Supplementary Materials for *PhD Learning: Learning with Pompeiu-hausdorff Distances for Video-based Vehicle Re-Identification*

Jianan Zhao[1], Fengliang Qi[1], Guangyu Ren[2*], Lin Xu[1†]
[1]Shanghai Em-Data Technology Co., Ltd. [2]Imperial College London

This supplementary material includes additional details of this paper in three parts, regarding: 1) More visualization results to show the rich spatio-temporal information and visual diversities (resolution, viewpoint, occlusion, and illumination) on our video-based vehicle re-identification (*VVeRI-901*) benchmark. 2) More empirical results on the *VVeRI-901* to verify the advantages of video-based vehicle re-ID over image-based ones. 3) Visualization of noise samples selected via the proposed Pompeiu-hausdorff Distance (PhD) loss to demonstrate its noise resistance ability.

1. Distinctive Characteristics of *VVeRI-901*

In this section, we present more detailed visualization results of video tracklets on our *VVeRI-901* to illustrate its distinctive characteristics. The *VVeRI-901* is collected from unconstrained capture conditions involving huge diversity of resolutions, viewpoints, occlusions, and illuminations in multiple traffic intersections as illustrated in Figure 1. Figure 1.(a-1) to (a-4) exhibit the visual diversities or variations in different levels, from simple resolutions to complicated mixed cases on the *VVeRI-901*. The spatio-temporal clues (e.g., visual appearances) in video tracklets are rich and change successively, which are beneficial for re-identifying a vehicle under complex surveillance conditions in the wild. In Figure 1.(b-1) and (b-2), we also display some samples from image-based vehicle re-ID benchmark (e.g., *VERI-wild* [2]). We can find that each vehicle identity only has a small number of sampled images in the dataset. The necessary spatio-temporal information is insufficient for complex recognition tasks.

2. Merits of Video-based Vehicle Re-ID

There are four re-ID strategies in terms of probe-to-gallery pattern [3], i.e., image-to-image, image-to-video, video-to-image, and video-to-video. The image-to-image strategy represents the image-based re-ID, where both query and gallery are images. The image-to-video mode can be regarded as a special case of multi-shot learning [3], and the video-to-image mode is a multiple-query method in image retrieval problems [1]. Among the four modes, the video-to-video mode uses the whole sequences as both query and gallery. We compared the performance of these four re-ID strategies on our *VVeRI-901* dataset. Figure 2 shows that the video-to-video outperforms the other three modes by a large margin in both mAP and Rank-1, while the image-to-image strategy shows the weakest performance. The empirical results demonstrate that the recognition performance can be boosted significantly by the video-based vehicle re-ID method, while the complexity is slightly increased. E.g., the video-based vehicle re-ID can converge and achieve 41% mAP with 400 epochs (52.17 seconds/epoch), while the image-based vehicle re-ID can only achieve 26% mAP with 400 epochs (51.88 seconds/epoch), as shown in Figure 2. It is worth noting that compared with the image-based re-ID, the additional burden brought by video is only the extra consumption of graphics memory, which can be covered by most prevailing GPU.

3. Advantages of PhD loss in noise resistance

Two kinds of partial and full occlusion induced noisy data, i.e., outliers (samples of visual appearances changed by partial occlusions) and label noise (samples of other categories introduced by full occlusions), are illustrated in the top four lines and the bottom four lines in Figure 3 respectively. We take two examples for illustration, each of which has four tracklets (e.g., $S1$, $S2$, $S3$, and $S4$) involving two IDs (e.g., $S1$ and $S2$ belong to an ID, while $S3$ and $S4$ belong to another ID), where each tracklet consists of four images. For video-based PhD loss, let $S1$ be the anchor set, and anchor-positive pairs, anchor-negative pairs are marked in yellow and red squares, respectively. Even though partially occluded samples (e.g., the second image of $S1$ in Figure 3.(a)) and fully occluded samples (e.g., the second and third images of $S2$ in Figure 3.(b)) are involved, the PhD learning method excludes these detrimental samples automatically due to the max-min optimization between two sets. The reasonable positive and negative pairs are chosen for the two cases from hard mining perspectives, indicating the promising performance of PhD loss in the noise resis-

