PERSON RE-IDENTIFICATION WITH REINFORCEDATTRIBUTE ATTENTION SELECTION

Ayesha Sameen
CSE Department,
Shadan women's college of Engineering and Technology,
Hyderabad, India
ayeshasameen781@gmail.com

Dr. Mohd Umar Farooq
CSE Department,
Shadan women's college of Engineering and Technology
Hyderabad, India
umarfarooq.mohd@gmail.com

Dr. K. Palani
CSE Department,
Shadan women's college of Engineering and Technology
Hyderabad, India
Principalswcet2020@gmail.com

***Abstract*— Person re-identification (Re-ID) aims to match pedestrian images across various scenes in video surveillance. There are a few works using attribute information to boost Re-ID performance. Specifically, those methods leverage attribute information to boost Re-ID performance by introducing auxiliarytasks like verifying the image level attribute information of two pedestrian images or recognizing identity level attributes. Identitylevel attribute annotations cost less manpower and are well- fitted for person re-identification task compared with image-level attribute annotations. However, the identity attribute informationmay be very noisy due to incorrect attribute annotation or lack of discriminativeness to distinguish different persons, which is probably unhelpful for the Re-ID task.**

***Index Terms*— Person re-identification, attentional module, sequential decision making.**

I. INTRODUCTION

GIVEN a query image, the goal of person re-identification (Re-ID) is to identify the images containing the same person from the gallery images captured by non-overlapping surveillance cameras. The previous re-id works either learndiscriminative features which are invariant across differentcamera views and illumination conditions, or learn the distance metric which can preserve the ranking order oftraining samples. In recent years, rising deep learningtechniques have been used for the Re-ID task and achieved excellent performance by learning discriminative features and good metric simultaneously in an end-to-end system.Despite the rapid progress in Re-ID, Re-ID models suffer from misalignment issue due to the pose variation, wrong bounding boxes, and the variations of the camera views. Recently, several methods have employed attentional models, which extract features from regions of interest,to alleviate misalignment issue and supremely improve the performance. Nevertheless, these methods only learn global attention on the entire human body and cannot capture fine-grained attention related to attributes (*e.g.*, cloth colors, age,gender) while such attribute information has been proveduseful for the Re-ID task in recent works [17], [18]. Inspiredby fully attentional block proposed in [14], we propose anovel Attribute Attentional Block (AAB), which combines both global and attribute attention to generate attention map. The attributes mentioned above are ID-level attributes because we assume that all the images belonging to the sameID should have the same attribute value. In Figure 1(a),we represents some examples of ID-level annotations for pedestrian images. In contrast, image-level attributes may be independent of persons, indicating that different images atdifferent surveillance videos from the same person could havedifferent attribute values. Thus, such attributes are unsuitable for Re-ID purpose. Also, image-level attribute annotations cost much more manpower compared with ID-level attribute annotations.

However, there exist several major issues when utilizing attribute information. First, Re-ID images are often of low quality (*e.g.*, low resolution, distracting persons, and wrong bounding boxes), which is likely to induce annotation errors of human annotators. Second, provided attribute annotation is usually ID-level attributes instead of image- level attributes because collecting image-level attributes is very time-consuming and expensive. For different images fromthe same ID, some attributes may vary along with the time (*e.g.*, vanishing attributes due to the occlusion caused by camera view change), resulting in the inconsistency of ID-levelattribute on the image level. Third, some of the attributesare rare (*e.g.*, a few persons wear hats in a cloudy environment) or not discriminative enough to distinguish different IDs(*e.g.*, black upper clothes are similar to each other in appearance), which might be unhelpful for the Re-ID task. Next, we show some examples of the noisy attributes. In Figure 1(b), we show two examples of annotation errors. Here the bounding boxes for the persons may be distracted by the persons whoare walking together, making the attribute annotation for the images transferred to the attribute values from the persons whoare also in the image. In Figure 1(c), we show two examplesof attribute inconsistency forID-level attributes caused by object occlusions.

II. RELATED WORKS

In this section, we provide a brief review of existing Re-ID methods from the following aspects.

