Supplementary Material for "Ranking Models in Unlabeled New Environments"

Xiaoxiao Sun, Yunzhong Hou, Weijian Deng, Hongdong Li, Liang Zheng
Australian National University
{first name.last name}@anu.edu.au

In the supplementary material, we 1) include the details of models ranked in the main paper, 2) provide experimental results when the DukeMTMC-reID dataset is used as the source, 3) provide more visual examples of the searched proxy sets, and 4) provide further discussion.

1. Person Re-identification Models

The main paper uses 280 models for ranking, which come from 28 representative baselines and approaches in person re-ID. These methods are selected from three popular Github repositories: Person_reID_baseline[1], reid-strong-baseline[2] and deep-person-reid[3]. Furthermore, for each method, we record 10 different versions corresponding to different epochs during training. Therefore, a total of $28 \times 10 = 280$ models are used.

The names of the 28 methods are shown in Table 1. Note that, although some methods use the same CNN architecture, such as ResNet50, their model accuracies are different because they use different training strategies or hyper-parameters (*e.g.,* learning rate, dimension of the FC layer output). Fig. 1 shows the mAP scores of the 280 models when trained and tested on a given dataset, such as the MSMT17 or Market-1501. Results show that these models have different image representation ability for person re-ID, so ranking them is feasible to reflect their relative representing performance on both target and proxy set.

Although the mAP scores of some models may be the same on a certain dataset, it will not influence the rank correlation evaluation since Kendall's tau can draw accurate generalizations for rankings with repeated rank [1].

2. DukeMTMC-reID as Source

Table 2 compares the quality of proxy sets in terms of Spearman's ρ and Kendall's τ when the DukeMTMC-reID and Market-1501 datasets are used as source and target, respectively. The result have similar trends to those in the main paper. For example, a weak correlation between the source and target sets is shown by the rank correlation coef-

Person_reID_baseline	reid-strong-baseline	deep-person-reid
IDE, PCB, DenseNet IDE-lr0.05, PCB-lr0.02, DenseNet-lr0.05, IDE-fix-bn, PCB-fix-bn, DenseNet-fix-bn	ResNet18, ResNet34 ResNet50, ResNet101, ResNet152, SeResNet50, SeResNet101, SeResNet152, SeResNeXt50, softmax, softmax-triplet, softmax-triplet-with-center, IBN-Net50-a	osnet-x0-25, osnet-x0-50, osnet-x0-75, osnet-x1-0, osnet-x1-0-cosinelr, resnet50-fc512, resnet50

Table 1: Names of methods that are used for model ranking in the main paper. "lr" represents learning rate.

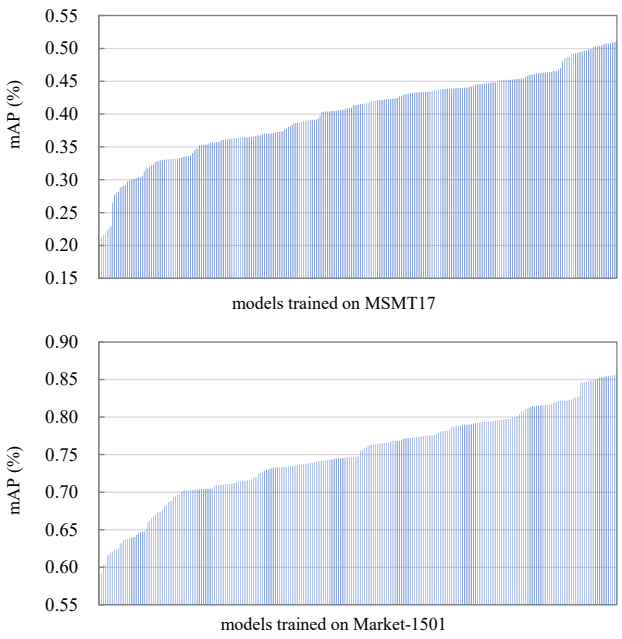

Figure 1: mAP (%) scores of 280 models trained and tested on the same dataset: **A:** MSMT17 and **B:** Market-1501.

