Robust Re-Identification by Multiple Views Knowledge Distillation

Angelo Porrello[0000−0002−9022−8484], Luca Bergamini[0000−0003−1221−8640], Simone Calderara[0000−0001−9056−1538]

AImageLab, University of Modena and Reggio Emilia
{angelo.porrello, luca.bergamini24, simone.calderara}@unimore.it

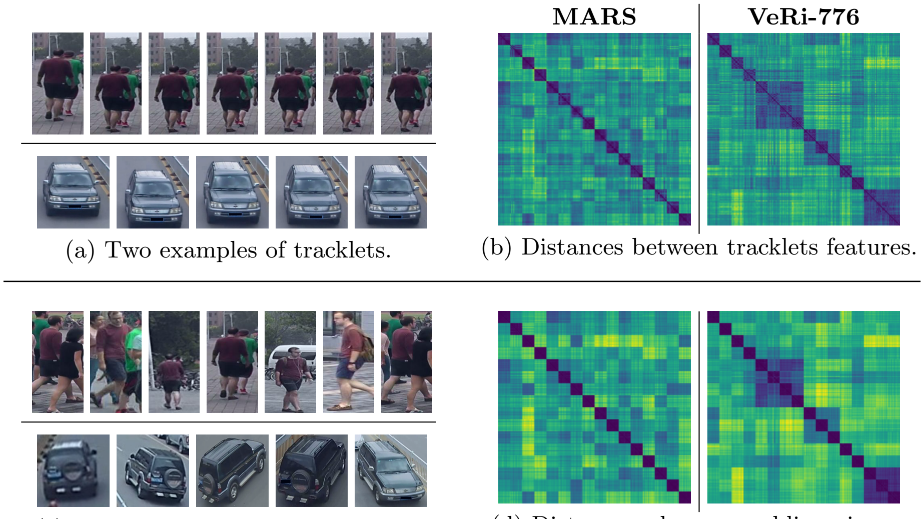
Abstract. To achieve robustness in Re-Identification, standard methods leverage tracking information in a Video-To-Video fashion. However, these solutions face a large drop in performance for single image queries (e.g., Image-To-Video setting). Recent works address this severe degradation by transferring *temporal information* from a Video-based network to an Image-based one. In this work, we devise a training strategy that allows the transfer of a superior knowledge, arising from a set of views depicting the target object. Our proposal – Views Knowledge Distillation (VKD) – pins this *visual variety* as a supervision signal within a teacher-student framework, where the teacher educates a student who observes fewer views. As a result, the student outperforms not only its teacher but also the current state-of-the-art in Image-To-Video by a wide margin (6.3% mAP on MARS, 8.6% on Duke and 5% on VeRi-776). A thorough analysis – on Person, Vehicle and Animal Re-ID – investigates the properties of VKD from a qualitatively and quantitatively perspective. Code is available at https://github.com/aimagelab/VKD.

Keywords: Deep Learning, Re-Identification, Knowledge Distillation

1 Introduction

Recent advances on Metric Learning [38,41,47,45] give to researchers the foundation for computing suitable distance metrics between data points. In this context, Re-Identification (Re-ID) has greatly benefited in diverse domains [56,16,37], as the common paradigm requires distance measures exhibiting robustness to variations in background clutters, as well as different viewpoints. To meet these criteria, various deep learning based approaches leverage videos to provide detailed descriptions for both query and gallery items. However, such a setting – known as Video-To-Video (V2V) Re-ID – does not represent a viable option in many scenarios (e.g. surveillance) [54,50,30,10], where the query comprises a single image (Image-To-Video, I2V).

As observed in [10], a large gap in Re-ID performance still subsists between V2V and I2V, highlighting the number of query images as a critical factor in achieving good results. Contrarily, we advise the learnt representation should not be heavily affected when few images are shown to the network (*e.g.* only



(a) Two examples of tracklets.

(b) Distances between tracklets features.

(c) Two examples of multiview sets.

(d) Distances when ensambling views.