A. Re-ID Based on Deep Learning

There are mainly two research lines for the Re-ID task: the first one is representation learning , which learns a discriminative ID-level representation and then calculate fea- ture distance between query and gallery images. While metric learning is to optimizes the distance metric. For example, to pull the images with

the same ID together by minimizing the feature distance and push away the images with different IDs by maximizing the feature distance

B. Re-ID Based on Attention Models

Re-ID models suffer from misalignment issue due to the pose variation, wrong bounding boxes, and variation of the camera views. To alleviate these problems, some methods try to match the invariant parts of the persons. PurifyNet refines the wrong ID labels for noisy images generated by wrong bounding boxes and reuses these images to train the model with refined labels to boost Re-ID perfor- mance. The attention mechanism.

C. Re-ID Based on Attribute

The attribute can be categorized into ID-level attribute and image-level attribute, as discussed in Section I. For image-level attribute, some previous works use the attribute information toselect tuples for metric learning, or mine the rules forverifying persons.There are more works using ID-level attribute.

D. Re-ID Based on Reinforcement Learning

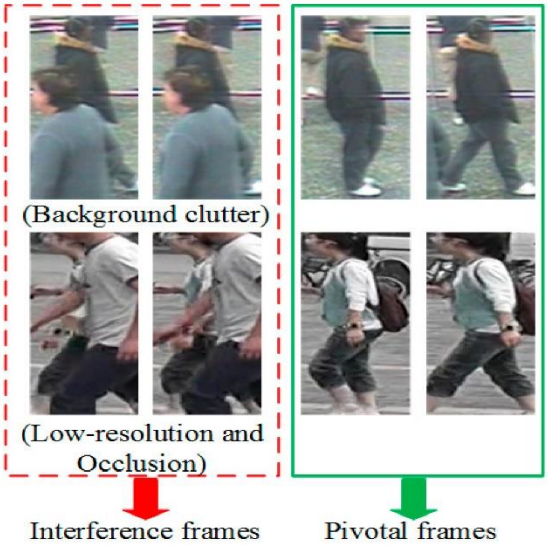
Reinforcement Learning (RL) trains an agent to exploit the experience during exploration and simultaneously maximize the cumulative reward. There are only a few works using RL for person Re-ID. For instance, Lan *et al*. proposed a method to refine the bounding box of persons to improve Re-ID task performance by RL.

III. OUR PROPOSED METHOD

Given a dataset composed of pedestrian images (each image contains a detected bounding box enclosing a pedestrian) fromdifferent persons captured by various surveillance cameras, we split the dataset into a training set and test set according to person ID, so that training IDs and test IDs have no overlap. Each training ID is associated with a suite of ID-level attributelabels. In the testing stage, the test set is divided into a setof query images and a set of gallery images. Inspired by FAB, we propose an Attribute Attentional Block (AAB), which utilizes attribute information to obtain fine-grained attribute attention. Following, we apply our proposed AAB module after the second, the third and fourth block of the ResNet50 backbone. The comparison between FAB and our AAB is shown in Figure 3. With an input feature tensor $\mathbf{F}$ with size $C \ H \ W$, we apply a 1 1convolution layer with $(M \ 1) \ C_r$ outputs and a ReLU layer to generate the grouped attribute feature $\mathbf{F}_g$ with size $(M \ 1) \ C_r \ H \ W$, in which we set $C_r \ C/16$ to reduce the number of parameters.

IV. EXPERIMENTS

We evaluate our method and baselines on three datasets: Market-1501, Duke MCMT-Re-ID and CUHK03 The detailed setting for each dataset can be found in Supplementary. Here, we will demonstrate how we get the attribute annotations for each dataset. *Market-1501:* Market-1501 [19] dataset contains 32668 images of 1501 IDs with annotated bounding boxes are detected and collected using the Deformable Part Model (DPM) pedestrian detector. View overlapping exists among different cameras, including 5 high-resolution cameras, and one low-resolution camera. The dataset into a training set with 12936 images of 751 persons and a testing set of 750 persons containing 3368 query images and 19732 gallery images.



