Improved Mutual Mean-Teaching for Unsupervised Domain Adaptive Re-ID

Yixiao Ge, Shijie Yu, and Dapeng Chen

The Chinese University of Hong Kong
yxge@link.cuhk.edu.hk

Abstract. In this technical report, we present our submission to the VisDA Challenge in ECCV 2020 and we achieved one of the top-performing results on the leaderboard. Our solution is based on Structured Domain Adaptation (SDA) [9] and Mutual Mean-Teaching (MMT) [7] frameworks. SDA, a domain-translation-based framework, focuses on carefully translating the source-domain images to the target domain. MMT, a pseudo-label-based framework, focuses on conducting pseudo label refinery with robust soft labels. Specifically, there are three main steps in our training pipeline. (i) We adopt SDA to generate source-to-target translated images, and (ii) such images serve as informative training samples to pre-train the network. (iii) The pre-trained network is further fine-tuned by MMT on the target domain. Note that we design an improved MMT (dubbed MMT+) to further mitigate the label noise by modeling inter-sample relations across two domains and maintaining the instance discrimination. Our proposed method achieved 74.78% accuracies in terms of mAP, ranked the 3rd place out of 54 teams. [†]

1 Introduction

Accurate person re-identification (re-ID) is at the core of smart city systems, which aims at retrieving the same person's images across multiple cameras. Although deep learning-based methods have achieved satisfying re-ID performances by training with large-scale datasets, inevitable domain gaps between different camera systems prevent the trained networks from being directly deployed on a new scene. Unsupervised domain adaptation (UDA) is therefore proposed to transfer the learned knowledge from the labeled source domain (dataset) to the unlabeled target domain (dataset).

In ECCV 2020, the VisDA Challenge introduced a synthetic→real UDA task on person re-ID. The synthetic source-domain data generated by the Unity engine [15] have the same style as PersonX [18] and their labels can be naturally provided. The unlabeled target-domain images are collected from real-world scenarios. It is a much challenging problem due to the fact that 1) there are larger domain gaps between synthetic and real scenarios than those between real and real scenarios and 2) the distribution of unlabeled target-domain data is much

---

[†]Code of this work is available at https://github.com/yxgeee/VisDA-ECCV20.
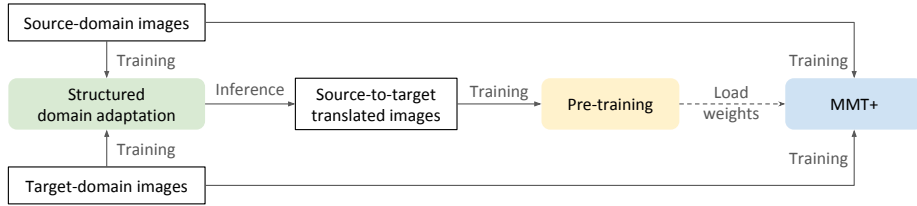
Fig. 1. The training pipeline of our proposed method, which consists of three steps: structured domain adaptation (SDA), pre-training with source-to-target translated images and fine-tuning on the target domain with the improved MMT framework.

more realistic than previous benchmarks. Specifically, existing benchmarks for UDA re-ID (*e.g.* Market-1501 [27]→DukeMTMC-reID [16]) generally construct the target-domain dataset by simply removing the labels from public datasets, however, such public datasets have already been manually organized, which could not well match the distribution of unlabeled datasets in practical use. In contrast, VisDA provides a noisy target-domain dataset with more practical settings.

Existing methods tackling the problem of UDA re-ID can be divided into two main categories, domain-translation-based methods [5, 20, 2, 9] and pseudo-label-based methods [17, 22, 26, 7, 24, 30, 23, 19]. Domain-translation-based methods target at translating the source-domain images to have the target-domain style while well preserving their original IDs [5, 20] or inter-sample relations [9]. Such kind of methods provides a plausible way to make use of source-domain images and their valuable ground-truth identities. Pseudo-label-based methods aim at learning the distribution of unlabeled target-domain data with pseudo labels, where the pseudo labels are generated by either clustering instance features [17, 22, 26, 7, 24] or measuring similarities with exemplar features [30, 23, 19]. Although pseudo-label-based methods could achieve superior performance than domain-translation-based methods, we argue that they are complementary to each other and can work together to achieve optimal performance.