Fig. 1. Visual comparison between tracklets and viewpoints variety, on person (MARS [55]) and vehicle (VeRi-776 [25]) re-id. Right: pairwise distances computed on top of features from ResNet-50. Inputs batches comprise 192 sets from 16 different identities, grouped by ground truth identity along each axis.

one). To bridge such a gap, [10,5] propose a teacher-student paradigm, in which the student – in contrast with the teacher – has access to a small fraction of the frames in the video. Since the student is educated to mimic the output space of its teacher, it will show higher generalisation properties than its teacher when a single frame is available. It is noted that these approaches rely on transferring *temporal* information: as datasets often come with tracking annotation, they can guide the transfer from a tracklet into one of its frames. In this respect, we argue the limits of transferring temporal information: in fact, it is reasonable to assume an high correlation between frames from the same tracklet (Fig. 1a), which may potentially underexploit the transfer. Moreover, limiting the analysis to the temporal domain does not guarantee robustness to variation in background appearances.

Here, we make a step forward and consider which information to transfer, shifting the paradigm from *time* to *views*: we argue that more valuable information arises when ensembling diverse views of the same target (Fig. 1c). This information often comes for free, as various datasets [55,49,25,4] provide images capturing the same target from different camera viewpoints. To support our claim, Fig. 1 (right) reports pairwise distances computed on top of ResNet-50, when trained on Person and Vehicle Re-ID. In more details: matrices from Fig. 1b visualise the distances when tracklets are provided as input, whereas Fig. 1d shows the same for sets of views. As one can see, leveraging different views leads to a more distinctive blockwise pattern: namely, activations from the same identity are more consistent if compared to the ones computed in the tracklet scenario. As shown in [44], this reflects a higher capacity to capture the semantics of the dataset, and therefore a *graceful* knowledge a teacher can transfer to a student. Based on the above, we propose Views Knowledge Distillation (**VKD**), which transfers the knowledge lying in several views in a teacher-student fashion. VKD devises a two-stage procedure, which pins the visual variety as a teaching signal for a student who has to recover it using fewer views. We remark the following contributions: *i)* the student outperforms its teacher by a large margin, especially in the Image-To-Video setting; *ii)* a thorough investigation shows that the student focuses more on the target compared to its teacher and discards uninformative details; *iii)* importantly, we do not limit our analysis to a single domain, but instead achieve strong results on Person, Vehicle and Animal Re-ID.

2 Related Works

Image-To-Video Re-Identification The I2V Re-ID task has been successfully applied to multiple domains. In person Re-ID, [46] frames it as a point-to-set task, where image and video domains are aligned using a single deep network. The authors of [54] exploit time information by aggregating frames features via a Long-Short Term Memory. Eventually, a dedicated sub-network aggregates video features and match them against single image query ones. Authors of MGAT [3] employ a Graph Neural Network to model relationships between samples from different identities, thus enforcing similarity in the feature space. Dealing with vehicle Re-ID, authors from [26] introduce a large-scale dataset (VeRi-776) and propose PROVID and PROVID-BOT, which combine appearance and plate information in a progressive fashion. Differently, RAM [24] exploits multiple branches to extract global and local features, imposing a separate supervision on each branch and devising an additional one to predict vehicle attributes. VAMI [59] employs a viewpoint aware attention model to select core regions for different viewpoints. At inference time, they obtain a multiview descriptor through a conditional generative network, inferring information regarding the unobserved viewpoints. Differently, our approach asks the student to do it implicitly and in a lightweight fashion, thus avoiding the need for additional modules. Similarly to VAMI, [7] predicts the vehicle viewpoint along with appearance features; at inference, the framework provides distances according to the predicted viewpoint.

Knowledge Distillation has been first investigated in [35,13,53] for model compression: the idea is to instruct a lightweight model (student) to mimic the capabilities of a deeper one (teacher): as a gift, one could achieve both an acceleration in inference time as well as a reduction in memory consumption, without experiencing a large drop in performance. In this work, we benefit from the techniques proposed in [13,44] for a different purpose: we are not primarily engaged in educating a lightweight module, but on improving the original model itself. In this framework – often called *self-distillation* [9,51] – the transfer occurs from