1) Duke MCMT-Re-ID: The Duke MTMC-re-ID dataset is a subset of the Duke MTMC dataset specifically collected for person re-id. 1404 identities appear in more than two cameras while 408 (distractor) identities appear in only one camera. The training set consists of 16522 images with 702 identities and a testing set which consists of 2228 query images of the other 702 identities and 17661 gallery images of 702 iden- tities plus 408 distractor identities. We train our model with annotated ID-level attribute labels for 23 attributes providedin for the Duke MTMC-re-ID dataset. Similar to the Market-1501 dataset, we merge 8 attributes for lower-body clothing color and 7 attributes for upper-body clothing color, resulting in totally 10 attributes.

2) CUHK03: Following, we split all IDs into 767 training IDs and 700 test IDs. From each camera, one image is selected as the query for each ID and the rest of images are used to construct the gallery set. There are two ways of anno- tating bounding box, including labeled by human or detected by a detector. In this paper, we are using the detected bounding boxes. The dataset consists of 7365 training, 1400 query, and 5332 gallery images. Note that CUHK03 dataset does nothave attribute annotations, so we explore discovering latent attributes in an unsupervised manner. Without loss of general-ity, we assume that there are 10 latent attributes (M 10). The training process is the same as the other two datasets except that λ_{attr} 0 in Eqn. (16) due to the

absence of attribute supervision. In this case, our ASM can also learn to drop noisy latent attributes.

V. Implementation Details

Recall that we have three training stages. We train each stage for 400000 iterations. The initial learning rate is 0.001 for the first two stages and 0.0001 for the last stage. In each stage, after 200000 and 300000 iterations, the learning rate willbe discounted by 0.1. The training batch size is 64, including4 different images from 16 different IDs. Random flip and random erasing augmentation [56] are applied to the input images. Hard example mining is used for triplet lossWe adopt two metrics for evaluation: top-1 accuracy and mean Average Precision (MAP) which are commonly used forRe-ID evaluation.

To evaluate our proposed AAB and ASM modules, we con-duct extensive ablation studies in this section. We evaluate 21 special cases including our full-fledged method on Duke MTMC-re-ID. These special cases can be divided into three groups based on the stages used to train the models (see Section III-D). "Stage 1" means the models are only trained with Stage 1, "Stage
1 2" means the models are trained with first two stages, and "Stage 1 2 3" means the models are trained with all the three stages.

TABLE
NUMBER OF PARAMETERS (IN MEGA-BYTES) AND PERFORMANCE FOR DIFFERENT MODELS. RESULTS OBTAINED ON DUKEMTMC-REID

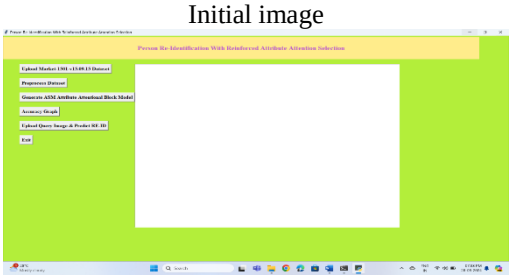
Method	mAP(%)	R1(%)	#Parameters(M)
w/o AAB, w/o PBM	66.6	80.2	23.5
w/o AAB	76.3	86.5	26.7
w/o ASM	78.7	88.0	30.2
APR [17]	55.6	73.9	24.4
AANet [18]	72.6	86.4	29.9
CA3Net [54]	70.2	84.6	31.0
Ours	80.4	89.9	30.7

attribute information, our method could achieve about 7% MAP improvement and about 3% R1 improvement on Market-1501 and Duke MTMC-re-ID datasets. This is because our proposed model with AAB can select key attributes and only leverage beneficial attribution information. Although CUHK03 dataset does not have attribute annotation, our method performs surprisingly well, which implies the potential of our method in discovering latent attributes in an unsupervised manner. Moreover, our method outperforms all the baselines and achieves the state-of-the-art results on all three datasets.