To effectively take advantage of both domain-translation-based and pseudo-label-based methods, we introduce a training pipeline with Structured Domain Adaptation (SDA) [9] and Mutual Mean-Teaching (MMT) [7] frameworks. SDA, one of the state-of-the-art domain-translation-based methods [5, 20, 2, 9], adopts CycleGAN [31] architecture to perform image-to-image translation. An online relation-consistency regularization is introduced to maintain inter-sample relations during the training of SDA. Domain translation is crucial in this task due to the evident domain gaps between synthetic data and real-world data.

MMT, one of the state-of-the-art pseudo-label-based methods [17, 22, 26, 7, 24, 30, 23, 19], proposes to conduct pseudo label refinery with reliable soft labels, which are generated online in a mutual teaching pipeline. We further improve the MMT framework, namely MMT+, by jointly training with both source-domain images and target-domain images to model complex inter-sample relations across two domains. To further mitigate the effects caused by noisy pseudo labels, we add the MoCo [10] loss to maintain the instance discrimination. Since the mean

networks in the MMT framework are similar to the momentum encoders in MoCo [10], the MoCo loss can be easily utilized without extra costs.

As shown in Fig. 1, there are three main training stages in our introduced pipeline. (i) An SDA framework is trained to translate source-domain images to the target domain. (ii) Source-to-target translated images serve as training samples to pre-train the network with ground-truth identities. The network can then be roughly adapted to the target domain. (iii) The pre-trained network is further fine-tuned on the target domain with the proposed MMT+ framework. Both labeled source-domain raw images and unlabeled target-domain images are used for training.

The contributions of this work could be summarized as three-fold. (1) We introduce a joint pipeline to properly make use of both the domain-translation-based and pseudo-label-based frameworks, which are complementary to each other. (2) We propose to improve the state-of-the-art MMT [7] framework by modeling the inter-/intra-identity relations across two domains and preserving the instance discrimination to mitigate the effects caused by noisy pseudo labels. (3) Our proposed method achieved one of the top-performing results on the leaderboard, yielding 74.78% accuracies in terms of mAP, which ranked the 3rd place in the VisDA Challenge.

2 Preliminary

2.1 Source-domain Pre-training

Given source-domain images $\mathbb{X}^s$, the neural network $\mathcal{F}^s$ is trained to transform each sample $x^s \in \mathbb{X}^s$ into a feature vector $\boldsymbol{f}^s = \mathcal{F}^s(x^s)$, which could be used to predict its ground-truth identity y^s with a learnable classifier $\mathcal{C}^s : \boldsymbol{f}^s \to \{1, \cdots, p^s\}$, where p^s is the number of identities in the source domain. A classification loss in the form of cross-entropy loss ℓ_{ce} and a softmax-triplet loss in the form of binary cross-entropy loss ℓ_{bce} are adopted jointly for training,

$$\mathcal{L}^s_{\mathrm{cls}}(\mathcal{F}^s, \mathcal{C}^s) = \mathbb{E}_{x^s \sim \mathbb{X}^s} \left[\ell_{\mathrm{ce}}(\mathcal{C}^s(\boldsymbol{f}^s), y^s) \right], \tag{1}$$

$$\mathcal{L}^s_{\mathrm{tri}}(\mathcal{F}^s) = \mathbb{E}_{x^s \sim \mathbb{X}^s} \left[\ell_{\mathrm{bce}}(\mathcal{T}(\boldsymbol{f}^s), \mathbf{1}) \right], \tag{2}$$

where

$$\mathcal{T}(\boldsymbol{f}^s) = \frac{\exp(\|\boldsymbol{f}^s - \boldsymbol{f}^s_n\|)}{\exp(\|\boldsymbol{f}^s - \boldsymbol{f}^s_p\|) + \exp(\|\boldsymbol{f}^s - \boldsymbol{f}^s_n\|)}, \tag{3}$$

and the subscripts $_{p,n}$ denote the mini-batch's hardest positive and negative feature indexes for the anchor $\boldsymbol{f}^s$. The overall loss function for source-domain pre-training is $\mathcal{L}^s = \mathcal{L}^s_{\mathrm{cls}} + \mathcal{L}^s_{\mathrm{tri}}$.

