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

Abstract. Person re-identification (re-ID) is the problem of visually identifying a person given a database of identities. In this work, we focus on image-to-video re-ID which compares a single query image to videos in the gallery. The main challenge is the asymmetry association of an image and a video, and overcoming the difference caused by the additional temporal dimension. To this end, we propose an attention-aware discriminator architecture. The attention occurs across different modalities, and even different identities to aggregate useful spatio-temporal information for comparison. The information is effectively fused into a united feature, followed by the final prediction of a similarity score. The performance of the method is shown with image-to-video person re-identification benchmarks (DukeMTMC-VideoReID, and MARS).

Keywords: Image and Video Understanding · Identity Retrieval · Re-identification · Attention

1 Introduction

Computer vision is all about teaching machines to identify and understand objects in images and videos. Among all, identification is to distinguish between objects within the same category. Practical applications of identification include unmanned surveillance, human-robot interaction, and so on. Specifically, person re-identification (re-ID) is the problem of identifying a person from a given set of person identities. The task involves multiple views of a moving person taken from a single or multiple cameras, so person re-ID suffers from pose variance, illumination change, occlusion, and background clutter. For this reason, re-ID demands a fine-grained level of image understanding.

For example, domain knowledge such as body parts [38,14] or pose [25,26] can be used to extract fine-grained features like head, torso, and so on. To restrain using domain knowledge, attention based approaches [2,6,36,18] are proposed to adaptively find locations of interest. However, these methods do not consider the semantic relationship between objects since attention independently occurs within one object of a single identity.

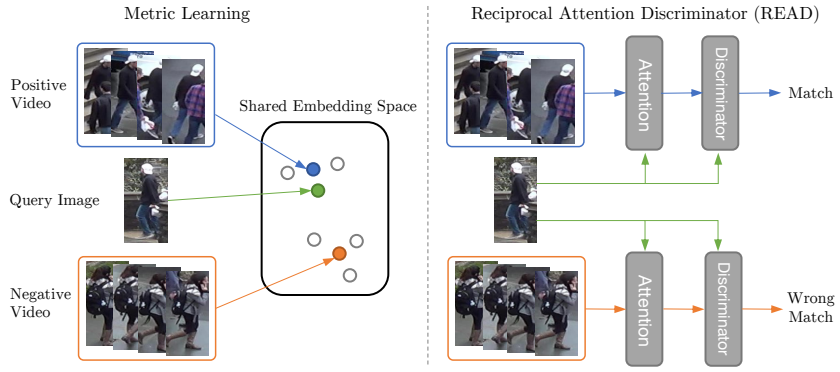

Fig. 1: Comparison between metric learning and the proposed Reciprocal Attention Discriminator in the image-to-video person re-identification setting. Metric learning projects images and videos into a shared embedding space, but those features are independent and do not refer to each other at the moment of comparison. Our Reciprocal Attention Discriminator creates a reciprocally coalesced feature using attention that occurs across images and videos.

To compare two images, it is natural for a human to perceive images side-by-side, just like spot-the-difference puzzle game. For instance, if a person is wearing a eye-catching furry red scarf, we can first try to find whether a person in another image is also wearing the red scarf. However, existing person re-ID methods focus on extracting one person's feature independently without consideration of the others.

Image-to-video (I2V) re-ID is a task of comparing a single image (query) to videos (gallery). The I2V re-ID task brings out another challenge for identification. Contrary to image-to-image (I2I) or video-to-video (V2V) re-ID, I2V re-ID is about connecting bridges between image representations [8,29] and video representations [30,1,5]. Existing I2V re-ID methods [35,32] suggest to project images and videos into a shared embedding space. Instead, another approach proposes to transfer representative power of video embedding network to image embedding network [7]. However, these metric learning based approaches encourage image and video information to resemble each other even though a video is innately different from an image, because of the temporal dimension (Fig. 1).

To this end, we devise the **Re**ciprocal **A**ttention **D**iscriminator (READ) for image-to-video re-identification. First, the READ is designed with a reciprocal attention structure. The attention is reciprocal because it not only uses self-attention within each gallery video, but also promotes the observation to occur across between query images and gallery videos' spatio-temporal dimension. In addition, this attention mechanism efficiently aggregates the temporal dimension of videos, which naturally solves the aforementioned asymmetry problem. Compared to average pooling multiple image features across the temporal di-

mension, such mechanism enables more expressive power of video embedding. The module aggregates videos' spatio-temporal information with attention, and then efficiently fuses image-video feature maps into a united feature. Finally, instead of measuring similarity based on the distance between image embedding and video embedding, the READ uses discriminator in order to actually observe query and gallery at the same time to calculate similarity score.

Extensive experiments show the effectiveness of the READ on large scale I2V re-ID benchmarks: DukeMTMC-VideoReID (Duke) and MARS. In I2V re-ID benchmarks, pedestrian image sequences from multiple camera views formulate gallery, while each query is a still image.

