1 Supplementary Materials

1.1 More Visualization of Samples from different viewpoints

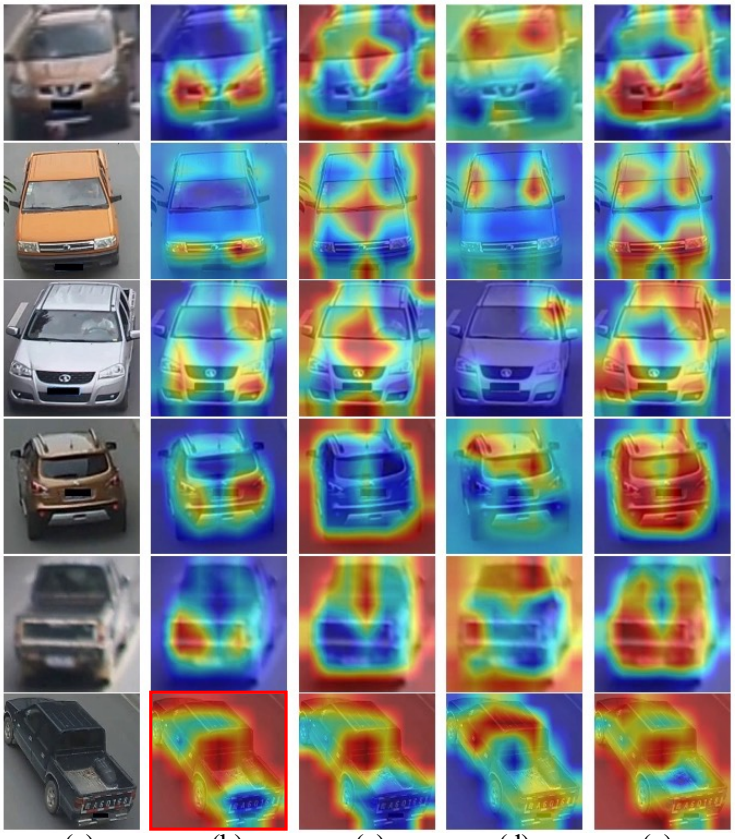

Fig. 1. More class activation mappings for the Rank-1 prediction of each decoupled feature. The first column shows the original images. The right four columns show the activation mappings learned by four decoupled features respectively. Red boxes mark some outlier examples.

As discussed in our main paper, we propose a transformer-based feature decomposing head and cluster-based decoupling constraint which aims to decouple the vehicle into different groups and each of them can focus on similar regions of interest (ROIs). The visualization presented in main paper mainly focus on the vehicle images captured from the front-side, and we add more visualization of vehicles from different viewpoints in Fig. 1.

We presented six examples in Fig. 1, where three of them are captured from the front-side and the other of them are captured from the backside. We visualize

Table 1. Ablation study of the influence of outliers during inference, where $Ex - outliers$ represents the exclusion of the similarity of outliers during inference, and we test on the raw decoupled features without enhancement from the global information.

	$Ex - outliers$	VeRi-776 (mAP)	VeRi-776 (CMC@1)
UFDN	×	77.3%	95.0%
UFDN	✓	77.5%	95.7%

the four decoupled features of these images, and each of them is based on the corresponding Rank-1 prediction. As shown in Fig. 1, the first column images are the raw images which will be fed to the UFDN, and we find that the decoupled features from the same part can focus on slimier information even though they are captured from different viewpoints, e.g., a) the first part of decoupled features (column b in 1) tend to focus on the lights and corners of vehicles; b) the second part of decoupled features (column c in 1) tend to focus on the contour information of vehicles; c) the third part of decoupled features (column d in 1) tend to focus on the upper or window information of vehicles; d) the forth part of decoupled features (column e in 1) tend to focus on the global information of vehicles and can provide more semantic information such as types, colors, or plates. The above phenomenon validates that our proposed UFDN provide several general cluster centers and vehicles from different viewpoints can be decoupled into parts according to them. Moreover, the decoupled feature (line 6. b in Fig. 1) marked with a red box is an outlier (which doesn't focus on vehicle light) and will be excluded during the alignment process.

It is worth noticing that the visualization results in 1 are based on UFDN (Res50) with a different seed while the ones in the main paper are based on UFDN (Swin-tiny), which shows our methods are consistently effective for both CNN-based and Transformer-based backbones.

1.2 Excluding of the Outliers During Retrieving

As mentioned above, we decouple a given vehicle feature into four parts, e.g., the lights of vehicles, the front part of vehicles, the upper part of vehicles, and the global information of vehicles. However, there exists some outliers as shown in Fig. 1 (images with red boxes) which can be concluded as: 1) some decoupled features lack the semantic information of the relative cluster-center due to different viewpoints or poses, such as the front window information is absent in the images captured from the backside. 2) some decoupled features contains some regions uncorrelated to the relative cluster center due to the conflicts during training.