2.2 Clustering-based Baseline Training

Clustering-based pipeline serves as a strong baseline model for pseudo-label-based methods, which alternates between (i) generating pseudo classes by clustering target-domain instances $\mathbb{X}^t$ and (ii) training the network $\mathcal{F}^t$ with generated pseudo classes. Specifically, after loading the pre-trained weights of $\mathcal{F}^s$ to

Fig. 2. Source-to-target translated images. Compared to SPGAN [5], the SDA [9] could better preserve inter-sample relations, *e.g.* maintaining consistent appearance for person images of the same identity.

$\mathcal{F}^t$, $\mathbb{X}^t$'s encoded features $\{\boldsymbol{f}^t\}$ are clustered into $\hat{p}^t$ classes and images within the same cluster are assigned the same label. Generally, the density-based clustering algorithm (*e.g.* DBSCAN) is adopted as we do not know the number of the target-domain identities. Note that $\hat{p}^t$ is automatically measured by density-based clustering. Similar to source-domain pre-training, a classification loss and a softmax-triplet loss are adopted,

$$\mathcal{L}_{\text{cls}}^t(\mathcal{F}^t, \mathcal{C}^t) = \mathbb{E}_{x^t \sim \mathbb{X}^t} \left[\ell_{\text{ce}}(\mathcal{C}^t(\boldsymbol{f}^t), \hat{y}^t) \right], \qquad (4)$$

$$\mathcal{L}_{\text{tri}}^t(\mathcal{F}^t) = \mathbb{E}_{x^t \sim \mathbb{X}^t} \left[\ell_{\text{bce}}(\mathcal{T}(\boldsymbol{f}^t), \mathbf{1}) \right], \qquad (5)$$

where $\hat{y}^t$ denotes the pseudo label for unlabeled data x^t. The overall loss function for clustering-based baseline training is $\mathcal{L}^t = \mathcal{L}_{\text{cls}}^t + \mathcal{L}_{\text{tri}}^t$. The generated pseudo labels are updated before each epoch.

3   Proposed Approach

We propose to properly make use of both the domain-translation-based method (*i.e.* SDA) and the pseudo-label-based method (*i.e.* MMT) with a three-stage training pipeline: (i) training a structured domain adaptation framework to carefully translate source-domain images to the target domain; (ii) the translated images serve as informative training samples to pre-train the network and roughly adapt the model onto the target domain; (iii) the pre-trained network is further fine-tuned on the target domain with an improved mutual mean-teaching framework.

3.1   Structured Domain Adaptation

Domain translation is crucial in the synthetic→real adaptation task since the gaps between source and target domains are significant. Directly pre-training

the network with raw source-domain data as described in Sec. 2.1 could only achieve limited performance on the target domain, which will result in inaccurate pseudo labels. We adopt the state-of-the-art domain-translation-based framework, Structured Domain Adaptation (SDA) [9], to translate source-domain images to have the target-domain style. SDA adopts CycleGAN [31] architecture as the backbone, and introduces an online relation-consistency regularization to maintain the inter-sample relations instead of simple IDs. Specifically, the inter-sample relations in SDA are measured online by source-domain and target-domain encoders. We use the pre-trained $\mathcal{F}^s$ (Sec. 2.1) as the source-domain encoder and $\mathcal{F}^t$ trained by clustering-based baseline (Sec. 2.2) as the target-domain encoder. The relation-consistency regularization can therefore be formulated as a soft binary cross-entropy loss