In summary, we make following contributions:

- We propose the READ, a novel attention based discriminator, to deal with asymmetric image and video information in I2V re-ID.
- We train the READ in an end-to-end manner by designing two losses and sampling strategy.
- We demonstrate the effectiveness of our method on two benchmark databases Duke and MARS. The READ outperforms previous I2V re-ID methods on both datasets.

2 Related Work

Person re-identification (re-ID) is a branch of identity retrieval task that usually involves multiple camera views to identify a bounding box pedestrian image from existing ID images in the gallery. The field of re-ID can be divided into three major branches, image-to-image (I2I) re-ID, image-to-video (I2V) re-ID, and video-to-video (V2V) re-ID. Considering its practicality, re-ID has a large body of literature so we focus on recent advances that relate to our work. We refer readers to [42] for more comprehensive review on re-ID.

Recent I2I re-ID focuses on data-driven approaches to learn features suitable for classifying IDs and computing distance between images [9,3,24]. Triplet based loss functions have been extensively studied [9,3], as they can formulate distances between features to follow similarities among IDs, i.e. features from images with similar appearances are close to each other. Spatial-attention approaches [39,20,15,36,2] or part-based models [38,25,28,13,26,27,11] are adopted to further guide CNNs to filter out unnecessary segments of images and concentrate on interesting parts, especially human bodies in the case of pedestrian images.

Zhang et al. [36] proposed key-value memory matching which utilizes an attention-based matching mechanism, for computing similarity between images represented by position-aware key-value memory. They showed that the attention module could attend semantically corresponding regions, e.g. body parts, bags, or shoes. As of constructing triplet samples for training, Zhang et al. [37] uses hard identity sampling and multi-stage training strategy for maximizing margins between distant identities. On the other hand, our reciprocal attention module explicitly observes each pair of a query and a gallery identity for comparison.

In other words, attention is not only applied within one identity, but it also occurs across identities to determine the best spatio-temporal locations suitable for distinguishing identity.

I2V re-ID is a problem domain of re-ID comparing a single image to a sequence of images. Early works [32,44,43,35] focused on embedding images and videos into a shared feature space. Gu et al. [7] proposed a training method to solve the lack of temporal information in image by transferring knowledge from video representation network to image representation network and building a unified feature space. In addition, non-local neural network [33] is intensively embedded into recent video embedding networks for re-ID [7,18], alongside with triplet loss [9]. While the video embedding networks are able to extract self-attention features, it cannot handle asymmetry features of I2V re-ID. In contrast, our model learns correlation between a query image and a video in an end-to-end manner, thus the information asymmetry between image and video is naturally solved and extra steps of training different networks are not demanded. In this paper, we search for a method to embrace the advances made so far while wisely handling immanent problems of I2V identity retrieval task.

3 Proposed Method

In this section, we explain our new I2V re-ID framework (Fig. 2); in the order of the image and video embedding sub-networks, the Reciprocal Attention Discriminator (READ), and its training strategy. The framework measures the matching probability given two types of inputs, an image query I_q, and a video V_i sampled from the gallery G. The gallery G consists of videos for each identity, $G = \{V_i | i \in [1, 2, ..., M]\}$, where M is the number of videos in the gallery. The image and video embedding sub-networks respectively encode each input as 2D or 3D feature maps, which serve as intermediate representations for reciprocal comparison in the READ. The comparison is to observe one image and one video at the same time, and to create an attention map to aggregate spatio-temporal information from the video. The query image information and query-specific understanding of the video is then combined together for determining the final similarity score.

3.1 Image Embedding Network

Spatial information of the query image should be maintained to compare each spatial location of query against global spatio-temporal location of gallery video. Given a query image I_q, the image embedding network extracts a 2D feature map $f_I \in \mathbb{R}^{H \times W \times C}$ from $res4$ layer of ResNet-50 [8] in order to maintain spatial information [28]. The parameters are initialized with the weights pretrained on ImageNet [4].

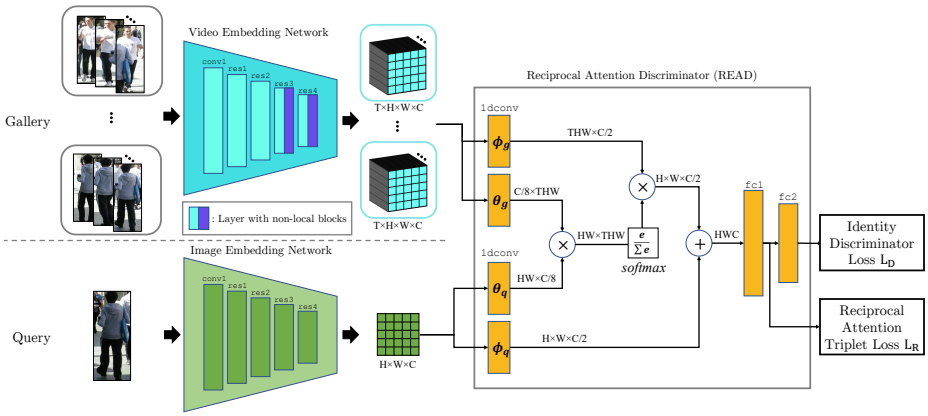

Fig. 2: Illustration of Reciprocal Attention Discriminator (READ), showing an example of comparing one query against multiple gallery videos. In the training phase, multiple queries and gallery sequences will form a minibatch and used for computing losses. ϕ and θ are 1D convolution blocks for channel size reduction. $\oplus$ denotes concatenation, and $\otimes$ denotes matrix multiplication.