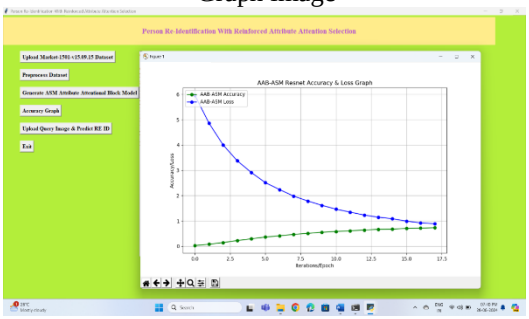
We visualize the attention map generated by our model withoutattribute supervision and report the number of dropped latent attributes. We observe that AAB can still generate diverse attention maps, but it is hard to provide a clear semantic explanation based on latent attributes.

VI. RESULT AND DISCUSSION

Initial image



Graph image



Result image

VII. FUTURE ENHANCEMENT

The scale of data has increased significantly in the re-ID community in recent years, e.g., from several hundred gallery images in VIPER and ILIDS to over 100k as in PRW and LSPS , which gives rise to the predominance of From both research and an application perspective, person re-ID in very large galleries should be a critical direction in the future. Attempts to improve both the accuracy and efficiency issues should be made. On the one hand, robust and large-scale learning of descriptors and distance metrics is much more important. This coincides with current research. Following large-scale image recognition, person re-ID will progress to large-scale evaluations. Although current methods address the re-ID problem between one or several pairs of cameras in a very limited time window, robustness in a camera network over a long time period has not been well considered. In, the re-ID consistency within a camera network is jointly optimized with pairwise matching accuracy, but the testing datasets (WARD and RAID) only have 3 and 4 cameras and less than 100 identities. In a network

with n cameras, the number of camera pairs. Considering the long recording time and lack of annotated data, it is typically prohibitive to train distance metrics or CNN descriptors in a pair-wise manner. As a consequence, training a global re-ID model with adaptation to various illumination condition and camera location is a priority. Toward this goal, an option is to design unsupervised descriptors which aim to find visually similar persons and treat visually dissimilar ones as false matches. But unsupervised methods may be prone to lighting changes.

VIII. CONCLUSION

In this paper, we have proposed an Attribute Attentional Module (AAB) to enhance the performance of person Re-ID task leveraging identity-level attribute information, in which diverse attention maps are learned based on iden-tity and identity-level attribute information. Moreover, noisy attributes are discarded by our designed Attribute Selection Module (ASM) sequentially using reinforcement learning. Comprehensive experiments on three benchmark datasets have demonstrated the superiority of our method.

IX. REFERENCE

[1] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep learning for person re-identification: A survey and outlook," 2020, *arXiv:2001.04193*. [Online]. Available: https://arxiv.org/abs/2001.04193

[2] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu, "Shape and appearance context modeling," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Rio de Janeiro, Brazil, Oct. 2007, pp. 1–8.

[3] I. Kviatkovsky, A. Adam, and E. Rivlin, "Color invariants for person reidentification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1622–1634, Jul. 2013.

[4] R. Zhao, W. Ouyang, and X. Wang, "Learning mid-level filters for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 144–151.

[5] Z. Zeng, Z. Wang, Z. Wang, Y. Zheng, Y. Chuang, and S. Satoh, "Illumination-adaptive person re-identification," *IEEE Trans. Multime- dia*, early access, doi: 10.1109/TMM.2020.2969782.

[6] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall, "Learning a Mahalanobis metric from equivalence constraints," *J. Mach. Learn. Res.*, vol. 6, no. 1, pp. 937–965, 2006.

[7] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith, "Learning locally-adaptive decision functions for person verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Portland, OR, USA, Jun. 2013, pp. 3610–3617.

[8] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *J. Mach. Learn. Res.*, vol. 10, pp. 207–244, Feb. 2009.

[9] Z. Wang *et al.*, "Person reidentification via discrepancy matrix and matrix metric," *IEEE Trans. Cybern.*, vol. 48, no. 10, pp. 3006–30.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classifica- tion with deep convolutional neural networks," in *Proc. Annu. Conf. Adv. Neural Inf. Process. Syst.*, Lake Tahoe, NV, USA, Dec. 2012, pp. 1106–1114.