*Work done while an intern at Shanghai Em-Data Technology Co., Ltd.
†Contact Author (Email: lin.xu5470@gmail.com)

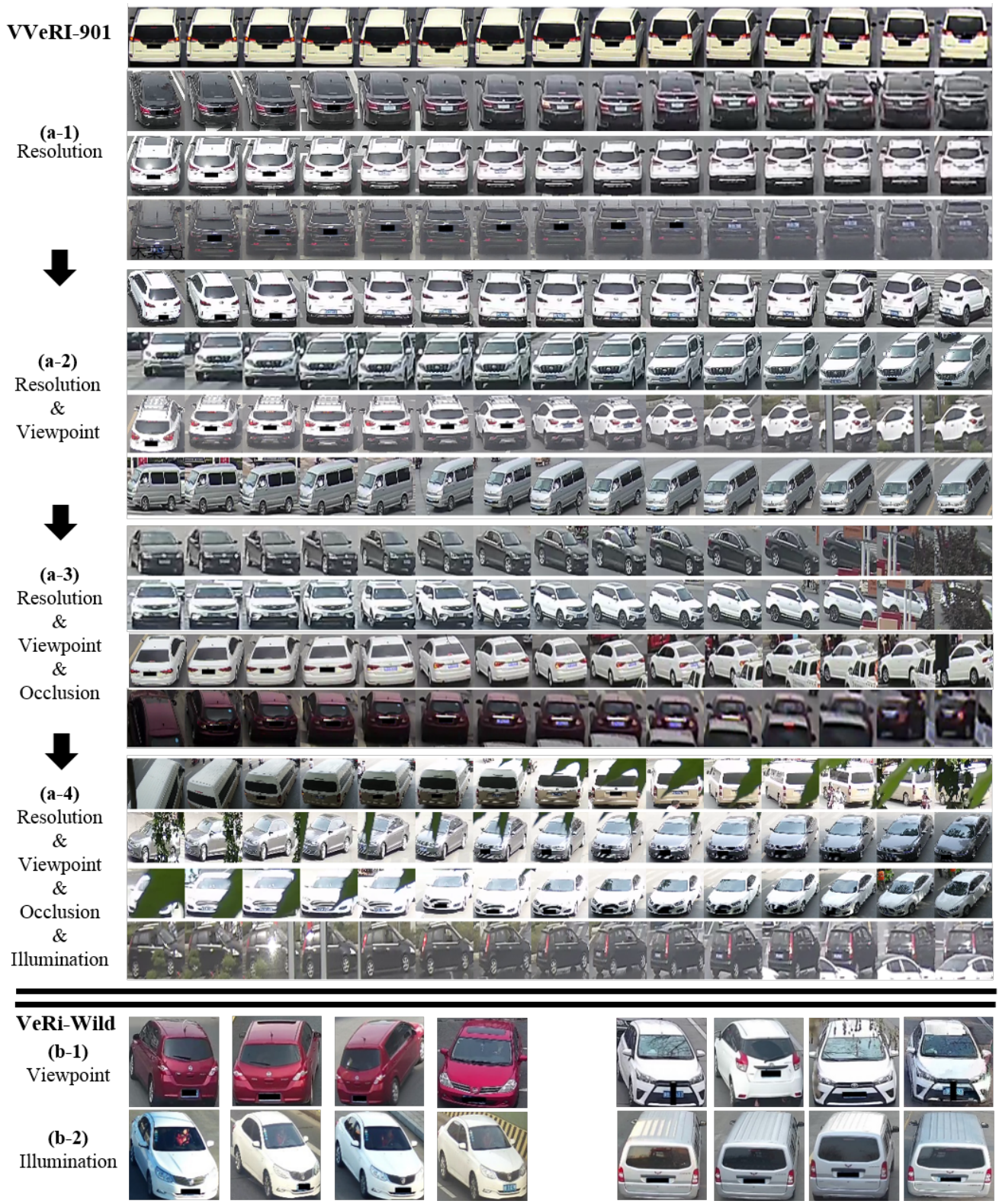

Figure 1. **Visualizations of some video tracklets on the proposed video-based vehicle re-identification (*VVeRI-901*) benchmark dataset**. The top subfigures (a-1) to (a-4) exhibit the sequential spatio-temporal diversities in resolution, viewpoint, occlusion, and illumination. The bottom (b-1) and (b-2) are instances from image-based