Supplementary Material: Person Re-identification via Attention Pyramid

Guangyi Chen, Tianpei Gu, Jiwen Lu, *Senior Member, IEEE,* Jin-An Bao, and Jie Zhou, *Senior Member, IEEE*

TABLE I
ABLATION STUDY OF THE PARAMETER λ ON THE MARKET-1501 DATASET.

λ	**mAP**	**R-1**	**R-5**
$\lambda = 0.1$	89.5	95.6	98.5
$\lambda = 0.5$	90.2	96.1	98.9
$\lambda = 1$ **(Ours)**	90.5	96.2	98.8
$\lambda = 2$	90.4	95.9	98.8

A. More Parameters Analysis

A. *Influence of parameter λ*

As shown in Table I, we investigate the inference of different values of parameter λ to the APNet-C model on the Market-1501 dataset, including $\lambda = 0.1$, $\lambda = 0.5$, $\lambda = 1$, and $\lambda = 2$. The parameter λ denotes the balance rate between the cross-entropy loss and triplet loss. We observe that the performance is stable when the rate λ changes from 0.5 to 2. It demonstrates the model is robust to the change of triplet loss rate. Besides, when we further reduce the rate λ to 0.1, we observe the obvious performance reduction. It might be because the impact of the triplet loss can not match the one of cross-entropy loss. It indicates that the triplet loss is also important to train the model by reducing the intra-class distances and enlarge the inter-class distances.

B. *Analysis of multi-part model*

We find that the multi-part trick in MGN [1] is very effective on the CUHK03 dataset. We conducted an ablation study to analyze it. The multi-part trick adds a new local branch which splits the feature map into two parts and learns the local features. Specifically, the feature maps are split from the third residual block along the height dimension. In the inference, we connected the original global feature with local features for final matching. As shown in Table II, the multi-part trick can effectively improve the performance on both labeled and detected CUHK03 datasets. It is because the scale of the CUHK03 dataset is small, where the multi-part trick indicates the prior knowledge of human images. However, the effectiveness of this trick is limited when we add the scale of datasets (e.g., MSMT17 [2] or Market-1501 [3]).

C. *Analysis of model size and inference time*

In Table III, we also compare the model size and inference time of our channel-wised APNet and other methods. Compared with the baseline SE-ResNet [4], the extra complexity requirement of our APNet is limited. Compared with other methods such as Pyramid ReID [5] and SCSN [6], our APNet is more efficient by the "split-attend-merge-stack" principle.

TABLE II
ABLATION STUDIES OF MULTI-PART TRICK IN MGN [1] ON THE CUHK03 DATASET.

CUHK03	**Labeled**				**Detected**			
	mAP	**R-1**	**R=5**	**R-10**	**mAP**	**R-1**	**R=5**	**R-10**
APN-C	82.1	84.2	93.9	95.9	79.5	82.4	92.6	95.9
APN-C + Multi-part	85.3	87.4	95.4	96.8	81.5	83.0	93.7	95.9
APN-S	77.0	79.9	93.8	95.7	75.6	77.4	91.2	94.7
APN-S + Multi-part	81.1	83.5	93.2	95.8	78.1	80.9	93.4	94.7

TABLE III
GFLOPS, MODEL SIZE AND INFERENCE TIME OF DIFFERENT METHODS.

Methods	**GFLOPS**	**Model-size (M)**	**Inference (image/s)**
APNet-C	6.24	26.37	395
SE-ResNet [4]	6.24	26.02	526
Pyramid ReID [5]	9.96	31.04	287
SCSN [6]	7.24	31.87	326

B. More Comparisons

A. *Comparisons with Non-local [7] and TransReID [8]*

We compared the proposed method with other attention-based approaches including Non-local [7] and TransReID [8]. The Non-local [7] method applies the self-attention model to capture long-range dependencies. Non-local also adds the attention model in each block of the backbone network. TransReID [8] applies the vision transformer for the person ReID task. Different from other person ReID methods, TransReID uses the extra supervisory signals (e.g. Camera ID) to train the models.

