Unsupervised Mutual Mean Teaching with Heterogeneous Models for Person Re-Identification

Wenlong Wang*, Xiaofu Wu, Suofei Zhang, Yuxin Zhang and Jun Yan

Nanjing University of Posts and Telecommunications, No.66 XinMoFan Road, Gulou District, Nanjing City, Jiangsu Province, China

*Email: 1019010623@njupt.edu.cn

Abstract. Person re-identification (Re-ID) aims at matching images of the same pedestrian under different cameras. State-of-the-art unsupervised domain adaptation (UDA) methods for person Re-ID transferred the learned knowledge from the source domain and further finetuned with pseudo-label target domain via clustering. To reduce the influence of noisy pseudo labels, the Mutual Mean-Teaching (MMT) framework was proposed by conducting pseudo label refinery to better model inter-sample relations in the target domain. Currently, MMT requires the use of the two same deep models for cooperative correction of noisy labels. In this paper, we extend the MMT framework to use two heterogeneous models for cooperative error correction. Experiments show that the proposed heterogeneous MMT framework can work well with two different deep models and its performance, however, is largely limited by the weaker model.

1. Introduction
The main purpose of Person re-identification (Re-ID) is to match images of the same pedestrian under different cameras. Due to its important applications in security and surveillance, Person Re-ID has gained widespread attention from academia world and industry world. Although the adoption of CNN (convolutional neural network [1-3]) has greatly improved the system performance, the distribution bias between the source dataset and the target dataset may cause a significant performance drop when migrating to the target domain [4,5]. Because it is inefficient to annotate every image from target datasets, the employment of UDA (Unsupervised Domain Adaptation) methods has become the most popular solution to this problem.

Most of UDA methods utilize cluster algorithm to produce pseudo label so that we can train neural networks accordingly. Recently, mutual mean-teaching (MMT) framework has shown its power on unsupervised person Re-ID, where two deep models can teach each other for pseudo-label correction and achieves state-of-the-art performance [6]. Currently, the MMT framework requires the two models to be the same, which limits its potential usage. In this paper, we extend the MMT framework to support two heterogeneous models.

2. Related Works

2.1. Teacher-Student Models
One of the most studied methods in semi-supervised learning is the so-called Teacher-Student model. The main idea of the teacher-student model is to use different models' predictions to create consistent
training supervision for labelled or unlabeled data. Deep mutual learning [7] adopted a pool of student models by training them with supervision from each other. MMT using soft labels to let the two models teach each other [6]. However, existing methods with teacher-student mechanism are mostly designed for models with same network structure, thus cannot be directly used for heterogeneous models.

2.2. Unsupervised Domain Adaptation (UDA) for Person Re-Identification
UDA (Unsupervised Domain Adaptation) methods attract a lot of attention because no manual annotation is required. Fan [5] proposed to assign labels to unlabeled data alternately, and then optimize the network using these generated targets. Lin [8] proposed a bottom-up clustering framework with a repelled loss. Yang [9] assigned hard pseudo labels for both global and local features. Ge [6] introduced on-line refined soft pseudo labels by using Mutual-Mean Teaching framework.

3. Proposed Approach

3.1. Mutual Mean Teaching Revisit
The MMT framework relies on the clustering algorithm to generate pseudo labels. It uses two soft classification losses, namely, \( L_{\text{Id}}^i(\theta_1 | \theta_2) \) and \( L_{\text{Id}}^2(\theta_2 | \theta_1) \), where:

\[
L_{\text{Id}}^i(\theta_1 | \theta_2) = \frac{1}{N_t} \sum_{t=1}^{N_t} \left( C_2^i \left( F(x_t^i | [E(T)](\theta_2)) \right) - \log C_1^i \left( F(x_t^i | \theta_1) \right) \right).
\]

(1)

We denote the feature-embedding functions for the two networks as \( F(\cdot | \theta_1) \) and \( F(\cdot | \theta_2) \). The corresponding pseudo label classifiers can be denoted as \( C_1^i \) and \( C_2^i \), respectively. The images are denoted as \( x_t^i \) and \( x_t^j \) for the two networks, and the confidences of their pseudo label can be predicted as \( C_1^i(F(x_t^i | \theta)) \) and \( C_2^i(F(x_t^j | \theta)) \).