3.2 Video Embedding Network

The video embedding network extracts a 3D feature map $f_{V_i} \in \mathbb{R}^{T\times H\times W\times C}$ from a video input V_i. Each video $V_i = \{I_j | j \in [1, 2, ..., L]\}$ includes images of a specific identity. The length L of the video sequences might vary within the gallery G, so we sample a fixed number of frames from videos (Sec 3.4). In the video embedding network, we add two non-local blocks to $res3$ and three non-local blocks to $res4$ on ResNet-50 [8]. The non-local blocks enable self-attention within the video, where each location combine information from global spacetime locations. Unlike [7], we prune out $res5$ layer just like the image embedding network to keep the number of channels small enough to keep computation feasible.

3.3 Reciprocal Attention Discriminator (READ)

Given a query embedding and a video embedding, the READ measures matching probability of two different embeddings. The discriminator D consists of a reciprocal attention block, and fully connected layers. We use a spatio-temporal non-local attention block [31,33,5,22,19] to compare query feature maps against gallery feature maps. In short,

$$F(I_q, V) = fc1([softmax(\theta_g(V)^T \cdot \theta_q(I_q)) \cdot \phi_g(V), \phi_q(I_q)]), \tag{1}$$

$$D(I_q, V) = fc2(relu(F(I_q, V))), \tag{2}$$

where θ and ϕ are 1D convolution layers as bottleneck of $C/8$ and $C/2$ respectively, and $[\cdot,\cdot]$ is concatenation. By Eq. 1, we can get a reciprocal attention feature $F(I_q, V) \in R^{d_F}$, where d_F is the output channel size of the first fully connected layer *fc1*. This feature $F(I_q, V)$ is used for Reciprocal Attention Triplet Loss described in Sec. 3.5. The attention block first applies parallel 1D convolutions θ_q and θ_g to the query and gallery with a bottleneck of $C/8$, followed by a dot-product similarity function normalized with softmax. The softmax is applied over dimension of gallery. Then, final feature map is generated by applying the softmaxed attention map to feature maps of gallery, concatenated by query feature maps, where those feature maps are generated with ϕ_g and ϕ_q, a bottleneck of $C/2$. Then, the output feature map is flattened to go through a pair of fully connected layers (*fc1* and *fc2*) with a ReLU activation in between.

Computing attention across two features, embedded from the query image and the gallery video with possibly different identity, might look astray. However, the proposed setting still acknowledge the concept of non-local operation, while effectively solving the asymmetry between image and video. When embedding a video, the idea of the non-local operation is to enrich information of a source pixel by integrating information from pixels in global spatio-temporal locations. In the case of previous re-ID methods, the *source* and *global locations* remain within a video of a single identity. Instead, we let the READ bring global information from gallery, across different modalities and no matter the identity of the gallery matches with a given query. In the end, the network first observes a query image and a gallery video, to determine which spatio-temporal locations are valuable for comparison.

3.4 Sampling

Training the READ involves two types of sampling, 1) sampling a subset of frames from videos, and 2) sampling pairs of query and gallery. In order to sample a subset of input video frames, we utilize restricted random sampling (RRS) [14]. Each video of variable length L is divided into T splits with equal duration, and one frame is randomly sampled from each split, resulting in T images per video. The randomness of the sampling method naturally leads to data augmentation and regularization. We empirically found RRS works better than sampling sequential frames. Following the observation, RRS is used throughout all experiments.

Since our model adopts discriminator, each training batch is a combined set of query images and gallery videos. In order to make the discriminator converge, we guarantee a query in each set to have at least one positive sample in the gallery of the set. Therefore, when we sample one query image, one gallery video with the same identity is sampled, to form an image-video pair of the same ID.

Furthermore, we test the sampling method used in [7,18] to sample the same identity multiple times in one minibatch. *#samples per person* denotes the number of samples with the same identity in a minibatch. If *#samples per person* is 2 and the size of minibatch is 32, there will be 16 identity query-gallery pairs in each minibatch, and we denote this as *#avgID*=16.

While our I2V framework requires only a single query frame for both training and testing, the existing re-ID benchmarks are originally made up for the V2V setting. However during training, it turns out that it is beneficial to sample multiple query frames and use them as a normalization to accelerate training. We sample T_q queries per training sample using RRS, and average $D(I_q, V)$ over the number of queries per video:

$$logit_{final} = \sum_{t=1}^{T_q} \frac{logit_t}{T_q}, \tag{3}$$

then $logit_{final}$ is passed to compute the discriminator loss (Sec. 3.5). Compared to only using a single randomly sampled query frame from the training set, this strategy further improve the speed and performance.