Table IV summarizes the performance of Se-ResNet50 [4], Non-local [7], TransReID [8], and our APNet. The performance of Se-ResNet50 [4] and Non-local [7] is comparable and better than the one of ResNet baseline. It demonstrates the attention model is effective for the person ReID task. Compared our APNet with Non-local, we observe a significant performance improvement which shows that APNet can further improve the accuracy of the model by mining the multi-scale salient clues. Compared transformer-based method TransReID [8] with other CNN-based methods, TransReID cannot obtain the improved performance and uses the extra supervisory signals. Besides, TransReID needs 2.8 times the inference time over the ResNet50. It shows that the inductive bias of CNN (local relation) might be useful to recognize the high-structured objects such as the person.

1) Comparison with Pyramid ReID [5]: We further explain from what aspects is the proposed attention pyramid better than Pyramid ReID [5] as shown in Table IV. Pyramid ReI-

TABLE IV
COMPARISON WITH NON-LOCAL [7] AND TRANSREID [8] ON THE MARKET-1501 DATASET.

Method	**mAP**	**R-1**	**R-5**
ResNet50	87.8	95.0	98.6
Se-ResNet50 [4]	88.6	95.5	98.5
ResNet-50 + Non-local* [7]	88.7	95.6	98.5
TransReID [8] ViT-B/16	89.5	95.2	-
APNet-C	90.5	96.2	98.8

D [5] obtains the pyramidal features by the split and adds the classifier for each feature. Pyramid ReID [5] applies the "split-classification" principle, while our APNet proposes the "split-attend-merge-stake" principle. Compared with the principles of our APNet and Pyramid ReID [5], only the "split" principle is similar, while the "attend-merge-stake" principle is the main aspect that APNet is better than Pyramid ReID from. We will explain these differences in the following: 1) Different from Pyramid ReID [5] obtaining pyramidal features for classification, our APNet split features to learn the pyramidal attentions. The object of the pyramid structure is different. Pyramid ReID directly uses the split feature for further encoding and classification, while APNet learns the sub-attentions of each split feature. 2) By the "merge" principle, APNet only obtains one single feature of one person image. While Pyramid ReID [5] use all pyramidal features for multiple losses. The "merge" principle reduces the computing cost of the model. 3) By the "stack" principle, APNet applies coarse attention to guide the learning of the fine-grained attention. In Pyramid ReID, all pyramidal features with different pyramid levels are fed into a CNN block and a classifier. The "stack" principle smooths the learning of the attention model. 4) APNet introduces a multi-stage strategy that applies the pyramidal attention at the top of each residual block. This multi-stage strategy gradually guides the deep network to discover the salient clues. In the contrast, the pyramidal features in Pyramid ReID are only used before classification. 5) APNet can be implemented with any basic attention model, e.g. channel-wise attention or spatial attention. Pyramid ReID is only used for the spatial split.

B. Comparison with SNR [9], LAG-Net [10], and CBDB-Net [11]

The comparison with SNR [9]: SNR aims to improve the generalization capability of the person ReID model by filtering out identity-irrelevant interference with instance normalization and learning domain-invariant person representations. APNet focuses on mining the multi-scale salient clues to improve the accuracy of the ReID system by attention pyramid. The objective and motivation of these two methods are different. To evaluate the generalization capability of the person ReID model, SNR conducts the experiments for unsupervised domain adaptation person ReID tasks, whose source domain is labeled and target domain is unlabeled. While APNet conducted the experiments on the traditional person ReID task to evaluate the representation ability. Therefore, these two methods can not be directly compared. As shown in Table IV, we select the results of SNR when the source domain and target domain are consistent. We observe that the performance of SNR on the single domain is limited. It is because the aim of SNR is the cross-domain generalization but not the single domain accuracy.

TABLE V
THE COMPARISON ON THE VIPER DATASET.