And the soft triplet loss can be calculated as:

\[
L_{\text{tripl}}^i(\theta_1 | \theta_2) = \frac{1}{N_t} \sum_{t=1}^{N_t} \text{bce} \left( T_t(\theta_1), T_t(E(T) | [\theta_2]) \right).
\]

(2)

where

\[
T_t(\theta) = \frac{\exp(\| F(x_t^i | \theta) - F(x_t^i_{\text{n}} | \theta) \|)}{\exp(\| F(x_t^i | \theta) - F(x_t^o | \theta) \|) + \exp(\| F(x_t^i | \theta) - F(x_t^i_{\text{n}} | \theta) \|)}.
\]

(3)

MMT also requires to use pseudo-labels for “supervised” training of two deep models, where both the hard ID loss \( L_{\text{Id}}^i(\theta_m), m = 1, 2 \) and the hard triplet loss \( L_{\text{tripl}}^i(\theta_m) \) are both used.

For any input image \( x \), MMT generates two embedding feature vectors with the same model, namely, \( f_1 = F(x | \theta_1) \) and \( f_2 = F(x | \theta_2) \). Due to the use of the two same models, \( f_1 \) and \( f_2 \) are homogeneous with the same dimension. Then, MMT proposed to employ its mean form, namely,

\[
f(x) = 0.5(f_1 + f_2) = 0.5[F(x | \theta_1) + F(x | \theta_2)].
\]

(4)

For producing pseudo-labels via a clustering algorithm \( C \) over the target domain \( D_t = \{x_t^i, i = 1, \cdots, n_t\} \). Mathematically, this can be concisely expressed as:

\[
Y_t = C \left( f(D_\tau) \right),
\]

(5)

where \( Y_t = \{y_{ti}, i = 1, \cdots, n_t\} \) denotes the generated pseudo labels for \( D_t \).

The overall loss is as follows:

\[
\mathcal{L}(\theta_1, \theta_2) = \left( 1 - \lambda_{\text{Id}}^i \right) \left( L_{\text{Id}}^i(\theta_1) + L_{\text{Id}}^i(\theta_2) \right) + \lambda_{\text{Id}}^i \left( L_{\text{Id}}^i(\theta_1 | \theta_2) + L_{\text{Id}}^i(\theta_2 | \theta_1) \right) + \left( 1 - \lambda_{\text{tripl}}^i \right) \left( L_{\text{tripl}}^i(\theta_1) + L_{\text{tripl}}^i(\theta_2) \right) + \lambda_{\text{tripl}}^i \left( L_{\text{tripl}}^i(\theta_1 | \theta_2) + L_{\text{tripl}}^i(\theta_2 | \theta_1) \right).
\]

(6)

3.2. Heterogeneous Mutual Mean Teaching
We propose a Heterogeneous Mutual Mean Teaching (HMMT) framework with two different deep models, namely, \( F_1(\cdot | \theta_1) \) and \( F_2(\cdot | \theta_2) \) with \( F_1 \neq F_2 \). The feature vector extracted from model-1 \( f_1 = \begin{align*}
L_{\text{Id}}(\theta_1, \theta_2) &= \left( 1 - \lambda_{\text{Id}}^i \right) \left( L_{\text{Id}}^i(\theta_1) + L_{\text{Id}}^i(\theta_2) \right) + \lambda_{\text{Id}}^i \left( L_{\text{Id}}^i(\theta_1 | \theta_2) + L_{\text{Id}}^i(\theta_2 | \theta_1) \right) + \left( 1 - \lambda_{\text{tripl}}^i \right) \left( L_{\text{tripl}}^i(\theta_1) + L_{\text{tripl}}^i(\theta_2) \right) + \lambda_{\text{tripl}}^i \left( L_{\text{tripl}}^i(\theta_1 | \theta_2) + L_{\text{tripl}}^i(\theta_2 | \theta_1) \right).
\end{align*}
(6)
$F_1(x|\theta_1)$ is essentially heterogeneous with the feature vector $f_2 = F_2(x|\theta_2)$ from model-2, therefore we cannot directly use its mean form as the MMT framework does. Instead of using the simple mean form, we propose to employ its concatenated form as:

$$\tilde{f}(x) = (f_1, f_2) = (F_1(x|\theta_1), F_2(x|\theta_2)).$$

Hence, the pseudo-label generation over $\mathcal{D}_t$ can be expressed as

$$\mathcal{U}_t = C(\tilde{f}(\mathcal{D}_t)).$$

which can be directly implemented for two cooperative but heterogeneous models.

![Figure 1](image.png)

**Figure 1.** The overall framework of HMMT. Net1 and Net2 use different model structure. The dimension of $c_1, c_2$ is the same as the output dimension of model.

