IMAGE-TO-VIDEO RE-IDENTIFICATION VIA MUTUAL DISCRIMINATIVE KNOWLEDGE TRANSFER

Pichao Wang, Fan Wang, Hao Li

Alibaba Group

ABSTRACT

The gap in representations between image and video makes Image-to-Video Re-identification (I2V Re-ID) challenging, and recent works formulate this problem as a knowledge distillation (KD) process. In this paper, we propose a mutual discriminative knowledge distillation framework to transfer a video-based richer representation to an image based representation more effectively. Specifically, we propose the triplet contrast loss (TCL), a novel loss designed for KD. During the KD process, the TCL loss transfers the local structure, exploits the higher order information, and mitigates the misalignment of the heterogeneous output of teacher and student networks. Compared with other losses for KD, the proposed TCL loss selectively transfers the local discriminative features from teacher to student, making it effective in the ReID. Besides the TCL loss, we adopt mutual learning to regularize both the teacher and student networks training. Extensive experiments demonstrate the effectiveness of our method on the MARS, DukeMTMC-VideoReID and VeRi-776 benchmarks.

Index Terms— Image-to-Video, Re-identification, knowledge distillation, triplet contrast loss, mutual learning

1. INTRODUCTION

Re-identification (Re-ID) targets to retrieve an object among the gallery set that has the same identity with the given query across non-overlapping cameras. Generally speaking, this task involves three main categories: image-based (I2I), video-based (V2V) and Image-to-Video (I2V) Re-ID. Contrary to I2I and V2V Re-ID, I2V Re-ID targets to build bridges between image representations and video representations. Recent works [1, 2] formulate this problem as a knowledge distillation process. Specifically, Gu et al. [1] transfer temporal information from a video-based teacher network to an image-based student one; Porrello et al. [2] distill the multiple views knowledge from teacher to student network. In terms of loss function for distillation, in TKP [1], feature distance and cross sample distance are used for propagation; in VKD [2], logits and pairwise distance are adopted for distillation. All of these distillation losses focus on the global matching, designed for classification tasks [3] or representative learning [4]. However, in the Re-ID task, the discriminative ability of features
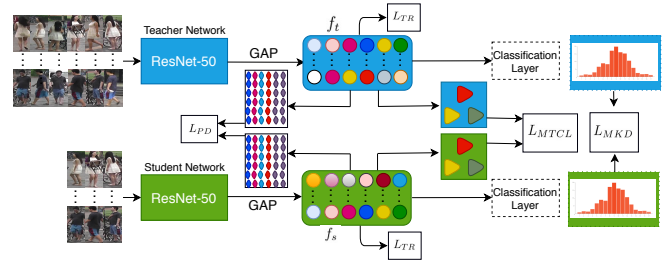

Fig. 1. The framework of MDKT.

is more important due to the non-overlapping ID labels of the training and testing set.

In this paper, inspired by the commonly used triplet loss [5] for ReID task, we design the triplet contrast loss (TCL) for the knowledge distillation between teacher network and student network. Instead of directly comparing the distance between anchor-positive and anchor-negative pairs, we propose to measure the probabilities of these distances. Unlike other losses [1, 2] that match the global information between teacher and student, the TCL loss focuses on the local structure consisting of nearest neighbors for each example. Coupled with hard triplet sampling in the student embedding, the discriminative ability of features are enhanced by the soft supervision of the teacher network. The TCL loss mitigates the misalignment of the heterogeneous outputs of teacher and student networks. It also encodes the higher order (third order) structured knowledge in the triplet, and introduces richer information of similarity comparison than the vanilla triplets in the context of knowledge distillation. The TCL loss is complementary to conventional KD losses and can be combined with them to further boost the performance. The TCL improves the performance of I2V Re-ID largely, even without the commonly used cross-entropy loss. Besides the proposed TCL, mutual learning [6] is adopted in order to regularize both teacher and student learning. Integrating these two techniques, the proposed method is dubbed as mutual discriminative knowledge transfer (MDKT). Extensive experiments demonstrate the effectiveness of the proposed method on MARS [7], DukeMTMC-VideoReID [8] and VeRi-776 benchmarks [9].

2. THE PROPOSED METHOD

The framework of MDKT consists of two-stages: 1) the teacher network is trained using the standard V2V Re-ID setting. 2) we feed frames representing different numbers of views as input to the teacher and students networks for view KD using three level distillation losses, as depicted in Figure 1.

