Supplementary Material

This Appendix provides a detailed illustration for Cross-camera Generalization Measure and detailed analyses of our proposed approach. Appendix 1 first reports more explanations and illustrations on CGM. We then provide analyses of HRCN in Appendix 2.

1. Explanation and Illustration on CGM

Detailed evaluation on CGM. In the evaluation of the query with ID V_{id}=i and camera C_{id}=A, we first remove images with camera A and id i in the retrieval. Then we calculate CGM in each camera with ID i independently, *e.g.*, to evaluate individual CGM with ID i and camera B in the retrieval, we remove images of the same ID i but with different cameras, *i.e.*, V_{id}=i, $C_{id} \neq B$, and then calculate it in Eq.10. The CGM of the query can be obtained by the average of individual CGM of each camera in Eq.11. All these evaluations are conducted with the same retrieval list but with different removes.

Position sensitivity on CGM. As mentioned in the main manuscript, the position sensitivity is one of advanced improvements on CGM. In Fig. 1, we conduct an experiment to explore the awareness of error position between AP and CGM. It can be observed that AP shows subtle value changes regardless of where the error position appears in the ranking list while the proposed CGM linearly transmits the error position penalties to correct backward samples. Moreover, gradients of the error penalties on the CGM decrease along with the increase of error samples, which ensures the importance of target images that occur in the head positions of a ranking list.

Representative case on CGM. To evaluate the superiority of the proposed CGM in cross-camera situations, we exhibit a representative case for cross-camera generalization capability between AP and CGM in Fig. 2. From Fig. 2 (a), although the image captured from camera 4 is in the last position of the sorted list, the AP measure is overwhelmingly high owing to excellent performances in the other cameras. However, considering the generalization in camera-level, the sorted results in (b) are better than those in (a). Hence we draw the a conclusion that the proposed CGM can well handle the cross-camera generalization problems.

Figure 1. Position-sensitive comparisons between AP and CGM. Assume an original list consisting of 30 correct samples and only one error sample is dynamically inserted into the list. We exhibit: (a) No static error sample (b) One static error sample in the middle position. AP values are not sensitive to the error position and tolerate a few error samples in the list. While CGM values are position-sensitive and also equally transmit the error effects to subsequent samples.

Table 1. Analyses of center pooling types on VeRi-776 dataset.

Pooling Type	mCGM	mAP	CMC@1	CMC@5
Annulus	0.611	0.817	0.968	0.986
Rectangle	0.627	0.828	0.970	0.985
Circle	**0.630**	**0.831**	**0.973**	**0.989**

Table 2. Analyses of pooling pyramid on VeRi-776 dataset. Pool.Num: the number of poolings.

Pool.Num	mCGM	mAP	CMC@1	CMC@5
2	0.628	0.828	0.967	0.987
4	**0.630**	**0.831**	**0.973**	**0.989**
8	0.612	0.817	0.969	0.985

2. Analyses of The Proposed Approach

Analysis of model parameters. We use python library THOP to compute parameters of our model. When the size of the first weight matrix in GRM are set to 512 and 1024, the parameters are respectively 35.30M (less than 38.19M in PAMTRI [29] and 58.03M (less than 38.19M in PVEN [25]). Moreover, we train a ResNet-101 with ibn-a blocks without these branches on VeRi-776. It only achieves 0.788 mAP, proving that huge parameters cannot lead to huge per-