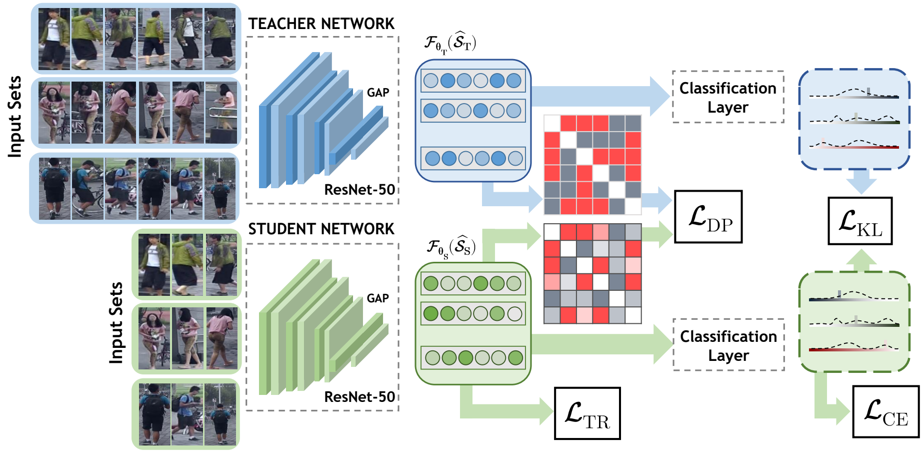


Fig. 2. An overview of Views Knowledge Distillation (VKD): a student network is optimised to mimic the behaviour of its teacher using fewer views.

the teacher to a student with the same architecture, with the aim of improving the overall performance at the end of the training. Here, we get a step ahead and introduce an asymmetry between the teacher and student, which has access to fewer frames. In this respect, our work closely relates to what [5] devises for Video Classification. Besides facing another task, a key difference subsists: while [5] limits the transfer along the temporal axis, our proposal advocates for distilling many views into fewer ones. On this latter point, we shall show that the teaching signal can be further enhanced when opening to diverse camera viewpoints. In the Re-Identification field, Temporal Knowledge Propagation (TKP) [10] similarly exploits intra-tracklet information to encourage the image-level representations to approach the video-level ones. In contrast with TKP: *i)* we do not rely on matching internal representations but instead their distances solely, thus making our proposal viable for cross-architecture transfer too; *ii)* at inference time, we make use of a single shared network to deal with both image and video domains, thus halving the number of parameters; *iii)* during transfer, we benefit from a larger visual variety, emerging from several viewpoints.

3 Method

We purse the aim of learning a function $\mathcal{F}_\theta(\mathcal{S})$ mapping a set of images $\mathcal{S} = (s_1, s_2, ..., s_n)$ into a representative embedding space. Specifically, $\mathcal{S}$ is a sequence of bounding boxes crops depicting a target (*e.g.* a person or a car), for which we are interested in inferring its corresponding identity. We take advantage of Convolutional Neural Networks (CNNs) for modelling $\mathcal{F}_\theta(\mathcal{S})$. Here, we look for two distinctive properties, aspiring to representations that are *i)* invariant to differences in background and viewpoint and *ii)* robust to a reduction in the number of query images. To achieve this, our proposal frames the training algorithm as a two-stage procedure, as follows:

- **First step** (Sec. 3.1): the backbone network is trained for the standard Video-To-Video setting.
- **Second step** (Sec. 3.2): we appoint it as the teacher and freeze its parameters. Then, a new network with the role of the student is instantiated. As depicted in Fig. 2, we feed frames representing different views as input to the teacher and ask the student to mimic the same outputs from fewer frames.

3.1 Teacher Network

Without loss of generality, we will refer to ResNet-50 [11] as the backbone network, namely a module $f_\theta : \mathbb{R}^{\mathrm{W}\times\mathrm{H}\times 3} \mapsto \mathbb{R}^{\mathrm{D}}$ mapping each image s_i from S to a fixed-size representation d_i (in this case $D = 2048$). Following previous works [28,10], we initialise the network weights on ImageNet and additionally include few amendments [28] to the architecture. First, we discard both the last ReLU activation function and final classification layer in favour of the BNNeck one [28] (*i.e.* batch normalisation followed by a linear layer). Second: to benefit from fine-grained spatial details, the stride of the last residual block is decreased from 2 to 1.