2.1. Teacher Network

Without loss of generality, ResNet-50 [10] is used as the backbone network. Following [1, 11], the network is initialized with the weights pretrained on ImageNet. A few amendments as in [11] are included in the architecture.

Video Representation. Given P person video clips $V = \{v_n\}_{n=1}^P$, each v_n contains T frames (T is set to 8 unless otherwise specified). In this paper, the video representation $f_t(V_n)$ for the teacher network is simply computed using the temporal average pooling over all frame features.

Teacher Network Optimisation. The teacher network is trained using the cross-entropy classification loss and the triplet loss [5]. The cross-entropy loss L_{CE}^t is formulated as:

$$L_{CE}^t = -\boldsymbol{y}_t \log \hat{\boldsymbol{y}}_t \qquad (1)$$

where t represents the teacher model, $\boldsymbol{y}_t$ and $\hat{\boldsymbol{y}}_t$ represent the one-hot labels and the output of the softmax respectively. The triplet loss L_{TR}^t is normally trained on a series of triplets $\{x_a, x_p, x_n\}$, where x_a and x_p are videos of the same person, denoted as anchor and positive samples, and x_n is from a different person, denoted as the negative sample. The triplet loss is designed to keep x_a closer to x_p than x_n and it is defined as:

$$L_{TR}^t = \sum_{a,p,n}^N [\|f_t(x_a) - f_t(x_p)\|_2^2 - \|f_t(x_a) - f_t(x_n)\|_2^2 + \alpha]_+ \qquad (2)$$

where $[z]_+ = max(z, 0)$, and $f_t(x_a), f_t(x_p), f_t(x_n)$ represent features of the three videos from the teacher network. α is a margin that is enforced between positive and negative pairs.

2.2. Mutual Discriminative Knowledge Transfer

After training the teacher network, following VKD [2], we transfer the knowledge lying in multiple views in a teacher-student distillation fashion. We allow the teacher to access frames $I_t = (i_1, i_2, \cdots, i_N)$ from different viewpoints and feed the student with subset of teacher's inputs $I_s = (i_1, i_2, \cdots, i_M)$, where the cardinality N >M (in the experiments, N = 8 and M = 2). The MDKT is formulated as an optimization problem using three level distillation losses:

Mutual Logits Distillation. Similar to VKD [2], the logits based knowledge distillation loss [3] is adopted to distill the multiple view knowledge from teacher to student:

$$L_{KD_{t2s}} = \tau_1^2 KL(y_t \| y_s) \qquad (3)$$

where s represents the student model; KL denotes the Kullback-Leibler divergence; $y_t = softmax(f_t(x)/\tau_1)$ and $y_s = softmax(f_s(x)/\tau_1)$ are the output distributions of teacher and student networks; τ_1 is a non-negative temperature, the large the value of τ_1, the smoother the output. $L_{KD_{t2s}}$ is a first-order distillation loss and it makes the student keep its predictions consistent with the teacher. Unlike VKD [2], another loss from student to teacher for logits distillation is added as follows:

$$L_{KD_{s2t}} = \tau_1^2 KL(y_s \| y_t) \qquad (4)$$

The final mutual logits distillation loss is:

$$L_{MKD} = L_{KD_{t2s}} + L_{KD_{s2t}} \qquad (5)$$

Pairwise Distance in Embedding. In addition to logits distillation, we also adopt an embedding-based knowledge distillation loss L_{PD} [2]:

$$L_{PD} = \sum_{(i,j) \in \binom{B}{2}} (D_t[i,j] - D_s[i,j])^2 \qquad (6)$$

where $D_t[i,j] = D(f_t(I_t[i]), f_t(I_t[j]))$, indicating the distance induced by the teacher between the i-th and j-th inputs (the same notation $D_s[i,j]$ also hold for the student); B denotes the batch size. The L_{PD} loss is a second order loss which encourages the student to mirror the pairwise distances spanned by the teacher.

Triplet Contrast Loss for Discriminative Transfer. The two abovementioned distillation losses mainly target to address the representation learning (global matching), but neglect the transfer of discrimination ability (local structure). For Re-ID tasks, the disrcriminative feature learning is more important as the labels between the training set and testing set are different. In order to address this in the context of I2V Re-ID, we propose a third order distillation loss, the triplet contrast loss (TCL), inspired by the vanilla triplet loss.