$$\mathcal{L}_{\mathrm{rc}}(\mathcal{G}^{s \to t}) = \mathbb{E}_{x^s \sim \mathbb{X}^s} \left[\ell_{\mathrm{bce}}(\mathcal{T}(\boldsymbol{f}^{s \to t}), \mathcal{T}(\boldsymbol{f}^s)) \right], \tag{6}$$

where $\mathcal{G}^{s \to t}$ is the source-to-target generator, $\boldsymbol{f}^{s \to t} = \mathcal{F}^t(x^{s \to t})$ and $\boldsymbol{f}^s = \mathcal{F}^s(x^s)$. Besides $\mathcal{L}_{\mathrm{rc}}$, conventional cycle generation losses used by CycleGAN are needed. As illustrated in Fig. 2, SDA can generate informative training samples, and the generated samples could be trained to achieve better pre-training performance.

3.2 Pre-training with Source-to-target Translated Images

Given the trained $\mathcal{G}^{s \to t}$ in SDA, we could translate the source-domain images $\mathbb{X}^s$ to the target domain, denoted as $\mathbb{X}^{s \to t}$. By adopting the similar classification loss and softmax-triplet loss in Sec. 2.1, we re-train the network $\mathcal{F}^s$ with $\mathbb{X}^{s \to t}$ and ground-truth labels $\mathbb{Y}^s$

$$\mathcal{L}_{\mathrm{cls}}^s(\mathcal{F}^s, \mathcal{C}^s) = \mathbb{E}_{x^{s \to t} \sim \mathbb{X}^{s \to t}} \left[\ell_{\mathrm{ce}}(\mathcal{C}^s(\hat{\boldsymbol{f}}^{s \to t}), y^s) \right], \tag{7}$$

$$\mathcal{L}_{\mathrm{tri}}^s(\mathcal{F}^s) = \mathbb{E}_{x^{s \to t} \sim \mathbb{X}^{s \to t}} \left[\ell_{\mathrm{bce}}(\mathcal{T}(\hat{\boldsymbol{f}}^{s \to t}), \mathbf{1}) \right], \tag{8}$$

where $\hat{\boldsymbol{f}}^{s \to t} = \mathcal{F}^s(x^{s \to t})$. By pre-training with translated images $\mathbb{X}^{s \to t}$, the network $\mathcal{F}^s$ can be roughly adapted to the target domain with much better performance than $\mathcal{F}^s$ pre-trained with $\mathbb{X}^s$ in Sec. 2.1.

3.3 Improved Mutual Mean-Teaching

The Mutual Mean-Teaching (MMT) [7] framework adopts a couple of networks, denoted as $\mathcal{F}_1^t$ and $\mathcal{F}_2^t$, and each network's mean-teacher network supervises the training of the other network by predicting robust soft labels. The original MMT framework is only trained with target-domain data and generated pseudo labels, however, we argue that properly modeling the inter-samples relations across two domains is critical to the final performance. We introduce an improved mutual mean-teaching framework (MMT+) by jointly training with two domains' raw images and minimize the domain gaps in mini-batches with domain-specific
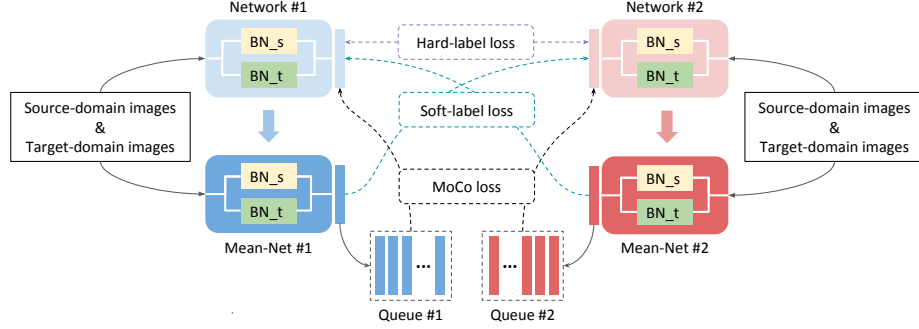
Fig. 3. We improve the Mutual Mean-Teaching (MMT) [7] framework by jointly train-ing with both source-domain and target-domain images to model complex inter-sample relations, and adding an instance discriminative MoCo [10] loss to further mitigate the effects caused by noisy pseudo labels.