Set representation Given a set of images S, several solutions [27,54,22] may be assessed for designing the aggregation module, which fuses a variable-length set of representations $d_1, d_2, \ldots, d_n$ into a single one. Here, we naively compute the set-level embedding $\mathcal{F}(\mathcal{S})$ through a temporal average pooling. While we acknowledge better aggregation modules exist, we do not place our focus on devising a new one, but instead on improving the earlier features extractor.

Teacher optimisation We train the base network - which will be the teacher during the following stage - combining a classification term $\mathcal{L}_{\mathrm{CE}}$ (cross-entropy) with the triplet loss $\mathcal{L}_{\mathrm{TR}}$[1]. The first can be formulated as:

$$\mathcal{L}_{\mathrm{CE}} = -\boldsymbol{y} \log \hat{\boldsymbol{y}} \tag{1}$$

where $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$ represent the one-hot labels (identities) and the output of the softmax respectively. The second term $\mathcal{L}_{\mathrm{TR}}$ encourages distance constraints in feature space, moving closer representations from the same target and pulling away ones from different targets. Formally:

$$\mathcal{L}_{\mathrm{TR}} = \ln(1 + e^{\mathcal{D}\left(\mathcal{F}_\theta(\mathcal{S}_a^i), \mathcal{F}_\theta(\mathcal{S}_p^i)\right) - \mathcal{D}\left(\mathcal{F}_\theta(\mathcal{S}_a^i), \mathcal{F}_\theta(\mathcal{S}_n^j)\right)}), \tag{2}$$

where $\mathcal{S}_p$ and $\mathcal{S}_n$ are the hardest positive and negative for an anchor $\mathcal{S}_a$ within the batch. In doing so, we rely on the batch hard strategy [12] and include P identities coupled with K samples in each batch. Importantly, each set $\mathcal{S}^i$ comprises images drawn from the same tracklet [22,8].

[1] For the sake of clarity, all the loss terms are referred to one single example. In the implementation, we extend the penalties to a batch by averaging.

3.2 Views Knowledge Distillation (VKD)

After training the teacher, we propose to enrich its representation capabilities, especially when only few images are made available to the model. To achieve this, our proposal bets on the knowledge we can gather from different views, depicting the same object under different conditions. When facing re-identification tasks, one can often exploit camera viewpoints [55,33,25] to provide a larger variety of appearances for the target identity. Ideally, we would like to teach a new network to recover such a variety even from a single image. Since this information may not be inferred from a single frame, this can lead to an ill-posed task. Still, one can underpin this knowledge as a supervision signal, encouraging the student to focus on important details and favourably discover new ones. On this latter point, we refer the reader to Section 4.4 for a comprehensive discussion.

Views Knowledge Distillation (**VKD**) stresses this idea by forcing a student network $\mathcal{F}_{\theta_S}(\cdot)$ to match the outputs of the teacher $\mathcal{F}_{\theta_T}(\cdot)$. In doing so, we: *i)* allow the teacher to access frames $\hat{\mathcal{S}_T} = (\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_N)$ from different viewpoints; *ii)* force the student to mimic the teacher output starting from a subset $\hat{\mathcal{S}_S} = (\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_M) \subset \hat{\mathcal{S}_T}$ with cardinality $M < N$ (in our experiments, $M = 2$ and $N = 8$). The frames in $\hat{\mathcal{S}_S}$ are uniformly sampled from $\hat{\mathcal{S}_T}$ without replacement. This asymmetry between the teacher and the student leads to a self-distillation objective, where the latter can achieve better solutions despite inheriting the same architecture of the former.

To accomplish this, VKD exploits the Knowledge Distillation loss [13]:

$$\mathcal{L}_{\text{KD}} = \tau^2 \ \text{KL}(\boldsymbol{y}_T \parallel \boldsymbol{y}_S) \tag{3}$$

where $\boldsymbol{y}_T = \text{softmax}(\boldsymbol{h}_T/\tau)$ and $\boldsymbol{y}_S = \text{softmax}(\boldsymbol{h}_S/\tau)$ are the distributions – smoothed by a temperature τ – we attempt to match[2]. Since the student experiences a different task from the teacher one, Eq. 3 resembles the regularisation term imposed by [19] to relieve *catastrophic forgetting*. In a similar vein, we intend to *strengthen* the model in the presence of few images, whilst not *deteriorating* the capabilities it achieved with longer sequences.