### 3.3. Training Procedure

#### 3.3.1. Supervised Pre-Training for Source Domain

The idea of UDA task in Person Re-ID is to transfer learned knowledge from model on the source domain to target domain. We need to pre-train our network on the source domain first. We denote our training datasets as $\mathcal{D}_s$, and the transformation function as $F(\cdot|\theta)$, it aims at transforming the input sample $x^s_i$ into a feature representation $F(x^s_i|\theta)$. After the model obtained the encoded features, classifier $c^s$ will produce an $M_s$-dimensional probability vector to predict the identities in the source datasets. Classification loss $L^{\text{id}}(\theta)$ and triplet loss $L^{\text{tri}}(\theta)$ are adopted to optimize the network. Note that the same data augmentation method is applied for pre-training the models, but when fine-tuning two models, we use different data augmentation methods.

$$L^{\text{id}}_{t_{id}}(\theta) = \frac{1}{N_t} \sum_{i=1}^{N_t} L^{\text{ce}}(y_{tp}, \hat{y}_{ti}^\tau).$$

$$L^{\text{tri}}_{t_{id}}(\theta) = \frac{1}{N_t} \sum_{i=1}^{N_t} \max(0, |F(x^s_i|\theta) - F(x^s_{tp}|\theta)| + m - |F(x^s_i|\theta) - F(x^s_{tn}|\theta)|).$$

The overall loss is therefore can be calculated as:

$$L^s(\theta) = L^{\text{id}}_{t_{id}}(\theta) + \lambda^{\text{tri}} L^{\text{tri}}_{t_{id}}(\theta).$$
3.3.2. Training with Heterogeneous Models
Our HMMT framework generates soft pseudo labels by jointly training two different networks with different augmentation methods. The overall framework is illustrated in figure 1. We first extract two feature vectors with different models, and then concatenate them using equation [7]. Then, we employ DBSCAN algorithm to obtain pseudo label and cluster center. Finally, we split cluster center and replace model’s classifier’s weight data with corresponding cluster center. The detailed optimization procedures are summarized in figure 2.

**Algorithm 1 HMMT Training Procedure**

for $n \in [1, epochs]$ do
  1. Extracting features $f_1, f_2$.
  2. Concatenate $f_1, f_2$ into $f$ with Eq 7.
  3. Use $f$ to generate hard pseudo labels $\gamma_i$ for each sample $x_i$ in $D_t$ by clustering algorithms DBSCAN. And obtain predicted number of class and cluster center $c$.
  4. Split $c$ into $c_1, c_2$ according to corresponding model’s classifier dimension.
  5. Replace model’s classifier weight data, $c_1$ for model-1, and $c_2$ for model-2.
for each mini-batch $B \subset D_t$, iteration $T$ do
  1. Optimizing models using hard pseudo label with Eq 9, Eq 10.
  2. Generate soft pseudo labels from the collaborative networks and optimizing models with Eq 6.
end for
end for

**Figure 2. HMMT Training Procedure.**

4. Experiment

4.1. Datasets
We evaluate our proposed HMMT on two most popular datasets, Market1501, DukeMTMC [10]. Market-1501 [11] dataset contains 32,668 images with 1,501 identities. All images are shot from 6 cameras. And DukeMTMC-reID consists of 16,522 annotated images that are shot from 8 cameras. Both training set and testing set have 702 identities.

**Table 1. Advantage of HMMT framework for the cooperation of the two same models.**

| Method            | Performance |
|-------------------|-------------|
|                  | mAP         | Top-1      |
| K-means + Mean Form | 71.2%       | 87.7%      |
| DBSCAN + Mean Form    | 72.5%       | 88.0%      |
| DBSCAN + Concatenate   | 73.0%       | 88.2%      |

4.2. Implementation Details
We employ ResNet-50, IBN-ResNet-50, OS-Net as the heterogenous backbone networks. Note that OS-Net employed is of the network scale of 1.0 and 0.75.

For pre-training on the source-domain and fine-tuning on the target-domain, we set batchsize at 64, with 16 identities (4 for each identity). The images are all resized to 256 x 128. Note that each epoch will re-generate hard pseudo label, so it is necessary to re-organize the batch with updated hard pseudo label.

In experiments, all hyper-parameters are the same for two datasets, where the weight decay of Adam optimizer is set at 0.0005. The employed data augmentation methods include Random Erasing [12], and it will only be adopted when fine-tuning the model in target-domain. When pre-training in source-domain, the batch of images, network parameters $\theta_1, \theta_2$ are updated independently by
optimizing Equation 11 with $\lambda^s = 1$. The total iteration is 80 epochs, the initial learning rate is 0.00035 and is multiplied by 0.1 when in 40th and 70th epoch. In the training with HMMT, the two models are updated by optimizing equation 6 with $\lambda^i_{id} = 0.5$ and $\lambda^t_{tr} = 0.8$. The learning rate is set at 0.00035 for the entire training process.