VIPeR	R-1	R-5	R-10
LOMO+ XQDA	40.0	-	80.5
GOG + XQDA	49.7	79.7	88.7
ResNet50	36.7	69.0	80.4
APNet-S	46.2	75.1	84.4
APNet-C	48.9	77.4	88.2

The comparison with LAG-Net [10]: LAG-Net includes three streams, where GF-Stream is the baseline backbone network which uses a global pooling to extract global features, PF-Stream is similar with Pyramid ReID [5] and APNet to extract the part features by splitting, LA-Stream introduces the Region-Interest Map (RIM) and a local attention system to generate diversity local features. LAG-Net also learns the multi-granularity features and employs the attention model. However, the multi-granularity feature and attention model are applied to different branches. Different from LAG-Net, our APNet aims to learn pyramidal attention where our attention model is multi-granularity. Besides, APNet doesn't apply the multi-stream trick and only learns the single global feature. As shown in Table IV, we also add the comparisons with LAG-Net [10]. With only one stream, our APNet achieves the significant improvement on the Market-1501 dataset and obtain comparable results on the DukeMTMC-reID dataset, over the three-stream LAG-Net. We believe that LAG-Net can achieve further improvement by using our APnet.

The comparison with CBDB-Net [11]: CBDB-Net proposes to drop a part of the feature to improve the robustness. To drop the feature part, CBDB-Net also applies the "split" strategy. Different from CBDB-Net, APNet applies the "split-attend-merge-stack" principle to build the attention pyramid. The CBDB-Net doesn't use the other principles (e.g. "attend", "stack"), build the pyramidal features, or apply the attention model. As shown in Table IV, we also add the comparisons with CBDB-Net [11]. APNet outperforms the CBDB-Net [11] by a large margin on all three datasets.

C. EVALUATIONS ON VIPER

We also evaluated our method on the VIPeR [12] dataset. The VIPeR dataset contains 632 person images captured from two camera views. As argued in early deep learning based person re-identification methods such as JSTL [13], Spindle Net [14] and PDC [15], the amount of images in VIPeR is not enough to train the deep neural network. Therefore, these methods use the extra data to train the network. However, the experimental protocols of these methods are not consistent. Spindle Net [14] uses seven person ReID datasets to jointly train one model. While PDC [15] uses three datasets for joint training. In these settings, we cannot investigate performance behavior on a short dataset. Therefore, we directly train the baseline network and our model on the single VIPeR [12]

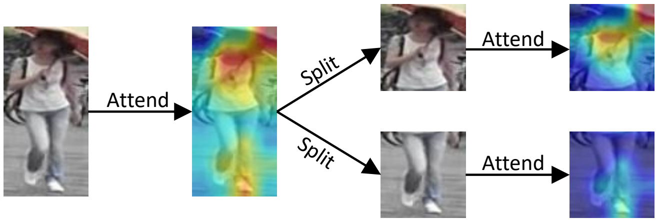

Fig. 1. Detailed visualization of attentions in two levels. We visualize the learning process of attention in two levels and show the differences. The level-1 attention is coarse-grained, which the foreground and salient information. While level-2 attention is more fine-grained. In this example, the level-1 attention mainly focuses on the salient umbrella, whole T-shirt, and legs, while level-2 attention further focuses on the umbrella, the logo of the T-shirt, and the shoes. The attentions of two levels are complementary in attention granularities, where level-2 attention is generated based on the level-1 attention.

dataset. As shown in Table V, we compare our methods (2-layer channel-wised APNet and spatial APNet) with the baseline network (ResNet50 without attention model) and the hand-crafted features including LOMO [16] and GOG [17]. We observe that the performance of learning based methods is lower than the hand-crafted ones. It is reasonable since the number of images in VIPeR is not enough to train the deep network. In addition, we observe that APNet improves the baseline network by a large margin. It indicates that the attention model is effective to guide the learning of deep neural networks with limited data and avoid over-fitting.

D. More Visualizations

To explain the difference between the two levels, we further provide a detailed visualization and analysis in Figure 1. First, level-1 attention is learned from the global image. This attention is always coarse-grained, which shows the foreground and salient information. In the example in Figure 1, the level-1 attention mainly focuses on the salient umbrella, whole T-shirt, and legs. These salient regions provide discriminative clues for person re-identification. In contrast, the level-2 attention focus on the fine-grained cues. In the upper half, the level-2 attention further focuses on umbrella and the logo of the T-shirt. In the other half, level-2 attention focuses on the shoes instead of the whole legs. The attentions of two levels are complementary in attention granularities. Besides, as shown in Figure 1, level-2 attention is generated based on level-1 attention. Thus, level-2 attention can be regarded as a fine-grained version of level-1 attention.