In addition to fitting the output distribution of the teacher (Eq. 3), our proposal devises additional constraints on the embedding space learnt by the student. In details, VKD encourages the student to mirror the pairwise distances spanned by the teacher. Indicating with $\mathcal{D}_T[i,j] \equiv \mathcal{D}(\mathcal{F}_{\theta_T}(\hat{\mathcal{S}_T}[i]), \mathcal{F}_{\theta_T}(\hat{\mathcal{S}_T}[j]))$ the distance induced by the teacher between the i-th and j-th sets (the same notation $\mathcal{D}_S[i,j]$ also holds for the student), VKD seeks to minimise:

$$\mathcal{L}_{\text{DP}} = \sum_{(i,j) \in \binom{B}{2}} (\mathcal{D}_T[i,j] - \mathcal{D}_S[i,j])^2, \tag{4}$$

where B equals the batch size. Since the teacher has access to several viewpoints, we argue that distances spanned in its space yield a powerful description

[2] Since the teacher parameters are fixed, its entropy is constant and the objective of Eq. 3 reduces to the cross-entropy between $\boldsymbol{y}_T$ and $\boldsymbol{y}_S$.

of corresponding identities. From the student perspective, distances preservation provides additional semantic knowledge. Therefore, this holds an effective supervision signal, whose optimisation is made more challenging since fewer images are available to the student.
Even thought VKD focuses on *self-distillation*, we highlight that both $\mathcal{L}_{\text{KD}}$ and $\mathcal{L}_{\text{DP}}$ allow to match models with different embedding size, which would not be viable under the minimisation performed by [10]. As an example, it is still possible to distill ResNet-101 ($D = 2048$) into MobileNet-V2 [36] ($D = 1280$).

Student optimisation The VKD overall objective combines the distillation terms ($\mathcal{L}_{\text{KD}}$ and $\mathcal{L}_{\text{DP}}$) with the ones optimised by the teacher - $\mathcal{L}_{\text{CE}}$ and $\mathcal{L}_{\text{TR}}$ - that promote higher conditional likelihood w.r.t. ground truth labels. To sum up, VKD aims at strengthening the features of a CNN in Re-ID settings through the following optimisation problem:

$$\underset{\theta_S}{\operatorname{argmin}} \quad \mathcal{L}_{\text{VKD}} \equiv \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{TR}} + \alpha \mathcal{L}_{\text{KD}} + \beta \mathcal{L}_{\text{DP}}, \tag{5}$$

where α and β are two hyperparameters balancing the contributions to the total loss $\mathcal{L}_{\text{VKD}}$. We conclude with a final note on the student initialisation: we empirically found beneficial to start from the teacher weights θ_T except for the last convolutional block, which is reinitialised according to the ImageNet pretraining. We argue this represents a good compromise between exploring new configurations and exploiting the abilities already achieved by the teacher.

4 Experiments

Evaluation Protocols We indicate the query-gallery matching as x2x, where both x terms are features that can be generated by either a single (I) or multiple frames (V). In the **Image-to-Image (I2I)** setting features extracted from a query set image are matched against features from individual images in the gallery. This protocol – which has been amply employed for person Re-ID and face recognition – has a light impact in terms of resources footprint. However, a single image captures only a single view of the identity, which may not be enough for identities exhibiting multi-modal distributions. Contrarily, the **Video-to-Video (V2V)** setting enables to capture and combine different modes in the input, but with a significant increase in the number of operations and memory. Finally, the **Image-to-Video (I2V)** setting [58,59,24,48,26] represents a good compromise: building the gallery may be slow, but it is often performed offline. Moreover, matchings perform extremely fast, as a query comprise only a single image. We remark that *i)* We adopt the standard "*Cross Camera Validation*" protocol, not considering examples of the gallery from the same camera of the query at evaluation and *ii)* even if VKD relies on frames from different camera during train, we strictly adhere to the common schema and switch to tracklet-based inputs at evaluation time.

Evaluation Metrics While settings vary between different dataset, evaluation metrics for Re-Identification are shared by the vast majority of works in the field. In the followings, we report performance in terms of top-k accuracy and Mean Average Precision (mAP). By combining them, we evaluate VKD both in terms of accuracy and ranking performance.