3.5 Training Objective

The whole framework aims to minimize two training objectives, the discriminator loss and the reciprocal attention triplet loss together:

$$\mathcal{L} = \mathcal{L}_D + \mathcal{L}_R, \tag{4}$$

so the two embedding networks and the READ are jointly trained in an end-to-end manner.

Discriminator Loss. The discriminator learns to classify a given image-video pair as positive if two IDs are same or as negative otherwise. The discriminator loss $\mathcal{L}_D$ is based on the binary cross-entropy loss defined as:

$$\mathcal{L}_D = -\mathbf{E}_{V_i \sim P} \log D(I_q, V_i) - \mathbf{E}_{V_i \sim N} \log (1 - D(I_q, V_i)), \tag{5}$$

where P and N are sets of positive and negative samples from the gallery mini-batch, respectively. We use $\mathcal{L}_D$ to impose the attention and embedding networks to generate features with distinctive differences when observing positive and negative pairs.

Reciprocal Attention Triplet Loss. The READ aims at constructing a manifold to classify the concatenated query-specific features. However, the features are combinations of high dimensional image/video visual information that are complex and not easily linearly separable. Extending the hard-mine triplet loss [9], we devise the Reciprocal Attention Triplet Loss (RATL) to provide extra constraints to ensure better manifold learning before the final classification:

$$\begin{aligned} \mathcal{L}_R = \frac{1}{|G|} \sum_{j=1}^{|G|} \sum_{k=1}^{|Q|} [m + & \max_{p \in P_{q_k}} d(F(I_p, V_j), F(I_{q_k}, V_j)) \\ & - \min_{n \in N_{q_k}} d(F(I_n, V_j), F(I_{q_k}, V_j))]_+, \end{aligned} \tag{6}$$

where P_{q_k} and N_{q_k} are the groups of positive and negative query samples of the query I_{q_k} (and $I_{q_k} \notin P_{q_k}$), m is the margin, $d(\cdot, \cdot)$ denotes Euclidean distance

and $[\cdot]_+$ denotes $\max(0, \cdot)$, and k and j are indices for query images and gallery videos in a minibatch, respectively. Step-by-step explanation about the RATL is detailed in the supplementary material.

Triplet loss is popularly used in identity retrieval tasks [27,18,37,21,16] to promote metric learning. After training, the distance in the shared embedding space becomes the criterion of similarity between two targets. On the contrary, computation of the RATL is not based on a shared embedding space, but a query-specific understanding against each gallery video. We use the RATL to encourage the reciprocal attention block to focus on specific spatio-temporal regions, where useful information is available for discriminating between gallery identities.

4 Experiments

4.1 Benchmark

We evaluate our method on video based person re-identification (re-ID) datasets: DukeMTMC-VideoReID (Duke) [34], and MARS [40]. Video frames from both datasets are cropped by the bounding boxes from a person detector. **Duke** contains 702 identities for training, 702 for testing, and 408 distractors. There are 2,196 videos for training, 2,636 videos for testing; and 6 cameras are used to capture the videos. **MARS** dataset contains 625 identities for training, and 635 identities for testing. Unlike Duke, MARS dataset's distractors do not have respective ID, so there are 625+635+1 (distractor) = 1,261 identities in total. Training split of MARS contains 8,298 tracklets, test split contains 11,310 tracklets (excluding 'junk' images provided in the orginal dataset that do not affect retrieval accuracy), and 6 cameras are used. It is worthwhile to note that query and gallery sets could share same camera views in the test split, however for each query, his/her gallery samples from the same camera are excluded during evaluation.

Following the standard evaluation metrics for both datasets, we report the the cumulative matching curve (CMC) at top-1, top-5, top-10 accuracy and the mean average precision (mAP). Identity retrieval tasks demand high top-1 accuracy compared to general image retrieval tasks since the goal is to precisely identify whom the query is. Yet, the database contains multiple matching answers for each query, so mAP is also used [41] to reflect recall as well as precision. The prediction of the READ is an affinity score (i.e., the probability of matching) between a query and its gallery sample. Hence, the list of ranked gallery samples is sorted in a descending order of the output probabilities instead of their feature distances in the embedding space. During testing, the first frame of each query video is used as the query image following the previous I2V re-ID context [32,7]. As for the test gallery videos, we follow [18] to sample the first frame from T equally-divided chunks which would also ensure consistent evaluation results over repeated tests.

4.2 Methods to be Compared

We analyze the effectiveness of architecture by comparing with two baseline models. One baseline model is designed without reciprocal attention, i.e., only comprising of the image embedding and the non-local video embedding. Since the output of the video embedding network has a time dimension T, we average the feature over its time dimension so the size of the image and video embedding would match. Image and video features are concatenated together then go through two fully connected layers, trained with the discriminator loss and the RATL as in Sec. 3.5. The other baseline is a metric learning architecture which has been tested in [7]. We report their performance in our experiment to show the differences with the discriminator architecture.

