Supplementary Material for READ: Reciprocal Attention Discriminator for Image-to-Video Re-Identification

Minho Shim[1][0000−0002−9637−4909], Hsuan-I Ho[2][0000−0001−8683−7538], Jinhyung Kim[3][0000−0002−2830−6365], and Dongyoon Wee[4][0000−0003−0359−146X]

[1] Seoul, South Korea minhoshim@minhoshim.com
[2] Department of Computer Science, ETH Zürich hohs@student.ethz.ch
[3] School of Electrical Engineering, KAIST kkjh0723@kaist.ac.kr
[4] Clova AI, NAVER Corp. dongyoon.wee@navercorp.com

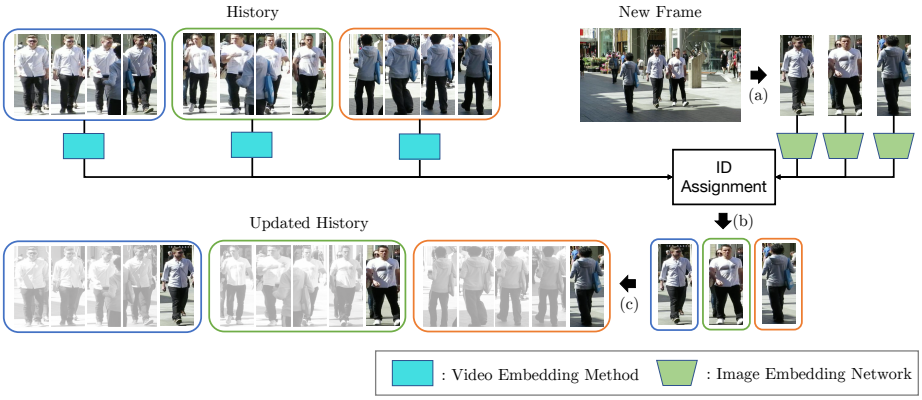

Fig. S1: Illustration of multi person tracking. One iteration consists of image feature extraction from a new frame, ID assignment, and history updates. See the main text for more details. Images are retrieved from MOT Challenge [3].

S1 Broader Impact

In this supplementary material, we present broader impact of our work in a practical point of view, portrayed under the multi person tracking setting. For multi person tracking, re-ID methods can be used for assigning ID to each detected person. Among three re-ID approaches, image-to-image (I2I) re-ID methods assign an ID to the given image by comparing it to the images from previous frames. Though it is simple to use, it cannot fully embrace the temporal information from previous frames. On the other hand, video-to-video (V2V) re-ID methods assign an ID to a query video by comparing it to gallery videos. However, it has limited usage, because IDs of images within each query video must be guaranteed to match. Resolving the problems above, image-to-video (I2V) re-ID can be an appropriate solution for multi-person tracking in real-world.

In Fig. S1, the I2V re-ID framework for multi-person tracking is illustrated. The tracking is an iterative process of:

Table S1: Experimental results of not using non-local blocks in the video embedding network. Experiments are conducted on DukeMTMC-VideoReID.

non-local	top-1	top-5	top-10	mAP
✗	86.0	94.6	95.9	82.4
✓	86.3	94.4	96.2	83.3

(1) For a given frame, a human detector detects every person in the frame as illustrated in Fig. S1(a). The cropped image of each detected person is converted into the image feature by the image embedding network.
(2) Each image feature is compared with the preceding sequence of images (history) of each ID to measure the similarity. For computing similarity, the distance between image feature and the history video feature of each ID can be used. In case of using discriminator, the output of discriminator can be used, otherwise. The ID with highest similarity is assigned to each person image, as depicted in Fig. S1(b). For newly detected person without any matching ID in history, new ID is assigned.
(3) Each image is appended into the history of the corresponding ID, as shown in Fig. S1(c).

Among several options to design the video embedding method, we deploy the non-local ResNet-50 [2,5] in the READ because its effectiveness is widely known for video embedding. For real-world tracking, however, use of the dedicated video embedding network in every iteration of tracking engages building tracklets of length T from history, followed by heavy inference through the video embedding network. This can be solved by removing non-local blocks from our video embedding network, i.e. exploiting the same architecture with the image embedding network so each video frame is embedded independently. When only the image embedding network is used, image features extracted by the image embedding network can be simply appended into history, then goes through our reciprocal attention discriminator with next query image features. So FLOP consumed by video embedding network can be saved in every iteration of tracking. The video embedding network is measured to consume 2.2x more FLOP compared to image embedding network, when dealing with video sample length of $T = 4$.

Concern on not using the dedicated video embedding network is when I2V re-ID heavily depends on non-local blocks, so the absence of the non-local blocks may lead to degradation of performance as reported in [1]. Thus we examine the impact of non-local block in Table S1. There are marginal 0.3 drop and 0.9 drop in top-1 accuracy and mAP respectively, without use of the non-local video embedding network. This is mainly because the reciprocal attention block in our work contains an attention mechanism across video frames and an image, that happens after the embedding networks. In sum, the READ is readily applicable to real-world applications such as multi person tracking while retaining its robustness to different video embedding methods.

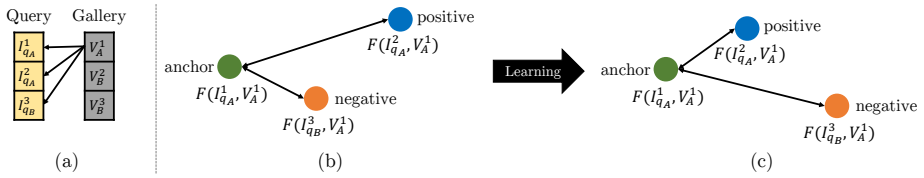

Fig. S2: Illustration of Reciprocal Attention Triplet Loss (RATL). With a slight abuse of notation, $I^1_{q_A}$ denotes the query image index 1 with identity A, and V^2_B denotes the gallery video index 2 with identity B.

S2 Step-by-step Details of the RATL

The RATL is illustrated in Fig. S2. (a) Here we show three query images and three gallery videos. Each operation of the RATL occurs on a basis of a gallery video. In the illustration, the basis is V^1_A. (b) Each circle is a feature generated with attention block, a query-specific understanding of V^1_A. In the viewpoint of $I^1_{q_A}$, $I^2_{q_A}$ is a positive image and $I^3_{q_B}$ is a negative. Accordingly, we let $F(I^2_{q_A}, V^1_A)$ and $F(I^3_{q_B}, V^1_A)$ be the positive and negative sample respectively against the anchor $F(I^1_{q_A}, V^1_A)$. (c) After training with the RATL, positive samples are encouraged to be more associated with the anchor, while increasing distance between anchor and negative samples. Above operation is iterated through each gallery video in the minibatch as a basis, V^2_B and V^3_B in the case of the illustration.

S3 mAP Calculation

When evaluating mAP, we used the implementation of the popular scikit-learn [4] package. As of version 0.19, the library has changed the way mAP is calculated. With the new version, the mAP scores 2-3 more than the old version in our setting. We emphasize that we use the version <0.19 that calculates the mAP in the same manner with the mAP implementation used by Gu et al. [1].

References

1. Gu, X., Ma, B., Chang, H., Shan, S., Chen, X.: Temporal knowledge propagation for image-to-video person re-identification. In: ICCV (2019)
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
3. Milan, A., Leal-Taixé, L., Reid, I., Roth, S., Schindler, K.: MOT16: A benchmark for multi-object tracking. arXiv preprint (2016)
4. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. JMLR **12**, 2825–2830 (2011)
5. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: CVPR (2018)