To show the learned attentions on the cross-dataset evaluation, we further visualize attentions under different training datasets (Market-1501 and DukeMTMC-reID) for the persons. In Figure 2, we show the attention maps trained with different training dataset on the same person in Market-1501. We observe that the attention maps is stable, even though the cross-dataset performance gap is huge. Besides, we find that the attentions trained with with cross-datasets are more diffuse, and the difference between level-1 and level-2 attentions become small. It is because of the domain shift, i.e., the attention trained with cross-dataset lacks of accurate supervisions to learn accurate and fine-grained attentions.

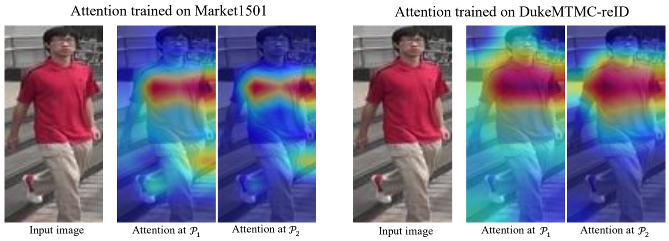
Attention trained on Market1501
Attention trained on DukeMTMC-reID
Input image
Attention at $\mathcal{P}_1$
Attention at $\mathcal{P}_2$
Input image
Attention at $\mathcal{P}_1$
Attention at $\mathcal{P}_2$

Fig. 2. The attention maps trained on Market-1501 and DukeMTMC-reID datasets. For two persons in Market-1501, we respectively display the attention maps trained on Market-1501 and DukeMTMC-reID datasets, from left to right. Despite the large domain gap, we observe that the attention maps trained on two different datasets are stable and accurate.

References

[1] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *ACMMM*, 2018, pp. 274–282.

[2] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in *CVPR*, 2018, pp. 79–88.

[3] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *ICCV*, 2015, pp. 1116–1124.

[4] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *CVPR*, 2018, pp. 7132–7141.

[5] F. Zheng, C. Deng, X. Sun, X. Jiang, X. Guo, Z. Yu, F. Huang, and R. Ji, "Pyramidal person re-identification via multi-loss dynamic training," in *CVPR*, 2019, pp. 8514–8522.

[6] X. Chen, C. Fu, Y. Zhao, F. Zheng, J. Song, R. Ji, and Y. Yang, "Salience-guided cascaded suppression network for person re-identification," in *CVPR*, 2020, pp. 3300–3310.

[7] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *CVPR*, 2018, pp. 7794–7803.

[8] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "Transreid: Transformer-based object re-identification," *arXiv preprint arXiv:2102.04378*, 2021.

[9] X. Jin, C. Lan, W. Zeng, Z. Chen, and L. Zhang, "Style normalization and restitution for generalizable person re-identification," in *CVPR*, 2020, pp. 3143–3152.

[10] X. Gong, Z. Yao, X. Li, Y. Fan, B. Luo, J. Fan, and B. Lao, "Lag-net: Multi-granularity network for person re-identification via local attention system," *TMM*, 2021.

[11] H. Tan, X. Liu, Y. Bian, H. Wang, and B. Yin, "Incomplete descriptor mining with elastic loss for person re-identification," *TCSVT*, 2021.

[12] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *PETSW*, vol. 3, no. 5, 2007, pp. 1–7.

[13] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *CVPR*, 2016, pp. 1249–1258.

[14] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang, "Spindle net: Person re-identification with human body region guided feature decomposition and fusion," in *CVPR*, 2017.

[15] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *ICCV*, 2017.

[16] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *CVPR*, 2015, pp. 2197–2206.

[17] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, "Hierarchical gaussian descriptor for person re-identification," in *CVPR*, 2016, pp. 1363–1372.