BatchNorms [1] (Fig. 3). Specifically, we denote the joint dataset as $\mathbb{X} = \mathbb{X}^s \cup \mathbb{X}^t$. The cross-domain inter-sample relations are modeled by constructing a joint la-bel system with both source-domain ground-truth IDs and target-domain pseudo IDs. Given an encoded feature $\boldsymbol{f} = \mathcal{F}^t(x)$, it is trained to predict its own label among all the identities across two domains.

Let's denote the mean networks as $\mathbb{E}[\mathcal{F}_1^t]$ and $\mathbb{E}[\mathcal{F}_2^t]$, logits predicted by one mean network serve as soft targets to train the other network by a soft cross-entropy loss

$$\mathcal{L}_{\text{soft}}^t(\mathcal{F}_1^t, \mathcal{F}_2^t, \mathcal{C}_1^t, \mathcal{C}_2^t) = - \mathbb{E}_{x \sim \mathbb{X}}\Big[\mathbb{E}[\mathcal{C}_2^t](\mathbb{E}[\mathcal{F}_2^t](x)) \cdot \log \mathcal{C}_1^t(\mathcal{F}_1^t(x))$$
$$+ \mathbb{E}[\mathcal{C}_1^t](\mathbb{E}[\mathcal{F}_1^t](x)) \cdot \log \mathcal{C}_2^t(\mathcal{F}_2^t(x))\Big], \tag{9}$$

where the logits need to be normalized by a softmax operation. Instead of con-ventional back-propagation, the weights of mean networks are updated with a moving average formulation $\mathbb{E}[\theta] = \alpha\mathbb{E}[\theta] + (1-\alpha)\theta$, where $\alpha = 0.999$ is the momentum hyper-parameter. Besides the soft-label loss in Eq. (9), hard-label loss similar to Eq. (4) is adopted

$$\mathcal{L}_{\text{hard}}^t(\mathcal{F}_1^t, \mathcal{F}_2^t, \mathcal{C}_1^t, \mathcal{C}_2^t) = \mathbb{E}_{x \sim \mathbb{X}}\left[\ell_{\text{ce}}(\mathcal{C}_1^t(\mathcal{F}_1^t(x)), y) + \ell_{\text{ce}}(\mathcal{C}_2^t(\mathcal{F}_2^t(x)), y)\right]. \tag{10}$$

In the experiments, we find that the pseudo labels are much noisy due to the noisy distribution of identities in the target domain. In order to further mitigate the effects caused by pseudo label noise, we propose to adopt a MoCo [10] loss to maintain the instance discrimination. Specifically, the MoCo loss needs a queue with a fixed length to cache the features encoded by the momentum encoder, which is almost the same as our mean network. So we save the features encoded by our mean networks in the queue, and such features act as negative samples in the MoCo loss. Due to the coupled networks in MMT, we need two queues to serve individually for each network. The overall MoCo loss is formulated as a

contrastive loss

$$\mathcal{L}_{\mathrm{moco}}^t(\mathcal{F}_1^t,\mathcal{F}_2^t) = -\mathbb{E}_{x\sim\mathbb{X}}\Bigg[\log\frac{\exp(\langle\mathcal{F}_1^t(x),\mathbb{E}[\mathcal{F}_1^t](x)\rangle/\tau)}{\exp(\langle\mathcal{F}_1^t(x),\mathbb{E}[\mathcal{F}_1^t](x)\rangle/\tau)+\sum_{\boldsymbol{k}_1^-}\exp(\langle\mathcal{F}_1^t(x),\boldsymbol{k}_1^-\rangle/\tau)}$$

$$+\log\frac{\exp(\langle\mathcal{F}_2^t(x),\mathbb{E}[\mathcal{F}_2^t](x)\rangle/\tau)}{\exp(\langle\mathcal{F}_2^t(x),\mathbb{E}[\mathcal{F}_2^t](x)\rangle/\tau)+\sum_{\boldsymbol{k}_2^-}\exp(\langle\mathcal{F}_2^t(x),\boldsymbol{k}_2^-\rangle/\tau)}\Bigg], \qquad (11)$$