In Sec. 3.4, various sampling related concerns are shared. Therefore, we examine how the parameter *#samples per person* in each training minibatch affects performance. However, the *#samples per person* parameter sometimes cannot be directly applied owing to database statistics. For example, Duke has a smaller number of tracklets per identity compared to MARS. To correctly show the difference, we record the average number of identities (*#avgID*) sampled in a minibatch. If there are plenty of tracklets, a minibatch of size 32 ($B = 32$) and 4 *samples per person* ($SP = 4$) will contain 8 identities in the batch. Duke does not have that many tracks, so the *#avgID* becomes 13.3 when $B = 32$ and $SP = 4$, as insufficient tracks are randomly sampled from different identities. To match the *#avgID* with Duke, we give randomness to the number of identities and set *#avgID* around 13.8 in the case of MARS.

4.3 Implementation Detail

We sample 4 frames (i.e. $T = 4$) from each video, and image height and width are resized to 256 and 128 pixels respectively. Adam [12] is adopted to optimize the parameters, with a weight decay of 5e-4 and a starting learning rate of 1e-4. The learning rate is divided by 10 after 60 or 100, and 180 or 200 epochs until it reaches 1e-6. We use a batch size of 32, and the margin of the RATL is $m = 0.3$. In addition, we apply random horizontal flip to the training input images or videos. We report the result of models with the best top-1 accuracy. Scikit-learn [23] version <0.19 is used to calculate mAP, the reason is detailed in the supplementary material.

5 Results

5.1 Ablation Study

Improvement by the READ. Table 1 shows the results of ablation experiments. Beginning from 'baseline (discriminator)', 'READ (*w/o triplet*)' contributes 8.9 and 15.3 top-1 accuracy improvements respectively on Duke and MARS. On top of 'READ (*w/o triplet*)', the RATL adds 1.7-2.2 and 1.2 top-1

Table 1: Results of ablation experiments. Results of baseline (metric) is brought from [7]. Note that three READ experiments in the middle (without triplet loss, or with/without random horizontal flip augmentation) defaults to *#avgID* of 32. The other experiments with specified *#avgID* defaults to use horizontal flip augmentation, even if the performance of not using the augmentation is slightly better, for the sake of readability. See details in Sec. 4.2 and Sec. 5.

Method		DukeMTMC-VideoReID				MARS			
		top-1	top-5	top-10	mAP	top-1	top-5	top-10	mAP
baseline (metric) [7]		67.5	-	-	65.6	67.1	-	-	55.5
baseline (discriminator)		75.2	88.9	92.9	71.7	65.0	81.0	85.7	53.2
READ (*w/o triplet*)		84.1	93.5	95.0	80.9	80.3	90.2	93.1	68.6
READ (*w/o hor. flip*)		85.8	93.0	95.7	82.0	81.5	92.1	93.8	70.4
READ (*w/ hor. flip*)		86.3	94.4	96.2	83.3	81.5	91.2	93.3	69.9
	#avgID								
READ	32	86.3	94.4	96.2	83.3	81.5	91.2	93.3	69.9
	16	86.0	93.7	95.3	83.4	76.6	86.9	89.6	64.6
	13.8	-	-	-	-	77.6	88.2	90.8	65.7
	13.3	84.9	94.3	96.6	82.9	-	-	-	-

accuracy depending on the database, while the horizontal flip augmentation does not seriously impact the results. Details follow about each ablation experiment.

Baseline. The difference between two baselines is the use of metric learning or a discriminator to distinguish identities. Also, the video embedding in 'baseline (discriminator)' does not have $res5$ layer compared to 'baseline (metric)', meaning a lower network capacity. The results show both baselines do not necessarily solve the issue of asymmetry. They both roughly perform pooling across the temporal dimension to match the image embedding dimension, and 'baseline (discriminator)' only concatenates those features. The READ is able to address the asymmetry of two different embeddings without dropping the temporal information.

RATL. The RATL also plays an important role for instructing reciprocal attention module, to learn where to focus on the gallery videos based on a given query. Without the RATL, performance degrades by 2.2 top-1 accuracy in Duke and 1.2 in MARS.

Augmentation. We examine the effect of random horizontal flip. The experimental results show that the random horizontal flip does not have significant influence to the performance. It implies the READ is robust to the direction of pedestrian given unflipped training data. Unless specified, all experimental results are derived from models trained with the horizontal flip augmentation.

#avgID. Applying various average numbers of identities in each minibatch (*#avgID*) displays different trends depending on the dataset. In the case of

Table 2: Benchmark comparison with state-of-the-art I2V re-ID methods.

(a) DukeMTMC-VideoReID.

Method	top-1	top-5	top-10	mAP
TKP [7]	77.9	-	-	75.9
READ (*ours*)	**86.3**	94.4	96.2	**83.4**

(b) MARS.