4.1 Datasets

Person Re-ID: **MARS** [55] comprises 19680 tracklets from 6 different cameras, capturing 1260 different identities (split between 625 for the training set, 626 for the gallery and 622 for the query) with 59 frames per tracklet on average. MARS has shown to be a challenging dataset because it has been automatically annotated, leading to errors and false detections [56]. The **Duke** [33] dataset was first introduced for multi-target and multi-camera surveillance purposes, and then expanded to include person attributes and identities (414 ones). Consistently with [10,40,22,29], we use the **Duke-Video-ReID** [49] variant, where identities have been manually annotated from tracking information[3]. It comprises 5534 video tracklets from 8 different cameras, with 167 frames per tracklet on average. Following [10], we extract the first frame of every tracklet when testing in the I2V setting, for both MARS and Duke.
Vehicle Re-ID: **VeRi-776** [25] has been collected from 20 fixed cameras, capturing vehicles moving on a circular road in a 1.0 km^2 area. It contains 18397 tracklets with an average number of 6 frames per tracklet, capturing 775 identities split between train (575) and gallery (200). The query set shares identities consistently with the gallery, but differently from the other two sets it includes only a single image for each couple (id, camera). Consequently, all recent methods perform the evaluation following the I2V setting.
Animal Re-ID: The **Amur Tiger** [18] Re-Identification in the Wild (ATRW) is a recently introduced dataset collected from a diverse set of wild zoos. The training set includes 107 subjects and 17.6 images on average per identity; no information is provided to aggregate images into tracklets. It is possible to evaluate only the I2I setting through a remote http server. As done in [21], we horizontally flip the training images to duplicate the number of identities available, thus resulting in 214 training identities.

Implementation details Following [12,22] we adopt the following hyperparameters for MARS and Duke: *i)* each batch contains $P = 8$ identities with $K = 4$ samples each; *ii)* each sample comprises 8 images equally spaced in a tracklet. Differently, for image-based datasets (ATRW and VeRi-776) we increase P to 18 and use a single image at a time. All the teacher networks are trained for 300 epoch using Adam [17], setting the learning rate to 10^{-4} and multiplying it by 0.1 every 100 epochs. During the distillation stage, we feed $N = 8$ images to the teacher and $M = 2$ ones (picked at random) to the student. We found beneficial

[3] In the following, we refer to Duke-Video-ReID simply as Duke. Another variant of Duke named Duke-ReID exists [34], but it does not come with query tracklets.

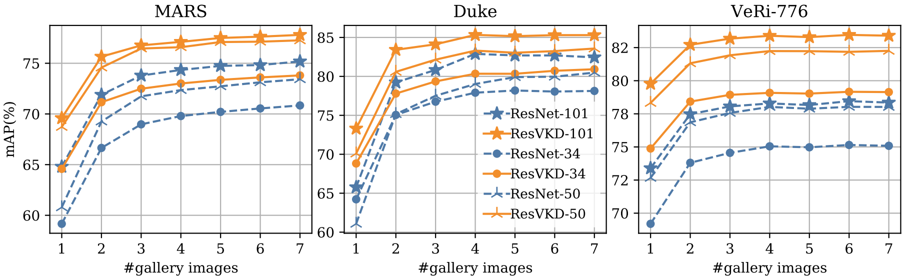


Fig. 3. Performance (mAP) in the Image-To-Video setting when changing at evaluation time the number of frames in each gallery tracklet.

Table 1. Self-Distillation results across datasets, settings and architectures.