In vanilla triplet loss (Eq. 2), the distances between anchor-positive and anchor-negative pairs are used for discriminative feature learning. However, in knowledge distillation, the outputs of two networks are heterogeneous, the absolute distances between these pairs are not well aligned. In order to mitigate the misalignment, we propose to measure the probability of the two distances. Let $d_{a2p}^t = \|f_t(x_a) - f_t(x_p)\|_2^2$ and $d_{a2n}^t = \|f_t(x_a) - f_t(x_n)\|_2^2$, and the probability is defined as:

$$p_{apn_{\tau_2}} = \frac{\exp(-d_{a2p}^t/\tau_2)}{\exp(-d_{a2p}^t/\tau_2) + \exp(-d_{a2n}^t/\tau_2)} \qquad (7)$$

p_{apn} measures how much the anchor is closer to the positive than the negative. If the anchor is closer to the positive

Models	Losses				MARS				Duke-video			
	L_{TR}	L_{MKD}	L_{PD}	L_{MTCL}	I2V		V2V		I2V		V2V	
					cmc1	mAP	cmc1	mAP	cmc1	mAP	cmc1	mAP
TR	√	×	×	×	76.77	66.85	84.55	74.23	78.24	70.66	88.24	84.96
TCL	×	×	×	√	80.71	71.56	86.82	78.04	82.69	79.26	93.38	92.01
TR+TCL	√	×	×	√	81.16	72.91	86.36	78.68	83.65	80.32	95.01	93.22
KD+PD+TCL	×	√	√	√	84.70	77.56	89.19	82.53	86.32	84.57	**95.58**	93.94
TR+KD+PD	√	√	√	×	83.96	77.43	88.89	82.47	85.04	83.97	95.01	93.69
TR+KD+TCL	√	√	×	√	83.59	76.28	88.69	81.83	84.90	83.89	94.87	93.56
TR+PD+TCL	√	×	√	√	85.33	77.90	89.01	82.65	84.90	83.74	95.30	**93.94**
ALL	√	√	√	√	**85.65**	**78.02**	**89.48**	**82.90**	**86.78**	**84.82**	95.26	93.83

Table 1. Ablation study on the impact of loss terms on MARS and Duke-video datasets using ResNet-50.

Models	MARS			
	I2V		V2V	
	cmc1	mAP	cmc1	mAP
freeze teacher	85.10	77.65	89.44	82.79
without mutual	85.33	77.77	89.22	82.80
with mutual	**85.65**	**78.02**	**89.48**	**82.90**
	Duke-video			
	I2V		V2V	
	cmc1	mAP	cmc1	mAP
freeze teacher	86.65	84.58	95.09	93.70
without mutual	86.63	84.72	95.10	93.51
with mutual	**86.78**	**84.82**	**95.26**	**93.83**

Table 2. Ablation study on the impact of mutual learning.

than the negative, p_{apn} is large, otherwise it is small. Compared with the vanilla triplet which accumulates the absolute differences, the p_{apn} brings higher order (third order) similarity comparison, encoding the structural information in the triplet. To transfer the local structure, batch hard sample mining [5] is adopted to make the local structure consist of nearest neighbors for each example.

With the $p^t_{apn_{\tau_2}}$ from teacher and $p^s_{apn_{\tau_2}}$ from student, we define the distribution $P^t_{apn_{\tau_2}} = [p^t_{apn_{\tau_2}}, 1 - p^t_{apn_{\tau_2}}]$ and $P^s_{apn_{\tau_2}} = [p^s_{apn_{\tau_2}}, 1 - p^s_{apn_{\tau_2}}]$. Thus, the TCL loss between teacher and student is formulated as:

$$L_{TCL_{t2s}} = \sum_{a,p,n}^{N} KL(P^t_{apn_{\tau_2}} \| P^s_{apn_{\tau_2}}) \quad (8)$$

Similar to the mutual logits distillation, we also adopt the matching between student and teacher for TCL loss:

$$L_{TCL_{s2t}} = \sum_{a,p,n}^{N} KL(P^s_{apn_{\tau_2}} \| P^t_{apn_{\tau_2}}) \quad (9)$$