[1] https://github.com/layumi/Person_reID_baseline_pytorch
[2] https://github.com/michuanhaohao/reid-strong-baseline
[3] https://github.com/KaiyangZhou/deep-person-reid

Source	Target		Individual Dataset							Other Dataset Generation Methods				Ours	
			CUHK03	Duke	Market	MSMT17	RandPerson	PersonX	UnrealPerson	Random	Attr. descent [4]	StarGAN [2]	pseudo-label [3]	w/o cam	w/ cam
Duke	Market	ρ	0.568	0.314	-	0.835	0.745	0.705	0.837	0.642	0.574	0.741	0.827	**0.866**	**0.893**
		τ	0.400	0.225	-	0.646	0.568	0.519	0.668	0.504	0.424	0.562	0.623	**0.698**	**0.706**

Table 2: Comparison of different proxy sets when using DukeMTMC-reID as source and Market-1501 as target.

A. source: Market-1501, target: MSMT17

B. source: DukeMTMC-reID, target: Market-1501

Figure 2: Image samples and compositions of searched proxy sets for different source and targets sets.

ficients $\rho = 0.314$ and $\tau = 0.225$. Further, the UnrealPerson dataset, when used as proxy, has higher correlation values of $\rho = 0.837$ and $\tau = 0.668$ with the target than the other individual datasets. Comparing with individual proxy sets and proxies generated by other methods, our proxy sets have higher rank correlation coefficients with the target set.

3. Image Samples of Proxy Sets

Fig. 2 shows the image samples and composition statistics of the searched proxy sets. We observe that the proposed method finds images with similar styles with the target, such as background color and illumination. For example, the searched images have various illumination conditions, and the illumination in MSMT17 also exhibits such characteristics (Fig. 2 **A**). Further, we observe that real-world data take up a larger proportion (*e.g.*, about 70% when using MSMT17 as the target) than synthetic data in the composition of the searched set. A possible reason is that real-world images have a small domain gap with the real-world target data.

4. Further Discussion

Can the proposed method generalize to other tasks? We discuss this question on image classification task by us-

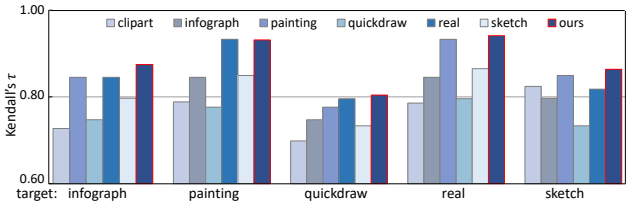

Figure 3: Comparison of different proxy sets on different targets of the DomainNet datasets over 70 models. Kendall's Rank Correlation τ is used as metric.

ing the DomainNet dataset, which has 6 domains, *i.e.,* Clipart, Infograph, Painting, Quickdraw, Real and Sketch, and 345 categories. We took Clipart as source and the other 5 in turn as target. The results are shown in Fig. 3. Our searched proxy achieves best results on four out of five targets and second best on the other target (painting).

The results suggest that 1) the chosen searching metric (FID and variance gap) is also effective in classification, 2) the proposed method a potential solution for ranking models of other tasks, such as image classification. However, most other tasks expect re-ID would require additional assumptions for evaluation, making the choices in candidate datasets limited based on existing datasets. For example, image classification requires the source and target domains to have the same classes. Therefore, the largest domain adaptation dataset, DomainNet, might still be suboptimal for investigating this problem because 1) it only offers 4 datasets (besides source and target) to construct the database pool, and 2) the distributions of the 4 domains (*e.g.,* sketch, real) are tremendously different.

Above limitations prevent our method from giving a clear margin over individual datasets are proxy, because the target will be approximated by mainly sampling images from one candidate rather than multiple. We will include above discussion and further study this problem by collecting data of other tasks in our future work.

Best models selected by proxy sets. Table 3 shows the mAP scores of the best models selected by different proxy sets (MSMT17 as source and DukeMTMC-reID as target).

proxy	Market	UnrealPerson	pseudo-label	random	ours	oracle
mAP (%)	36.98	37.16	36.70	36.05	38.10	38.12

Table 3: mAP scores of best models selected by proxy sets.

The model selected by the searched proxy (our) has the best performance on the target set, verifying the effectiveness of our approach.

References

[1] Haldun Akoglu. User's guide to correlation coefficients. *Turkish journal of emergency medicine*, 18(3):91–93, 2018. 1

[2] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Computer Vision and Pattern Recognition*, pages 8789–8797, 2018. 2

[3] Hehe Fan, Liang Zheng, Chenggang Yan, and Yi Yang. Unsupervised person re-identification: Clustering and fine-tuning. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 14(4):1–18, 2018. 2

[4] Yue Yao, Liang Zheng, Xiaodong Yang, Milind Naphade, and Tom Gedeon. Simulating content consistent vehicle datasets with attribute descent. In *Proceedings of the European Conference on Computer Vision*, 2020. 2