where k^- denotes the features from queue and $\tau = 0.7$ is the temperature hyper-parameter. The overall loss for the proposed MMT+ is

$$\mathcal{L}_{\mathrm{mmt+}}^t = \lambda_{\mathrm{soft}}\mathcal{L}_{\mathrm{soft}}^t + (1-\lambda_{\mathrm{soft}})\mathcal{L}_{\mathrm{hard}}^t + \lambda_{\mathrm{moco}}\mathcal{L}_{\mathrm{moco}}^t, \qquad (12)$$

where $\lambda_{\mathrm{soft}} = 0.5$, $\lambda_{\mathrm{moco}} = 0.1$ are the weighting parameters.

4   Experiments

4.1   Dataset and Evaluation Metric

There are four subsets provided by the VisDA Challenge[‡]: source_train, target_train, target_val and target_test, where only source_train and target_train can be used for training. The synthetic source-domain dataset is generated based on Unity [15]. Specifically, the source_train set consists of 20,280 images out of 700 identities shot from 6 cameras in total. The target-domain dataset consists of real-world images captured from 5 cameras, *i.e.* 13,198 images for training, 377 images for the query of target_val, 3,600 images for the gallery of target_val, 1,578 images for the query of target_test and 24,006 images for the gallery of target_test. Mean Average Precision (mAP) and Cumulated Matching Characteristics (CMC) accuracies are adopted to test the methods' performances, where only the top-100 matches are considered for evaluation.

4.2   Implementation Details

We implement our framework in PyTorch [13] and all the person images are resized to 384×128. We adopt ImageNet [4]-pretrained networks up to the global average pooling layer as the backbone, where GeM pooling [14] is adopted to replace the global average pooling layer for optimal performance. Adam optimizer is adopted with a weight decay of 0.0005.

Stage I: Structured domain adaptation. We adopt 8 GPUs for training the SDA, where each mini-batch contains 32 source-domain images of 8 ground-truth classes and 32 randomly sampled target-domain images. The training scheme iterates for 100 epochs and there are 200 iterations in each epoch, where the learning rate (lr = 0.0002) is constant for the first 50 epochs and then gradually decreases to 0 for another 50 epochs following the formula lr = lr × (1.0 − max(0, epoch − 50)/50).

[‡]https://github.com/Simon4Yan/VisDA2020

Stage II: Pre-training with source-to-target translated images. We adopt 1 GPU for pre-training, where each mini-batch contains 64 source-to-target images of 16 ground-truth classes. Auto augmentation [3] is adopted for pre-training. The initial learning rate is set to 0.00035 and is decreased to 1/10 of its previous value on the 40th and 70th epoch in the total 120 epochs. Each epoch has 200 iterations.

Stage III: Improved mutual mean-teaching. We adopt 4 GPUs for MMT+ training, where each mini-batch contains 64 source-domain images of 16 ground-truth classes and 64 target-domain images of 16 pseudo classes. Random erasing [29] is adopted. Domain-specific BNs [1] are adopted in this stage to minimize the domain gaps in each mini-batch. The learning rate is fixed to 0.00035 for overall 50 training epochs, where each epoch has 200 iterations. We use DBSCAN [6] and Jaccard distance [28] with k-reciprocal nearest neighbors for clustering before each epoch, where $k = 20$. As for DBSCAN, the maximum distance between neighbors is set as 0.6 and the minimal number of neighbors for a dense point is set as 4. The length of queue in MoCo loss is set as 12,800. Arcface loss or Cosface loss is adopted in this step to replace the simple classification loss.