VERI-Wild [2], which only provides a small number of sampled static images for each vehicle identity. Please also refers to the electronic edition (2×2 enlargement in PDF) for better visual effect.

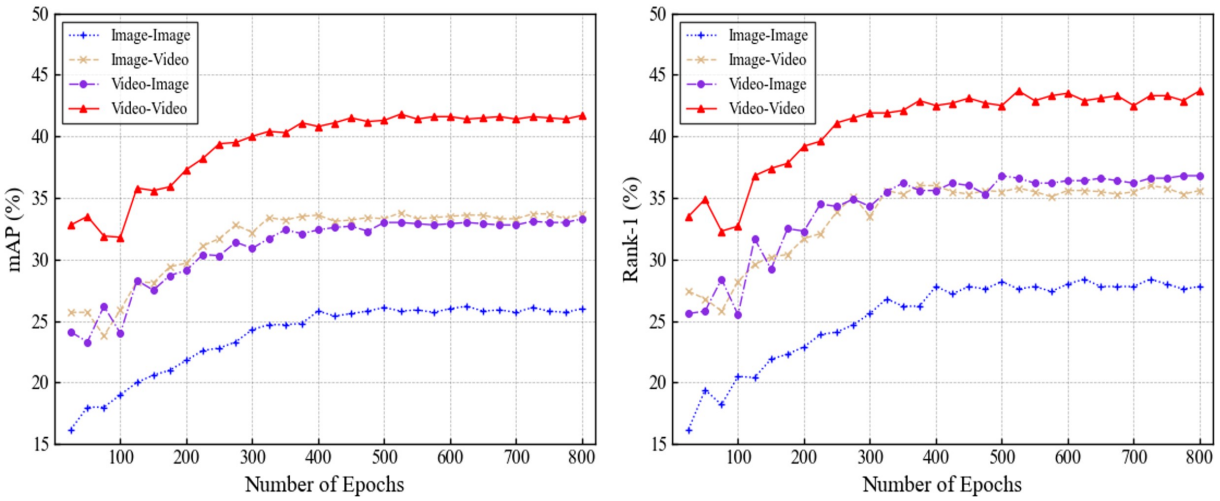

Figure 2. **The performance comparisons of the four re-ID modes on the *VVeRI-901* with the combination of triplet loss and ID loss**. The mAP and Rank-1 performance curves verify the advantages of video-based vehicle re-ID over other modes on the *VVeRI-901*.