Method	top-1	top-5	top-10	mAP
P2SNet [32]	55.3	72.9	78.7	-
ResNet-50+XQDA [17]	67.2	81.9	86.1	54.9
TKP [7]	75.6	87.6	90.9	65.1
READ (*ours*)	**81.5**	**92.1**	**93.8**	**70.4**

Duke, the range of disparity between different *#avgID* is small and does not seem significant, since changing *#avgID* from 32 to 16 causes 0.3 drop of top-1 accuracy. MARS yet outputs highly variable results. There is a 3.9 top-1 accuracy gap between *#avgID*=32 and *#avgID*=16. This possibly results from the distribution difference as MARS has larger number of tracklets for each identity. Thus, this parameter should be carefully selected based on the dataset.

5.2 Comparison

Table 2 presents the results of comparison with state-of-the-art I2V re-ID methods on DukeMTMC-VideoReID (Duke) and MARS benchmark datasets. The READ shows a significant improvement over the state-of-the-art methods on both datasets. On Duke dataset, the READ improves top-1 accuracy and mAP by around 8 and 6 respectively compared to TKP [7]. On MARS dataset, the READ outperforms all models from P2SNet [32], ResNet-50+XQDA [17], to TKP [7], by a large margin of at least 5.9 top-1 accuracy and 5.3 mAP.

5.3 Analysis

In this section, we examine the effect of various options on the performance on DukeMTMC-VideoReID (Duke) dataset.

Normalization. Without the query sample normalization (Eq. 3) for training, only a single image is randomly sampled as a query. Table 3(a) shows the effect of the query sample normalization with various *#avgID*. When *#avgID*=32, the top-1 accuracy is dropped by 6. The gap is huge considering the network does see all training query images even without the normalization. However, there are less positive samples without normalization in each minibatch, so it could have also

Table 3: Experimental results of analysis. Experiments in (a)-(d) are performed with DukeMTMC-VideoReID database.

(a) Without normalization.

#avgID	top-1	top-5	top-10	mAP
32	80.3	91.2	94.6	77.8
16	83.6	92.7	95.0	81.0
13.3	83.0	93.3	95.4	81.1

(b) Peformance on variable T.

Metric	$T = 2$	$T = 4$	$T = 6$	$T = 8$
top-1	83.2	86.3	85.2	85.9
top-5	92.6	94.4	94.4	94.4
mAP	79.0	83.4	82.2	82.7

(c) #samples for RATL.

norm.	*#avgID*	$\|P_q\|$	$\|N_q\|$
yes	32	4	124
yes	16	8	120
no	32	1	31
no	16	2	30

(d) Direction of RATL.

Direction	top-1	top-5	top-10	mAP
query $\rightarrow$ *gallery*	83.0	93.0	95.0	79.9
gallery $\rightarrow$ *query*	86.3	94.4	96.2	83.3

negatively impacted the RATL. Thus, the experiments with smaller *#avgID* of 16 and 13.3 and without normalization are additionally conducted. Those tests improved top-1 accruacy to 83.6 and 83.0 from 80.3; and mAP from 77.8 to 81.0 and 81.1. Combined with the results in Table 1, we can conclude that the query sample normalization greatly helps the discriminator loss $\mathcal{L}_D$, and the RATL $\mathcal{L}_R$ requires balance between the number of positive and negative samples, $|P_q|$ and $|N_q|$, respectively. Table 3(c) shows $|P_q|$ and $|N_q|$ in the perspective of the RATL.

Sample Length. Table 3(b) shows the experiments conducted by varying the length T of samples from the gallery videos. We test four variants of T from 2 to 8 with a stride of 2, and the results show the performances do not increase beyond $T = 4$. This result is consistent with [7]. In contrast, it is different to non-local neural networks applied to action recognition [33], where longer input video clips coherently shows better performance. This inconsistency is possibly on grounds of differences between action recognition and re-identification. In our re-ID benchmarks, gallery videos are guaranteed to contain an identity's visual information from a dedicated camera viewpoint. Hence, $T = 4$ can be the point where additional information does not further contribute.

Direction of RATL. A single operation of the RATL is described to be operated within the pool Q of query samples, in the perspective of a gallery sample $V_j \in G$ (Eq. 6 of Sec. 3.3). We analyze another case of direction, where the RATL operated in the pool of gallery samples with a query sample as its basis. The results are presented in Table 3(d). Compared to our default setting, the top-1 accuracy drops by 3.3, and mAP is degraded by 3.5. This result can be interpreted in the similar context of sample distribution. The asymmetry of I2V

