# Camera based person re-identification for VisDA 2020 challenge

Xiangyu Zhu, Zhenbo Luo
Ruiyan Technology
zhuxiangyu@ruiyanai.com

## Abstract

*In this paper, we present a simple but strong baseline for unsupervised domain adaptation (UDA) person re-identification. First, it generates pseudo label for unlabeled target domain images via cluster method. Second, both source domain and target domain pseudo labels are jointed to train a model. Finally, domain bias posed by camera is eliminated by camera-aware ReID model. Codes are aviable at https://github.com/Xiangyu-CAS/Yet-Another-reid-baseline*

## 1. Introduction

Pseudo-label-based method is simple but strong baseline approach in unsupervised domain adaptation (UDA). It shows competitive performance as well as maintaining easy to follow workflow. In this work, we propose an pseudo-label-based baseline with bag of tricks, including state-of-the-art data augmentation methods and losses. Besides, we explore the domain bias posed by camera and eventually prove it's even beneficial to eliminate the identity irrelevant bias in domain adaptation, which was initially proposed in our previous work VOC-ReID [1].

## 2. Proposed Approach

In section 2.1, we introduce our main framework, as well as bag of tricks and losses adopted to enhanced baseline. In section 2.2, workflows to generate pseudo label are listed. While in section 2.3, we describe the intuition and proper approach to eliminate camera bias.

### 2.1. Training Network

**Modeling.** Baseline architecture is almost the same with our previous work [1], including a backbone with IBN [2] structure, following with a generalized-mean pooling (GeM) layer and a batch normalization layer.

**Loss.** A combination of classification loss and metric loss is widely adopted in person re-identification, typically it's Cross-Entropy loss alongside triplet loss. We follow the formation but instead upgrade Cross-Entropy loss to popular large margin softmax, such as cosface, arcface and circle loss. In this challenge we choose cosface as classification loss due to slightly better performance compared to arcface and circle loss.

**Data Augmentation.** Strong data augmentation is crucial for generalized recognition and cross domain, which prevent the model from overfitting a specific domain. As reported by lots of works, color-jitter greatly improve performance when training and testing data come from different domains and vary tremendously in appearance. Otherwise, some advanced data augmentation methods are introduced in training, including Augmix [3], Auto-augmentation [4] and Random-Erase

### 2.2. Generate Pseudo Label

Reliable cluster methods are key component for generating pseudo labels. We compare several classic cluster methods in machine learning, which are described below.

**K-means.** Giving a hyper-parameter K, the algorithm will iteratively assign all samples with one of the K value, until reaching minimum variance. One of the major disadvantages of K-means is that most of the time we are not able to know how many clusters really exist in samples.

**HAC.** Hierarchical agglomerative clustering(HAC) is a bottom-up approach to iteratively merge close small clusters into large one. Similar to K-means, number of clusters is set as hyper-parameter of algorithm.

**DBSCAN.** DBSCAN is a density based method and doesn't require to set number of clusters as hyper-parameter. On the other hand, it leaves sparse distribution as noises which makes it relatively robust to noisy samples and finally becomes the most popular method in UDA person re-identification. Two distance metrics are compared using DBSCAN, euclidean distance and jaccard distance used in re-rank [5] .

### 2.3. Camera Bias

**Motivation.** Images captured from different cameras show variant appearance, such as background, image style and view point. We argue each camera can be treated as a

| method | param | recall | precision | f-score |
|---|---|---|---|---|
| K-means | 500 clusters | 14.21% | 43.58% | 21.42% |
| HAC | 500 clusters | 17.01% | 44.4% | 24.61% |
| DBSCAN(Euclidean) | minPts=4, eps=0.8 | 62.7% | 11.35% | 19.92% |
| DBSCAN(Jaccard) | minPts=4, eps=0.6 | 36.5% | 34.11% | 34.27% |

Table 1. Evaluation of cluster method on validation set.

individual domain in the broad sense. It would pose domain biases on similarity when cross-camera images are compared, which is identity irrelevant. Both identity relevant and identity irrelevant features are encoded in person re-identification and contribute to final similarity. Previous works aim to minimize identity irrelevant features by network attention mechanism itself, highlighting fine-grained parts in feature map and assigning less weight to background. However, it burden the network without using any prior knowledge. Since we know camera bias is a kind of identity irrelevant feature, it's a clear task to take advantage of this prior knowledge to emphasize identity relevant feature by simple minus. It is illustrated as equation written below, $f_{all}$ denotes global feature, $f_{id}$ and $f_{ir}$ denotes identity relevant and identity irrelevant features respectively.

$$f_{all} = f_{id} + f_{ir} \tag{1}$$

$$f_{id} = f_{all} - f_{ir} \tag{2}$$

**Camera ReID.** In this section we reform learning of camera domain bias as a camera ReID problem. The definition of camera ReID can be easily inferred from person ReID, it aims to retrieve images captured from one camera and share the same camera ID label. Similarity between features scores high value when two images come from same camera, otherwise the contrary. The next step is to eliminate camera bias, we have to subtract camera feature from global feature. Nevertheless, it's not possible to fuse global feature and camera feature directly as equation (1) described, because features generated by different network are not corresponded in channels. Luckily, similarity is one dimension scalar rather than vector, subtraction between similarities is reasonable. In equation below, $sim_{id}$ represents identity relevant similarity, while $sim_{all}$ and $sim_{cam}$ denotes similarity computed by global feature and camera feature. $\lambda$ is a coefficient of camera similarity, typically set as 0.1.

$$sim_{id} = sim_{all} - \lambda sim_{cam} \tag{3}$$

# 3. Experiments

## 3.1. Dataset

In VisDA 2020 challenge, there are four separated datasets, synthetic dataset as source domain (personx), unlabeled trainset from target domain, a small validation set from target domain and testset from target domain.