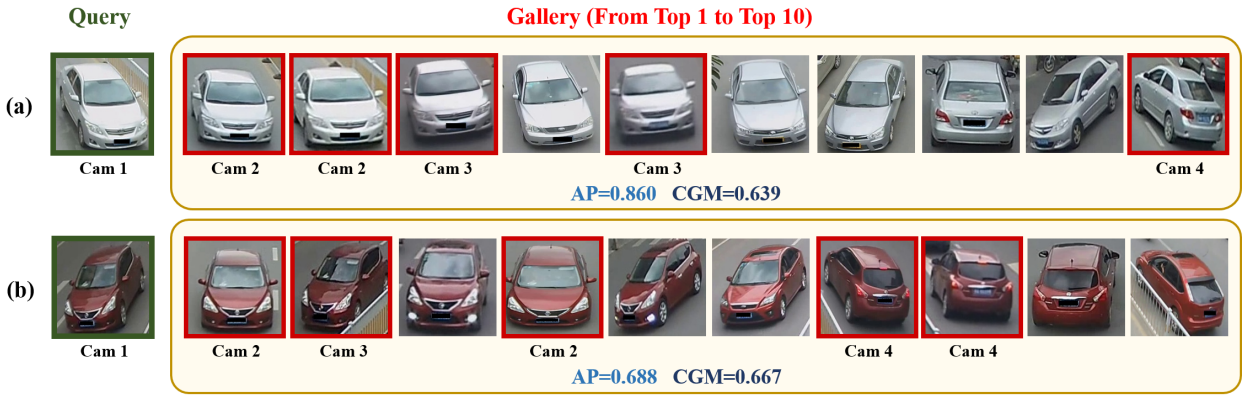

Figure 2. A representative case for cross-camera generalization. CGM in (a) is computed by $\frac{1}{3}\times(\frac{1}{2}\times(1+1)+\frac{1}{2}\times(1+\frac{1}{2})+1\times\frac{1}{6}) = 0.639$ and CGM in (b) is computed by $\frac{1}{3}\times(\frac{1}{2}\times(1+\frac{1}{2})+1\times 1+\frac{1}{2}\times(\frac{1}{4}+\frac{1}{4})) = 0.667$. Three items in outermost brackets represent camera 2, 3 and 4 respectively. Despite that the cross-camera generalization in the (b) is better than it in the (a), AP in the (a) is still much higher. Compared with AP, the proposed CGM can evaluate more accurately in the cross-camera generalization situations.

Table 3. Analyses of the number of features in each stage on VeRi-776 dataset.S_n denotes the n th stage.

Stage			mCGM	mAP	CMC@1	CMC@5
S_2	S_3	S_4				
3	3	1	0.601	0.813	0.967	0.988
3	1	3	0.626	0.825	0.968	0.986
2	3	3	0.620	0.826	0.972	0.987
3	3	3	**0.630**	**0.831**	**0.973**	**0.989**
4	6	3	0.619	0.825	0.968	0.985

formance improvement.

Analyses of center pooling types. To verify the effectiveness of the proposed circular pooling type, we conduct experiments to compare it with two pooling types, which are rectangular and annular in Tab. 1. For the rectangular center pooling, the value of the length and width is as same as the value of the diameter of circle. For annular center pooling, the first annulus is equivalent to the first circle while other bigger annuluses are formed by the subtraction of two adjacent circular center poolings. As shown in Tab. 1, the performance of circular center pooling is higher since the rectangular will bring more extra noise and it is easier for the annulus to split a discriminative region into several parts.

Analyses of center pooling pyramid. Considering the resolution of final feature maps from the ResNet-50 is 16×16, we select 2 pooling, 4 pooling and 8 pooling features, to construct a pooling pyramid and explore the optimal performance. In Tab. 2, it can be observed that the pyramid with 4 poolings shows better performance than the others.

Analyses of the number of features in each stage. We conduct experiments on the composition of features from stage S_2 to stage S_4. As shown in Tab. 3, it can be found that the account of features from stage S_4 shows the greatest influence and it does not seem to take effect to increase the number of features from stage S_2 and stage S_3. Moreover, unifying the number of features in each stage shows achieve the best performance in our experiments, which demonstrates the effectiveness of our proposed approach.

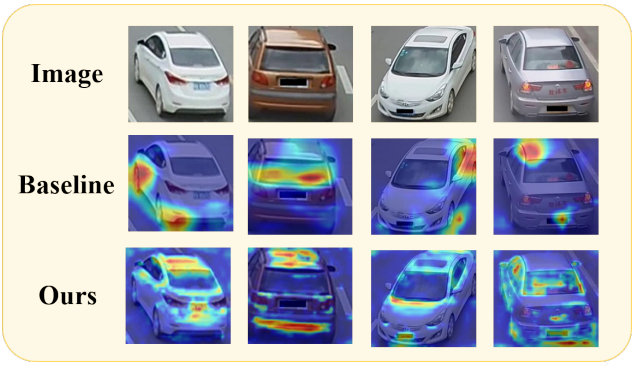

Figure 3. Visualization of class activations. Compared with baseline, the representation with heterogeneous relation complement in our proposed approach can focus on more discriminative regions.

Visualization analysis. In Fig. 3, it can be observed that the final representation in the baseline just focuses on some limited regions. In these concerned regions, vital discriminative parts are always ignored. In our method, we fuse high-level features with heterogeneous complementary features, which are lower level features and region-specific features, based on their relation. The final representation pays attention to more discriminative regions.