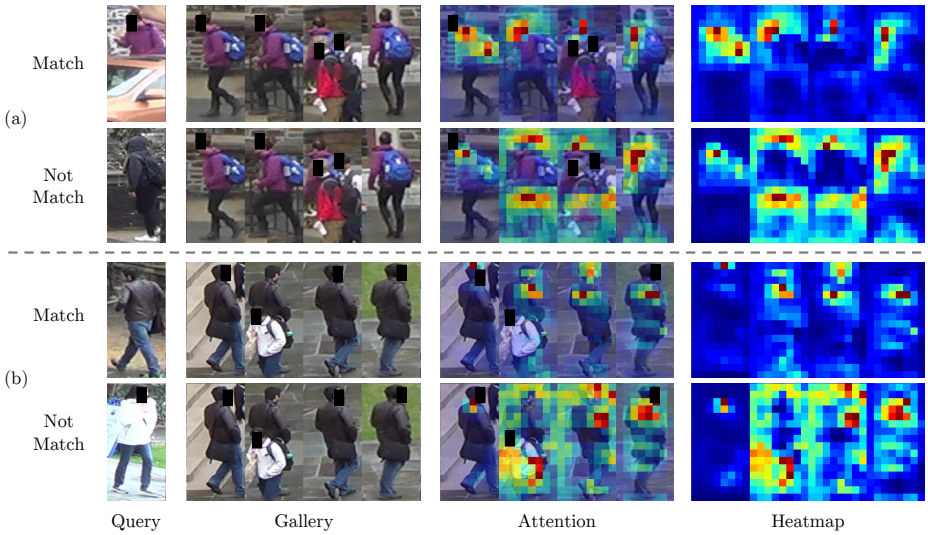

Fig. 3: Visualization of attention. Softmax normalized attention maps in accordance with the image-video pair are visualized. The attention focuses on the different spatio-temporal region of the gallery depending on the cases when the pair has same IDs (Match) and different IDs (Not Match). The detailed analysis can be found in the main text. Best viewed in color.

allows abundant sampling of query images compared to the amount of gallery video samples, thus *#avgID* parameter and normalization is exploited in our experiments. When a gallery sample is the basis of the RATL, there are 128 query images in our default setting. On the other side, there are only 32 gallery videos when a query sample is the basis. So the *gallery* $\rightarrow$ *query* direction allows each gallery sample to observe more diverse counterpart query samples.

5.4 Visualization

We qualitatively evaluate our proposed method by visualizing the attention map on Duke database as in Fig. 3. Attention created by the READ occurs across global spatio-temporal dimension. So as to visualize the attention as 2D map, the softmaxed attention of dimension $HW \times THW$ is aggregated by averaging over HW, then is reshaped to $T \times H \times W$.

To analyze the effect of reciprocal attention with different query images, we compare the attention map generated by a matched image and a non-matched image given the same gallery video. The attention focuses on the target when the identity of image-video pair matches. On the other hand, attention often spotlights other person or background if the pair has different IDs. For instance in Fig. 3(a), the attention focuses on the upper body of the matching target, whereas different people and backgrounds are attended when IDs do not match.

Table 4: Computational cost for forwarding the image/video embedding networks, and the reciprocal attention discriminator. Four TESLA P40 GPUs are used with the batch size of 32.

	N	M	N Image Embedding	M Video Embedding	N*M READ
-	32	32	49ms	168ms	4ms
Duke	702	2,636	1s	14s	12s
MARS	1,980	11,310	3s	59s	140s

Similarly, in Fig. 3(b), the target is attended with an image-video pair of the same ID. On the contrary, in case the query with white jacket is chosen, a person with a white jacket in the second frame of the gallery is focused. After all, the READ tries to find the information that matches the query if unexpectedly different gallery video is given. These results show that the proposed attention mechanism operates in a way of searching query related information from the video.

5.5 Computational Cost

We provide additional network forwarding time analysis in Table 4. Similar to the existing re-ID pipelines, our method is able to prefetch image/video embeddings in N+M forward passes. The remaining cost is O(NM) forward passes of the READ, which is a feasible overhead even in MARS and the cost similarly exists in other discriminator based re-ID works [36]. Also in the light of the READ, the expensive video embedding can be fully replaced by an image embedding network (x2.2 FLOPs smaller) with a marginal performance drop as discussed in the supplementary material.

Besides retrieving from a massive database (e.g. image search engine), re-ID aims at matching subjects across multiple camera views where occlusion and visual degradation might occur. The underlying motivation of this paper is an application on real-world MOT tasks, e.g. tracking person in dance videos [10], similarly described in the supplementary material. In such scenario, the scale of N*M stays feasible for real-time speed.

6 Conclusion

In this paper, we propose the Reciprocal Attention Discriminator (READ), the novel attention-based discriminator framework for I2V re-ID task along with two losses, the discriminator loss and the Reciprocal Attention Triplet Loss (RATL), for training the model. The READ can successfully integrate asymmetric information of image-video pair using non-local operation. We validate the effectiveness of our method quantitatively and qualitatively on two widely-used databases. Our method surpasses other previous arts by a wide margin. We also reported extensive ablation studies to verify the design choices.