Our final mutual TCL loss for final optimization is defined as:

$$L_{MTCL} = L_{TCL_{t2s}} + L_{TCL_{s2t}} \quad (10)$$

2.3. The Objective Function

Besides the three level distillation losses, we also adopt the widely used triplet loss for both teacher and student networks, formulated as:

$$L_{TR} = L^t_{TR} + L^s_{TR} \quad (11)$$

Different from TKP [1] and VKD [2], we discard the cross-entropy classification loss, as we find that our proposed MDKT can better learn discriminative features even without the classification loss. The final objective function is formulated as the combination of the four losses:

$$L = L_{TR} + \alpha L_{MKD} + \beta L_{PD} + \gamma L_{MTCL} \quad (12)$$

2.4. I2V Re-ID Testing

In the test stage, each query is a static image and the gallery set consists of video clips. In our work, one single network is used for both image and video feature extraction. Herein, we treat the still image as a video. After feature extraction, the distances between the query feature and each gallery video features are calculated and the I2V retrieval is conducted according to the distances.

3. EXPERIMENTS

3.1. Datasets and Implementation Details

We evaluate our method on MARS [7], DukeMTMC-VideoReID (Duke-video) [8] and VeRi-776 (VeRi) [9] benchmarks. The same training hyper-parameters in VKD [2] are adopted in this paper. We keep the $\tau_1 = 10$ (Equation 3), $\tau_2 = 4$ (Equation 7), $\alpha = 10^{-1}$ (Equation 12), $\beta = 10^{-4}$ (Equation 12) and $\gamma = 1000$ (Equation 12) in all experiments. The teacher is set to training mode during distillation.

3.2. Impact of Loss Terms

We perform a thorough ablation study on the final loss terms (Equation 12) on MARS and Duke-video datasets, and the

Models	MARS			
	I2V		V2V	
	cmc1	mAP	cmc1	mAP
with CE loss	84.29	76.94	88.69	82.23
without CE loss	**85.65**	**78.02**	**89.48**	**82.90**
	Duke-video			
	I2V		V2V	
	cmc1	mAP	cmc1	mAP
with CE loss	85.83	84.66	94.07	93.44
without CE loss	**86.78**	**84.82**	**95.26**	**93.83**

Table 3. Ablation study on the impact of cross-entropy loss.

Method	top1	top5	mAP
P2SNet [12]	55.3	72.9	-
TMSL [13]	56.5	70.6	-
TKP [1]	75.6	87.6	65.1
STE-NVAN [14]	80.3	-	68.8
NVAN [14]	80.1	-	70.2
MGAT [15]	81.1	92.2	71.8
READ [16]	81.5	92.1	70.4
VKD [2]	83.9	93.2	77.3
MDKT	**85.7**	**93.3**	**78.0**

Table 4. Comparison with SOTA methods on MARS dataset.

results are shown in Table 1. It can be seen that results are largely degraded only using triplet loss (TR model) for training. The reason is that with fewer views input, the batch hard sample mining could not pick up rich triplets for training. Compared with the TR model, the TCL model shows greatly improved performance by using only the mutual triplet contrast loss. The three level distillation losses play a crucial role for performance gain (KD+PD+TCL model). The distillation losses coupled with triplet loss achieves the best results for I2V Re-ID setting.

3.3. Impact of Mutual Learning

In our proposed MDKT, we adopt mutual learning to regularize both the teacher and student training. In this section, we investigate the role of mutual learning. Three models are designed, namely, freezing the teacher parameters, without mutual learning in Equation 5 and 10 (without student to teacher loss term), and with mutual learning. The results are illustrated in Table 2. From this Table it can be seen that mutual learning can improve the performances both for I2V and V2V settings.

3.4. Impact of Cross-Entropy Loss

Unlike TKP [1] and VKD [2], in our final objective function (Equation 12), we do not adopt the commonly used cross-

Method	top1	top5	mAP
STE-NVAN [14]	42.2	-	41.3
TKP [1]	77.9	-	75.9
NVAN [14]	78.4	-	76.7
VKD [2]	85.6	93.9	83.8
READ [16]	86.3	94.4	83.4
MDKT	**86.8**	**94.9**	**84.8**

Table 5. Comparison with SOTA methods on Duke-video dataset.

Method	top1	top5	mAP
PROVID [17]	76.8	91.4	48.5
VFL-LSTM [18]	88.0	94.6	59.2
RAM [19]	88.6	-	61.5
VANet [20]	89.8	96.0	66.3
PAMTRI [21]	92.9	92.9	71.9
SAN [22]	93.3	97.1	72.5
VKD [2]	95.2	98.0	82.2
MDKT	**96.0**	**99.3**	**83.4**

Table 6. Comparison with SOTA methods on VeRi dataset.

entropy classification loss. We find that by introducing the TCL loss, coupled with other three loss terms, the network can learn discriminative features for Re-ID. After adding the cross-entropy classification loss, the final performance will be negatively affected. The results are shown in Table 3.