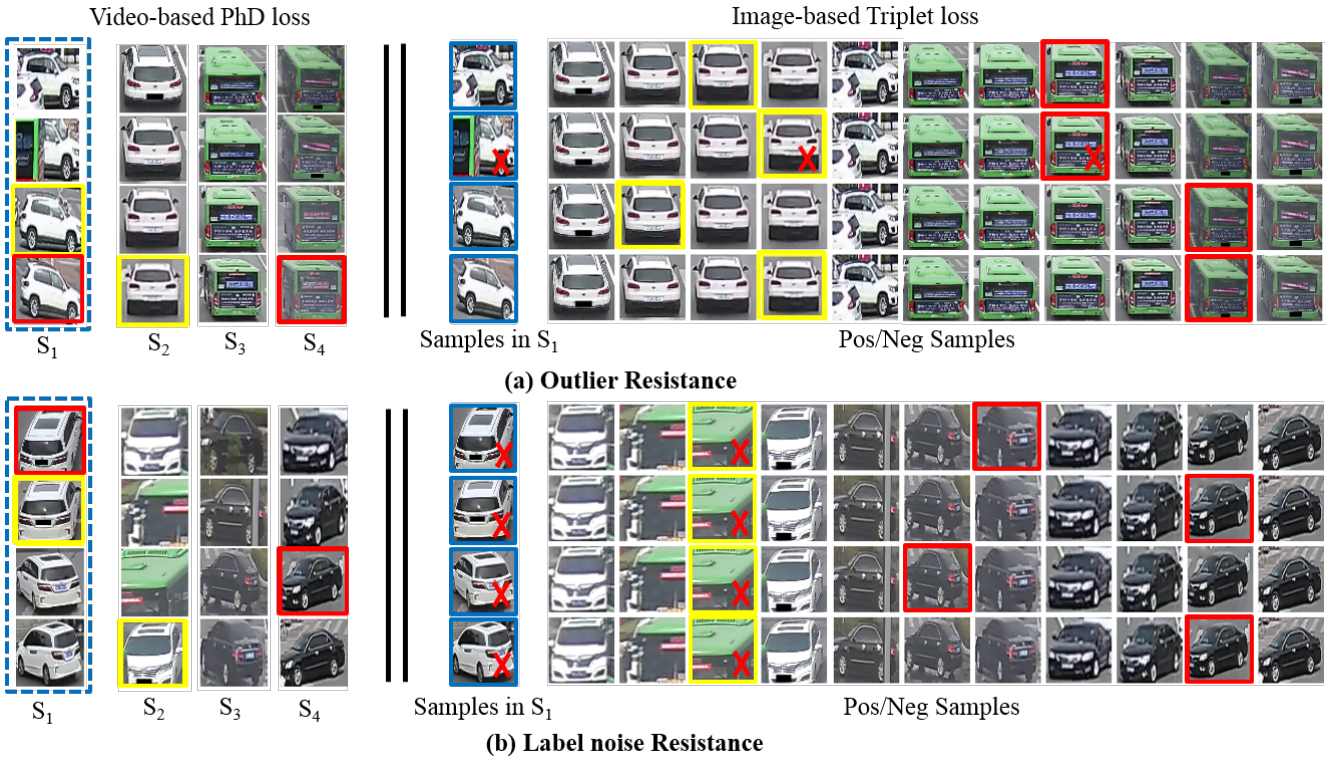

Figure 3. **Visualizations of positive and negative samples selected via PhD loss and Triplet loss objectives from a batch of samples on *VVeRI-901***. The blue, yellow, and red squares represent the anchor candidates for the positive and negative samples. Left side: set-based (i.e., video-based) PhD loss. Right side: point-based (i.e., image-based) triplet loss. The triplet loss is vulnerable to the occlusion induced noisy data(outliers and label noise), while the proposed PhD loss avoids selecting these noisy anchor-positive pairs automatically.

tance. However, for image-based triplet loss, anchors are individual samples in $S1$. The point-to-point triplet loss is susceptible to the noisy samples (as marked in the red cross) caused by occlusion. It might deteriorate the parameter optimization process during the training stage.

References

[1] R. Arandjelović and Andrew Zisserman. Multiple queries for large scale specific object retrieval. In *BMVC*, 2012. 1

[2] Yihang Lou, Yan Bai, Jun Liu, Shiqi Wang, and Lingyu Duan. Veri-wild: A large dataset and a new method for vehicle re-identification in the wild. In *CVPR*, June 2019. 1, 2

[3] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *ECCV*, pages 868–884, 2016. 1