### 4.3. Results

MMT framework uses the mean form of the features from both model-1 and model-2, which, however, fail to work for two heterogeneous models. Therefore, we resort to the concatenation of features and feed it to the clustering algorithm. Note that we employ DBSCAN for clustering, since it does not require to specify the number of classes and it also outperforms K-means. Eventually, this method achieves the performance improvement of 0.5%, 1.8% on mAP compared to the original MMT using K-means and DBSCAN, respectively (Please refer to table 1).

The performance of HMMT with two different models are shown in table 2. Clearly, HMMT works well and its performance, however, is largely limited by the weaker model.

| Models                  | Market1501-DukeMTMC | DukeMTMC-Market1501 |
|-------------------------|---------------------|---------------------|
|                         | Model-1 Performance | Model-2 Performance |
|                         | mAP  | top-1 | top-5 | top-10 | mAP  | top-1 | top-5 | top-10 |
| ResNet50-ResNet50-IBN-a | 58.7% | 73.7% | 84.6% | 88.5% | 57.6% | 73.8% | 84.7% | 88.4% |
| ResNet50-OSNet1.0       | 62.0% | 75.4% | 86.4% | 91.0% | 63.6% | 79.8% | 88.5% | 91.0% |
| ResNet50-IBN-a-OSNet1.0 | 61.3% | 75.9% | 86.9% | 89.9% | 62.2% | 78.0% | 87.1% | 90.5% |
| OSNet1.0-OSNet0.75      | 62.6% | 76.8% | 86.0% | 88.8% | 62.0% | 76.4% | 85.3% | 88.7% |

| Models                  | DukeMTMC-Market1501 |
|-------------------------|---------------------|
|                         | Model-1 Performance | Model-2 Performance |
|                         | mAP  | top-1 | top-5 | top-10 | mAP  | top-1 | top-5 | top-10 |
| ResNet50-ResNet50-IBN-a | 72.4% | 87.5% | 95.0% | 96.9% | 75.5% | 89.9% | 95.4% | 97.0% |
| ResNet50-OSNet1.0       | 67.4% | 83.7% | 82.6% | 94.9% | 66.7% | 80.4% | 90.9% | 93.3% |
| ResNet50-IBN-a-OSNet1.0 | 71.1% | 86.7% | 93.4% | 95.4% | 67.8% | 85.4% | 91.9% | 93.9% |
| OSNet1.0-OSNet0.75      | 65.6% | 82.6% | 90.0% | 92.7% | 65.3% | 83.1% | 90.4% | 92.9% |

### 5. Conclusion

In this paper, we successfully extend the MMT framework to use two heterogeneous models for cooperative error correction. With the proposed framework, we show that the concatenation of features may be helpful for improving the performance of MMT with the two same models. Experiments show that the proposed heterogeneous MMT framework can work well with two heterogeneous models and its performance, however, is largely limited by the weaker model.

### References

[1] Sun Y, Zheng L, Yang Y, Tian Q, and Wang S 2016 Beyond part models. Person etrieval with refined part pooling. arXiv preprint arXiv:1711.09349 (2016)

[2] Hermans A, Beyer L, and Leibe B 2017 In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737

[3] Wang G, Yuan Y, Chen X, Li J, and Zhou X 2019 Learning discriminative features with multiple granularities for person re-identification. In ACM Multimedia
[4] Deng W, Zheng L, Kang G, Yang Y, Ye Q, and Jiao J 2018 Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person reidentification. In IEEE CVPR

[5] Fan H, Zheng L, Yan C, and Yang Y 2018 Unsupervised person re-identification: Clustering and fine-tuning. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)

[6] Ge Y, Chen D, and Li H 2020 Mutual Mean-Teaching: Pseudo Label Refinery for Unsupervised Domain Adaptation on Person Re-identification. ICLR

[7] Zhang Y, Xiang T, Hospedales T M, and Lu H 2018B Deep mutual learning. In CVPR

[8] Lin Y, Dong X, Zheng L, Yan Y and Yang Y 2019 A bottom-up clustering approach to unsupervised person re-identification. In AAAI

[9] Yang F, Wei Y, Wang G, Zhou Y, Shi H, and Huang T 2019 Selfsimilarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. ICCV

[10] Ristani E, Solera F, Cucchiara R Z, and Thmasi C 2016 Performance measures and a data set for mutil-target, multi-camera tracking. In ECCVW

[11] Zheng L, Shen L, Tian L, Wang S, Wang J, and Tian Q 2015 Scalable person re-identification: A benchmark. In ICCV

[12] Zhong Z, Zheng L, Kang G, Li S, and Yang Y 2017b Random erasing data augmentation. arXiv preprint arXiv:1708.04896