3.5. Comparison with State-of-the-art Methods

The proposed approach is compared with state-of-the-art I2V Re-ID methods on the MARS, Duke-video and VeRi datasets. The results are presented in Table 4, Table 5 and Table 6, respectively. From these three Tables, we can see that our method clearly outperforms other competitors, almost on all metrics including top-1, top-5, and mAP. Specifically, in terms of top-1 and mAP, our method outperforms by 1.8% and 0.7% on MARS, and 0.5% and 1.4% on Duke-video. On the VeRi dataset, our method gains 1.2% boost on mAP.

4. CONCLUSION

In this paper, we propose a mutual discriminative knowledge transfer method for I2V ReID. The proposed method takes advantage of triplet for local discriminative feature learning and aligns the heterogeneous outputs of teacher and student networks. Coupled with the mutual learning, the proposed method achieves state-of-the-art results on three datasets, covering person and vehicle re-identification.

5. REFERENCES

[1] Xinqian Gu, Bingpeng Ma, Hong Chang, Shiguang Shan, and Xilin Chen, "Temporal knowledge propagation for image-to-video person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2019.

[2] Angelo Porrello, Luca Bergamini, and Simone Calderara, "Robust re-identification by multiple views knowledge distillation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020.

[3] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[4] Frederick Tung and Greg Mori, "Similarity-preserving knowledge distillation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2019.

[5] Alexander Hermans, Lucas Beyer, and Bastian Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[6] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu, "Deep mutual learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018.

[7] 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 *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016.

[8] Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang, "Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018.

[9] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma, "A deep learning-based approach to progressive vehicle re-identification for urban surveillance," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016.

[11] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)W*, 2019.

[12] Guangcong Wang, Jianhuang Lai, and Xiaohua Xie, "P2snet: Can an image match a video for person re-identification in an end-to-end way?," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 2777–2787, 2017.

[13] Dongyu Zhang, Wenxi Wu, Hui Cheng, Ruimao Zhang, Zhenjiang Dong, and Zhaoquan Cai, "Image-to-video person re-identification with temporally memorized similarity learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, 2017.

[14] Chih-Ting Liu, Chih-Wei Wu, Yu-Chiang Frank Wang, and Shao-Yi Chien, "Spatially and temporally efficient non-local attention network for video-based person re-identification," *arXiv preprint arXiv:1908.01683*, 2019.

[15] Liqiang Bao, Bingpeng Ma, Hong Chang, and Xilin Chen, "Masked graph attention network for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[16] Minho Shim, Hsuan-I Ho, Jinhyung Kim, and Dongyoon Wee, "Read: Reciprocal attention discriminator for image-to-video re-identification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020.

[17] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma, "Provid: Progressive and multimodal vehicle reidentification for large-scale urban surveillance," *IEEE Transactions on Multimedia*, vol. 20, no. 3, pp. 645–658, 2017.

[18] Saghir Ahmed Saghir Alfasly, Yongjian Hu, Tiancai Liang, Xiaofeng Jin, Qingli Zhao, and Beibei Liu, "Variational representation learning for vehicle re-identificati," in *IEEE International Conference on Image Processing*, 2019.

[19] Xiaobin Liu, Shiliang Zhang, Qingming Huang, and Wen Gao, "Ram: a region-aware deep model for vehicle re-identification," in *IEEE International Conference on Multimedia and Expo*, 2018, pp. 1–6.

[20] Ruihang Chu, Yifan Sun, Yadong Li, Zheng Liu, Chi Zhang, and Yichen Wei, "Vehicle re-identification with viewpoint-aware metric learning," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV))*, 2019.

[21] Zheng Tang, Milind Naphade, Stan Birchfield, Jonathan Tremblay, William Hodge, Ratnesh Kumar, Shuo Wang, and Xiaodong Yang, "Pamtri: Pose-aware multi-task learning for vehicle re-identification using highly randomized synthetic data," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2019.

[22] Jingjing Qian, Wei Jiang, Hao Luo, and Hongyan Yu, "Stripe-based and attribute-aware network: A two-branch deep model for vehicle re-identification," *Measurement Science and Technology*, 2020.