	MARS				Duke				VeRi-776			
	I2V		V2V		I2V		V2V		I2I		I2V	
	cmc1	mAP	cmc1	mAP	cmc1	mAP	cmc1	mAP	cmc1	mAP	cmc1	mAP
ResNet-34	80.81	70.74	86.67	78.03	81.34	78.70	93.45	**91.88**	92.97	70.30	93.80	75.01
ResVKD-34	**82.17**	**73.68**	**87.83**	**79.50**	**83.33**	**80.60**	**93.73**	91.62	**95.29**	**75.97**	**94.76**	**79.02**
ResNet-50	82.22	73.38	87.88	81.13	82.34	80.19	**95.01**	**94.17**	93.50	73.19	93.33	77.88
ResVKD-50	**83.89**	**77.27**	**88.74**	**82.22**	**85.61**	**83.81**	**95.01**	93.41	**95.23**	**79.17**	**95.17**	**82.16**
ResNet-101	82.78	74.94	88.59	81.66	83.76	82.89	**96.01**	**94.73**	94.28	74.27	94.46	78.20
ResVKD-101	**85.91**	**77.64**	**89.60**	**82.65**	**86.32**	**85.11**	95.44	93.67	**95.53**	**80.62**	**96.07**	**83.26**
ResNet-50bam	82.58	74.11	88.54	81.19	82.48	80.24	94.87	**93.82**	93.33	72.73	93.80	77.14
ResVKD-50bam	**84.34**	**78.13**	**89.39**	**83.07**	**86.18**	**84.54**	**95.16**	93.45	**96.01**	**78.67**	**95.71**	**81.57**
DenseNet-121	82.68	74.34	89.75	81.93	82.91	80.26	93.73	91.73	91.24	69.24	91.84	74.52
DenseVKD-121	**84.04**	**77.09**	**89.80**	**82.84**	**86.47**	**84.14**	**95.44**	**93.54**	**94.34**	**76.23**	**93.80**	**79.76**
MobileNet-V2	78.64	67.94	85.96	77.10	78.06	74.73	93.30	91.56	88.80	64.68	89.81	69.90
MobileVKD-V2	**83.33**	**73.95**	**88.13**	**79.62**	**83.76**	**80.83**	**94.30**	**92.51**	**92.85**	**70.93**	**92.61**	**75.27**

to train the student longer: so, we set the number of epochs to 500 and the learning rate decay steps at 300 and 450. We keep fixed $\tau = 10$ (Eq. 3), $\alpha = 10^{-1}$ and $\beta = 10^{-4}$ (Eq. 5) in all experiments. To improve generalisation, we apply data augmentation as described in [28]. Finally, we put the teacher in training mode during distillation (consequently, batch normalisation [15] statistics are computed on a batch basis): as observed in [2], this provides more accurate teacher labels.

4.2 Self-Distillation

In this section we show the benefits of self-distillation for person and vehicle re-id. We indicate the teacher with the name of the backbone (e.g. ResNet-50) and append "VKD" for its student (e.g. ResVKD-50). To validate our ideas, we do not limit the analysis on ResNet-*; contrarily, we test self-distillation on DenseNet-121 [14] and MobileNet-V2 1.0X [36]. Since learning what and where to look represents an appealing property when dealing with Re-ID tasks [8], we additionally conduct experiments on ResNet-50 coupled with Bottleneck Attention Modules [31] (ResNet-50bam).

Table 2. MARS **I2V**

Method	top_1	top_5	mAP
P2SNet[46]	55.3	72.9	-
Zhang[54]	56.5	70.6	-
XQDA[20]	67.2	81.9	54.9
TKP[10]	75.6	87.6	65.1
STE-NVAN[22]	80.3	-	68.8
NVAN[22]	80.1	-	70.2
MGAT[3]	81.1	92.2	71.8
ResVKD-50	83.9	93.2	77.3
ResVKD-50bam	**84.3**	**93.5**	**78.1**

Table 3. Duke **I2V**

Method	top_1	top_5	mAP
STE-NVAN[22]	42.2	-	41.3
TKP[10]	77.9	-	75.9
NVAN[22]	78.4	-	76.7
ResVKD-50	85.6	93.9	83.8
ResVKD-50bam	**86.2**	**94.2**	**84.5**

Table 4. VeRi-776 **I2V**

Method	top_1	top_5	mAP
PROVID[26]	76.8	91.4	48.5
VFL-LSTM[1]	88.0	94.6	59.2
RAM[24]	88.6	-	61.5
VANet[7]	89.8	96.0	66.3
PAMTRI[42]	92.9	92.9	71.9
SAN[32]	93.3	97.1	72.5
PROVID-BOT[26]	**96.1**	97.9	77.2
ResVKD-50	95.2	**98.0**	**82.2**
ResVKD-50bam	95.7	98.0	81.6