References

1. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR (2017)
2. Chen, B., Deng, W., Hu, J.: Mixed high-order attention network for person re-identification. In: ICCV (2019)
3. Chen, W., Chen, X., Zhang, J., Huang, K.: Beyond triplet loss: a deep quadruplet network for person re-identification. In: CVPR (2017)
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: CVPR (2009)
5. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: ICCV (2019)
6. Fu, Y., Wang, X., Wei, Y., Huang, T.: Sta: Spatial-temporal attention for large-scale video-based person re-identification. In: AAAI (2019)
7. Gu, X., Ma, B., Chang, H., Shan, S., Chen, X.: Temporal knowledge propagation for image-to-video person re-identification. In: ICCV (2019)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
9. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint (2017)
10. Ho, H.I., Shim, M., Wee, D.: Learning from dances: Pose-invariant re-identification for multi-person tracking. In: ICASSP (2020)
11. Hou, R., Ma, B., Chang, H., Gu, X., Shan, S., Chen, X.: Interaction-and-aggregation network for person re-identification. In: CVPR (2019)
12. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)
13. Li, D., Chen, X., Zhang, Z., Huang, K.: Learning deep context-aware features over body and latent parts for person re-identification. In: CVPR (2017)
14. Li, S., Bak, S., Carr, P., Wang, X.: Diversity regularized spatiotemporal attention for video-based person re-identification. In: CVPR (2018)
15. Li, W., Zhu, X., Gong, S.: Harmonious attention network for person re-identification. In: CVPR (2018)
16. Li, Y.J., Chen, Y.C., Lin, Y.Y., Du, X., Wang, Y.C.F.: Recover and identify: A generative dual model for cross-resolution person re-identification. In: ICCV (2019)
17. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: CVPR (2015)
18. Liu, C.T., Wu, C.W., Wang, Y.C.F., Chien, S.Y.: Spatially and temporally efficient non-local attention network for video-based person re-identification. In: BMVC (2019)
19. Liu, D., Wen, B., Fan, Y., Loy, C.C., Huang, T.S.: Non-local recurrent network for image restoration. In: NeurIPS (2018)
20. Liu, X., Zhao, H., Tian, M., Sheng, L., Shao, J., Yi, S., Yan, J., Wang, X.: Hydraplus-net: Attentive deep features for pedestrian analysis. In: ICCV (2017)
21. Liu, Y., Yuan, Z., Zhou, W., Li, H.: Spatial and temporal mutual promotion for video-based person re-identification. In: AAAI (2019)
22. Oh, S.W., Lee, J.Y., Xu, N., Kim, S.J.: Video object segmentation using space-time memory networks. In: ICCV (2019)
23. 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)

24. Shen, Y., Li, H., Yi, S., Chen, D., Wang, X.: Person re-identification with deep similarity-guided graph neural network. In: ECCV (2018)
25. Su, C., Li, J., Zhang, S., Xing, J., Gao, W., Tian, Q.: Pose-driven deep convolutional model for person re-identification. In: ICCV (2017)
26. Suh, Y., Wang, J., Tang, S., Mei, T., Mu Lee, K.: Part-aligned bilinear representations for person re-identification. In: ECCV (2018)
27. Sun, Y., Xu, Q., Li, Y., Zhang, C., Li, Y., Wang, S., Sun, J.: Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. In: CVPR (2019)
28. Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S.: Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: ECCV (2018)
29. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: CVPR (2015)
30. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: ICCV (2015)
31. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NIPS (2017)
32. Wang, G., Lai, J., Xie, X.: P2SNet: Can an image match a video for person re-identification in an end-to-end way? TCSVT **28**(10), 2777–2787 (2017)
33. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: CVPR (2018)
34. Wu, Y., Lin, Y., Dong, X., Yan, Y., Ouyang, W., Yang, Y.: Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In: CVPR (2018)
35. Zhang, D., Wu, W., Cheng, H., Zhang, R., Dong, Z., Cai, Z.: Image-to-video person re-identification with temporally memorized similarity learning. TCSVT **28**(10), 2622–2632 (2017)
36. Zhang, Y., Li, X., Zhang, Z.: Learning a key-value memory co-attention matching network for person re-identification. In: AAAI (2019)
37. Zhang, Y., Zhong, Q., Ma, L., Xie, D., Pu, S.: Learning incremental triplet margin for person re-identification. In: AAAI (2019)
38. Zhao, H., Tian, M., Sun, S., Shao, J., Yan, J., Yi, S., Wang, X., Tang, X.: Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In: CVPR (2017)
39. Zhao, L., Li, X., Zhuang, Y., Wang, J.: Deeply-learned part-aligned representations for person re-identification. In: ICCV (2017)
40. Zheng, L., Bie, Z., Sun, Y., Wang, J., Su, C., Wang, S., Tian, Q.: Mars: A video benchmark for large-scale person re-identification. In: ECCV (2016)
41. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: ICCV (2015)
42. Zheng, L., Yang, Y., Hauptmann, A.G.: Person re-identification: Past, present and future. arXiv preprint (2016)
43. Zhu, X., Jing, X.Y., Wu, F., Wang, Y., Zuo, W., Zheng, W.S.: Learning heterogeneous dictionary pair with feature projection matrix for pedestrian video retrieval via single query image. In: AAAI (2017)
44. Zhu, X., Jing, X.Y., You, X., Zuo, W., Shan, S., Zheng, W.S.: Image to video person re-identification by learning heterogeneous dictionary pair with feature projection matrix. TIFS **13**(3), 717–732 (2017)