Post-processing. (1) We adopt ResNet50-IBN [12] as the backbone for the training stage I to generate source-to-target training samples. ResNeSt50 [25], ResNeSt101 [25], DenseNet169-IBN [11, 12] and ResNeXt101-IBN [21, 12] are adopted for ensembling after training in stage II and III. Specifically, features encoded by the above backbones are concatenated and L_2-normalized. (2) Following [32], we train a camera classification network to predict the camera similarities between testing images. Let's denote the camera network as $\mathcal{F}_c$, then the image similarity between a query image q and a gallery image k is $s(q,k) = \|\boldsymbol{f}_q - \boldsymbol{f}_k\| - 0.1\|\mathcal{F}_c(q) - \mathcal{F}_c(k)\|$. $\mathcal{F}_c$ adopts ResNeSt50 [25] as the backbone. (3) We adopt the re-ranking technique [28] with $k_1 = 30, k_2 = 6, \lambda = 0.3$.

4.3   Quantitative Results

Table 1. Competition results of the VisDA Challenge in ECCV 2020. The results are evaluated on the target_test set. Our result is in **bold**.

Team Name	mAP(%)	top-1(%)
Logwhen	79.05	83.26
Vimar Team	76.56	84.25
Parrots	**74.78**	**82.86**
Xiangyu	72.39	83.85
Archer2	70.39	79.70

Comparison with other teams. As shown in Tab. 1, the introduced method achieved the 3rd place in terms of mAP accuracy, *i.e.* 74.78%.

Table 2. Ablation study on the effectiveness of source-to-target images translated by SDA. The results are evaluated on the target_val set. The pre-training for this ablation study adopts the backbone of ResNet50-IBN [12].

Images for pre-training	mAP(%)	top-1(%)
Raw source-domain images	61.0	71.6
Source-to-target images translated by SPGAN [5][a]	68.2	75.1
Source-to-target images translated by SDA [9]	**71.2**	**79.3**

[a]Downloaded from https://github.com/Simon4Yan/VisDA2020.

Effectiveness of SDA. To verify the effectiveness of source-to-target images translated by SDA for pre-training, we compared the pre-training performance with raw images and SPGAN translated images in Tab. 2.

Table 3. Ablation study on the effectiveness of the improved MMT. The results are evaluated on the target_val set. The experiments in this ablation study adopts the backbone of ResNet50-IBN [12]. All the post-processing techniques except the ensembling as described in Sec. 4.2 are used.

Pseudo-label-based method	mAP(%)	top-1(%)
Original MMT [7]	78.4	86.5
Our MMT+	**81.2**	**87.3**

Effectiveness of improved MMT. Compared to the original MMT, our proposed MMT+ achieves 2.6% improvements in terms of mAP (Tab. 3).

Performance of different backbones for MMT+. As shown in Tab. 4, we report the performances when training with different backbones. ResNeSt50 [25], ResNeSt101 [25], DenseNet169-IBN [11, 12] and ResNeXt101-IBN [21, 12] are adopted for ensembling in the final submission.

5   Conclusion and Discussion

In this work, we properly make use of both domain-translation-based and pseudo-label-based frameworks, which are both important for the unsupervised domain

Table 4. Performance of different backbones for MMT+. The results are evaluated on the target_val set. All the post-processing techniques as described in Sec. 4.2 are used.

Backbone for MMT+	mAP(%)	top-1(%)
ResNeSt50 [25]	83.6	89.4
ResNeSt101 [25]	82.7	89.1
DenseNet169-IBN [11, 12]	84.1	90.5
ResNeXt101-IBN [21, 12]	83.5	88.9
Ensemble	**86.3**	**91.2**

adaptation task on person re-ID. We also improve the state-of-the-art MMT framework to achieve better performance. Although we did not achieve the 1st place in the VisDA Challenge this year, we believe that the introduced methods have great potential in this research field. Also, there leaves some room for improvement and we will keep trying to make more progress.