Given the decoupled query images $Q \in R^{4\times m\times d}$ and gallery images $G \in R^{4\times n\times d}$, we try to exclude the influence of outliers as shown in Algorithm 1:

We conclude two main goals of the above algorithm: 1) if a query feature is an outlier of the relative cluster, then we will use a mask to exclude its distance from the final distance computation; 2) if a pair of query feature and gallery feature are all outilers, we will add their distance back since that they may both

ALGORITHM 1 Distance Computing for Inference

INPUT: Query images:$Q \in R^{4\times m\times d}$,
INPUT: Gallery images:$G \in R^{4\times n\times d}$,
INPUT: Cluster centers:$C \in R^{4\times d}$,
OUTPUT: $dist \in R^{m\times n}$
For $i \in (1, 4)$:
1) Compute the distance between query Q_i and gallery G_i:
$d_i = \sum_{j=1}^{d}(Q_i - G_i)^2$;
2) Compute the distance d_q between Q_i and C_i;
3) Compute the distance d_g between G_i and C_i;
4) mask the outlier features:
d_{m1}=where($d_q - m(d_q)$¿0,0,1).expend(m,n);
5) keep the distance when both the query and gallery feature are all outliers:
d_{m2}=where($d_q + d_g - m(d_q) - m(d_g)$¿0,1,0);
6) get the final mask $mask$:
$mask_i$=$d_{m1}|d_{m2}$;
7) compute the final distance matrix:
dist=$\frac{\sum_{i=1}^{4}(mask_i d_i)}{\sum_{i=1}^{4}(mask_i)}$.

taken from a backside. Take this into consideration drives our method to have a better performance as shown in Tab 1.

Discussions. Although the above algorithm achieve better performance, it needs additional operations on the final output features. Therefore, we use features without any post-processing in our main paper for a fair comparison with other works.

1.3 Comparison with TransReID

Table 2. Comparison with TransReID on three kinds of transformer-based backbones. It includes mAP and CMC@1 on VeRi-776; CMC@1 on VehicleID. For a fair comparison, all models are pre-trained on ImageNet-1K and no-extra annotation is employed.

			VeRi-776		VehicleID
Method	Backbone	input size	mAP	CMC@1	CMC@1 (S)
TransReID [2]	Vit-base	224 × 224	78.0	96.1	82.9
TransReID [2]	Vit-small	224 × 224	75.6	94.9	74.1
TransReID [2]	Swin-tiny	224 × 224	77.2	95.6	80.5
UFDN	Vit-base	224 × 224	78.5	96.4	83.2
UFDN	Vit-small	224 × 224	75.9	95.5	76.3
UFDN	Swin-Tiny	224 × 224	80.9	96.3	85.9

TransReID [2] is the first ReID network which brings the transformer into ReID, and we have compare our UFDN with it in the main paper. We provide more details about the performance of TransReID on both Vit-small and Vit-

base, which shows that TransReID (Vit-base) performs better than (Vit-small) and the error in our paper won't affect the final conclusion we achieved.

Moreover, we provide a comprehensive comparison between our methods and TransReID as shown in Tab. 2, which contains the Vit-small [1], Vit-base, and Swin-tiny [3] backbones. For a fair comparison, the JPM or other tricks of TransReID is not included in our experiments. When comparing our UFDN with TransReID under three different kinds of backbones, we find that our UFDN outperforms TransReID on all kinds of backbones, e.g., we obtain 80.9% on VeRi-776 while the pure TransReID without any extra information only obtains 77.2% which validates the effective and generalization of our proposed UFDN.

1.4 Discussion of Our Baseline.

We use TransReID [2] (78.2%) and Strong baseline [4] (79.6%) as baselines, which have been widely used in other works, e.g., partner learning [5], PGVR [6]. The significant improvements on strong baselines validate the effectiveness of UFDN. Our baseline does not introduce any extra parameters and uses default training settings.

References

1. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2020)
2. He, S., Luo, H., Wang, P., Wang, F., Li, H., Jiang, W.: Transreid: Transformer-based object re-identification. Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
3. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. arXiv preprint arXiv:2103.14030 (2021)
4. Luo, H., Gu, Y., Liao, X., Lai, S., Jiang, W.: Bag of tricks and a strong baseline for deep person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 1487–1495. Computer Vision Foundation / IEEE (2019). https://doi.org/10.1109/CVPRW.2019.00190, http://openaccess.thecvf.com/content_CVPRW_2019/html/TRMTMCT/Luo_Bag_of_Tricks_and_a_Strong_Baseline_
5. Qian, W., He, Z., Chen, C., Peng, S.: Partner learning: A comprehensive knowledge transfer for vehicle re-identification. Neurocomputing **480**, 89–98 (2022)
6. Qian, W., He, Z., Peng, S., Chen, C., Wu, W.: Pseudo graph convolutional network for vehicle reid. In: Proceedings of the 29th ACM International Conference on Multimedia. p. 3162–3171. MM '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3474085.3475462, https://doi.org/10.1145/3474085.3475462