, it can be seen that both CAL-GRL and CAL-CCE seem to better "mix" these camera classes than BL and DAL which do not consider any camera-level discrepancy. Furthermore, consistent with its lowest inter-camera distance reported in Table 6 of our paper, CAL-CCE displays an excellent "mixture" of different camera classes as expected, further illustrating its best capability in reducing the camera-level discrepancy in target domain.



(a) k (It decides k_p) (b) k_n

Figure A2. Parameter sensitivity (by BL+UOT in "Duke→Market1501").

A3. On the parameter sensitivity of UOT.

In this section, we conduct experiments to observe the parameter sensitivity of UOT. The results are reported in Fig. A2. Firstly, note that k_p is not directly preset but decided by k and the specific data (Recall that k_p is the number of positive samples of an anchor within the top-k positions and share its fragment). As in Fig. A2, when k is as small as 1, no positive samples (i.e., k_p is often zero) could be found and thus performance is poor. When k goes up to 5, the performance tends to plateau. Secondly, for k_n, if we only select one negative sample (i.e., k_n is set as 1), this sample often undesirably shares the same identity as the anchor. Meanwhile, setting k_n too large could include many easy negative samples, instead of hard negative ones preferred by UOT. In all experiments, we uniformly set k and k_n as 5 and 2.

References

[1] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research (JMLR)*, 2008. 2, 3