References

1. Chang, W.G., You, T., Seo, S., Kwak, S., Han, B.: Domain-specific batch normalization for unsupervised domain adaptation. In: CVPR (June 2019)
2. Chen, Y., Zhu, X., Gong, S.: Instance-guided context rendering for cross-domain person re-identification. In: ICCV. pp. 232–242 (2019)
3. Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V., Le, Q.V.: Autoaugment: Learning augmentation policies from data. arXiv preprint arXiv:1805.09501 (2018)
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database (2009)
5. Deng, W., Zheng, L., Ye, Q., Kang, G., Yang, Y., Jiao, J.: Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In: CVPR (2018)
6. Ester, M., Kriegel, H.P., Sander, J., Xu, X., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: KDD. vol. 96, pp. 226–231 (1996)
7. Ge, Y., Chen, D., Li, H.: Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. In: International Conference on Learning Representations (2020), https://openreview.net/forum?id=rJlnOhVYPS
8. Ge, Y., Li, Z., Zhao, H., Yin, G., Yi, S., Wang, X., et al.: Fd-gan: Pose-guided feature distilling gan for robust person re-identification. In: NeurIPS (2018)
9. Ge, Y., Zhu, F., Zhao, R., Li, H.: Structured domain adaptation with online relation regularization for unsupervised person re-id (2020)
10. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR (2020)
11. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
12. Pan, X., Luo, P., Shi, J., Tang, X.: Two at once: Enhancing learning and generalization capacities via ibn-net. In: ECCV (2018)
13. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: NeurIPS, pp. 8026–8037 (2019)
14. Radenović, F., Tolias, G., Chum, O.: Fine-tuning cnn image retrieval with no human annotation. IEEE transactions on pattern analysis and machine intelligence **41**(7), 1655–1668 (2018)
15. Riccitiello, J.: John riccitiello sets out to identify the engine of growth for unity technologies (interview). VentureBeat. Interview with Dean Takahashi. Retrieved January **18**, 3 (2015)
16. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: ECCVW (2016)
17. Song, L., Wang, C., Zhang, L., Du, B., Zhang, Q., Huang, C., Wang, X.: Unsupervised domain adaptive re-identification: Theory and practice. arXiv preprint arXiv:1807.11334 (2018)
18. Sun, X., Zheng, L.: Dissecting person re-identification from the viewpoint of viewpoint. In: CVPR (2019)

19. Wang, D., Zhang, S.: Unsupervised person re-identification via multi-label classification. In: CVPR (2020)
20. Wei, L., Zhang, S., Gao, W., Tian, Q.: Person transfer gan to bridge domain gap for person re-identification. In: CVPR (2018)
21. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. arXiv preprint arXiv:1611.05431 (2016)
22. Yang, F., Yunchao, W., Guanshuo, W., Yuqian, Z., Honghui, S., Thomas, H.: Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. ICCV (2019)
23. Yu, H.X., Zheng, W.S., Wu, A., Guo, X., Gong, S., Lai, J.H.: Unsupervised person re-identification by soft multilabel learning. In: CVPR (2019)
24. Zhai, Y., Lu, S., Ye, Q., Shan, X., Chen, J., Ji, R., Tian, Y.: Ad-cluster: Augmented discriminative clustering for domain adaptive person re-identification. In: CVPR (2020)
25. Zhang, H., Wu, C., Zhang, Z., Zhu, Y., Zhang, Z., Lin, H., Sun, Y., He, T., Mueller, J., Manmatha, R., et al.: Resnest: Split-attention networks. arXiv preprint arXiv:2004.08955 (2020)
26. Zhang, X., Cao, J., Shen, C., You, M.: Self-training with progressive augmentation for unsupervised cross-domain person re-identification. ICCV (2019)
27. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: ICCV (2015)
28. Zhong, Z., Zheng, L., Cao, D., Li, S.: Re-ranking person re-identification with k-reciprocal encoding. In: CVPR (2017)
29. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. In: AAAI (2020)
30. Zhong, Z., Zheng, L., Luo, Z., Li, S., Yang, Y.: Invariance matters: Exemplar memory for domain adaptive person re-identification. In: CVPR (2019)
31. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV. pp. 2223–2232 (2017)
32. Zhu, X., Luo, Z., Fu, P., Ji, X.: Voc-reid: Vehicle re-identification based on vehicle-orientation-camera. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 602–603 (2020)