## 3.2. Implementation Details

To meet the requirement of pair-wise loss, data is sampled by m-per-class sampler with parameter P and K, which denote numbers of classes and numbers of instances per class in a mini-batch. In our experiments they are set number of instance per class equals to 16 and classes per batch equals to 4. All models are trained on a single GTX-2080-Ti GPU in total 50 epochs, with feature layers frozen in the first 5 epoch, which works as a warm-up strategy. Cosine annealing scheduler is adopted to decay initial learning rate, beginning from 3.5e-4. For final submission, we ensemble 5 model with various backbones, including resnet50-ibn, resnet101-ibn, densent169-ibn, se-resnet101-ibn and resnest50.

## 3.3. Cluster Comparison

Refer to face cluster, pair-wise recall, pair-wise precision and f-score are utilized as evaluation metric. Table 1 shows several popular cluster method trained on source domain and evaluated on validation set, DBSCAN with Jaccard distance achieves the best f-score. Thus, we use it to generate pseudo labels in this work.

## 3.4. Ablation Study on Validation

In Table 2 , we reports results of each component on validation set. Personx is source domain trainset provided by challenge committees, while personx-spgan is personx applied with target domain transfer by GAN. pseudo label denotes generating pseudo labels via DBSCAN on unlabeled target domain trainset. To clarify, three iterations of DBSCAN are performed, the first iteration relies on the best model trained on personx-spgan, the second iteration relies on model trained on personx-spgan and the pseudo labels generated by previous iteration, the third time is the same as the second. BN finetune is an extra finetune process on target domain. Similar to domain specific batch normalization(DSBN), BN finetune aims to shift the distribution of batch normalization from joint dataset to target domain.

Besides, post processing like re-rank, removing camera bias and model ensemble are also adopted, which greatly improve the performance. Particularly, removing camera bias benefits rank1 with more than 8% and mAP with 6%.

| Method | Performance | | | | | | |
|---|---|---|---|---|---|---|---|
| personx | ✓ | | | | | | |
| personx-spgan | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| pseudo label | | | ✓ | ✓ | ✓ | ✓ | ✓ |
| BN finetune | | | | ✓ | ✓ | ✓ | ✓ |
| re-rank | | | | | ✓ | ✓ | ✓ |
| camera bias | | | | | | ✓ | ✓ |
| model ensemble | | | | | | | ✓ |
| mAP | 33.2% | 37.7% | 51.8% | 55.5% | 73.4% | 79.5% | 82.7% |
| rank1 | 59.2% | 63.7% | 77.7% | 81.4% | 80.9% | 89.1% | 90.7% |

Table 2. Ablation study on validation set

| Rank | Team Name | mAP | rank1 |
|---|---|---|---|
| 1 | vimar | 76.56% | 84.25% |
| 2 | **Xiangyu(Ours)** | **72.39%** | **83.85%** |
| 3 | log | 79.05% | 83.26% |
| 4 | yxge | 74.78% | 82.86% |
| 5 | Archer2 | 70.39% | 79.70% |
| 6 | PIO | 69.89% | 78.77% |

Table 3. Leaderboard of VisDA 2020 challenge

### 3.5. Performance on VisDA 2020 Challenge

In the ECCV 2020 VisDA challenge, our submission ranks the second place on leaderboard, ranking by rank1. Performance of top teams are shown in Table 3. Specifically, our submission got high rank1 score but inferior in mAP, which might be the consequence of removing camera bias. From our observation, removing camera bias benefits rank1 much more than mAP.

## 4. Conclusion

In this work, we apply a pseudo label based method to handle unsupervised domain adaptation problem and validate the effectiveness of removing camera bias on domain adaptation, which brings much more increment than normally person ReID.

### Reference

## References


[1] X. Zhu, Z. Luo, P. Fu, and X. Ji, "Voc-reid: Vehicle re-identification based on vehicle-orientation-camera," in *Proc. CVPR Workshops*, 2020. 1

[2] J. S. Xingang Pan, Ping Luo and X. Tang, "Two at once: Enhancing learning and generalization capacities via ibn-net," in *ECCV*, 2018. 1

[3] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, "AugMix: A simple data processing method to improve robustness and uncertainty," *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020. 1

[4] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "AutoAugment: Learning Augmentation Policies from Data," *arXiv e-prints*, p. arXiv:1805.09501, May 2018. 1

[5] S. Bai and X. Bai, "Sparse contextual activation for efficient visual re-ranking," *IEEE Transactions on Image Processing*, 2016. 1