Table 1 reports the comparisons for different backbones: in the vast majority of the settings, *the student outperforms its teacher*. Such a finding is particularly evident when looking at the I2V setting, where the mAP metric gains 4.04% on average. The same holds for the I2I setting on VeRi-776, and in part also on V2V. We draw the following remarks: *i)* in accordance with the objective the student seeks to optimise, our proposal leads to greater improvements when few images are available; *ii)* bridging the gap between I2V and V2V does not imply a significant information loss when more frames are available; on the contrary it sometimes results in superior performance; *iii)* the previous considerations hold true across different architectures. As an additional proof, plots from Figure 3 draw a comparison between models before and after distillation. VKD improves metrics considerably on all three dataset, as highlighted by the bias between the teachers and their corresponding students. Surprisingly, this often applies when comparing lighter students with deeper teachers: as an example, ResVKD-34 scores better than even ResNet-101 on VeRi-776, regardless of the number of images sampled for a gallery tracklet.

4.3 Comparison with State-Of-The-Art

Image-To-Video Tables 2, 3 and 4 report a thorough comparison with current state-of-the-art (SOTA) methods, on MARS, Duke and VeRi-776 respectively. As common practice [10,3,32], we focus our analysis on ResNet-50, and in particular on its distilled variants ResVKD-50 and ResVKD-50bam. Our method clearly outperforms other competitors, with an increase in mAP w.r.t. top-scorers of 6.3% on MARS, 8.6% on Duke and 5% on VeRi-776. This results is totally in line with our goal of conferring robustness when just a single image is provided as query. In doing so, we do not make any task-specific assumption, thus rendering our proposal easily applicable to both person and vehicle Re-ID.

Video-To-Video Analogously, we conduct experiments on the V2V setting and report results in Table 5 (MARS) and Table 6 (Duke)[4]. Here, VKD yields the following results: on the one hand, on MARS it pushes a baseline architecture as

[4] Since VeRi-776 does not include any tracklet information in the query set, following all other competitors we limit experiments to the I2V setting only.

Table 5. MARS **V2V**

Method	top$_1$	top$_5$	mAP
DuATN[40]	81.2	92.5	67.7
TKP[10]	84.0	93.7	73.3
CSACSE+OF[6]	86.3	94.7	76.1
STA[8]	86.3	95.7	80.8
STE-NVAN[22]	88.9	-	81.2
NVAN[22]	**90.0**	-	82.8
ResVKD-50	88.7	96.1	82.2
ResVKD-50bam	89.4	**96.8**	**83.1**

Table 6. Duke **V2V**

Method	top$_1$	top$_5$	mAP
DuATN[40]	81.2	92.5	67.7
Matiyali[29]	89.3	98.3	88.5
TKP[10]	94.0	-	91.7
STE-NVAN[22]	95.2	-	93.5
STA[8]	96.2	**99.3**	94.9
NVAN[22]	**96.3**	-	**94.9**
ResVKD-50	95.0	98.9	93.4
ResVKD-50bam	95.2	98.6	93.5

Table 7. ATRW **I2I**

Method	top$_1$	top$_5$	mAP
PPbM-a [18]	82.5	93.7	62.9
PPbM-b [18]	83.3	93.2	60.3
NWPU [52]	94.7	96.7	75.1
BRL [23]	94.0	96.7	77.0
NBU [21]	**95.6**	**97.9**	**81.6**
ResNet-101	92.3	93.5	75.7
ResVKD-101	92.0	96.4	77.2

ResVKD-50 close to NVAN and STE-NVAN [22], the latter being tailored for the V2V setting. Moreover – when exploiting spatial attention modules (ResVKD-50bam) – it establishes new SOTA results, suggesting that a positive transfer occurs when matching tracklets also. On the other hand, the same does not hold true for Duke, where exploiting video features as in STA [8] and NVAN appears rewarding. We leave the investigation of further improvements on V2V to future works. As of today, our proposals is the only one guaranteeing consistent and stable results under both I2V and V2V settings.

4.4 Analysis on VKD

In the absence of camera information. Here, we address the setting where we do not have access to camera information. As an example, when dealing with animal re-id this information often lacks and datasets come with images and labels solely: can VKD still provide any improvement? We think so, as one can still exploit the visual diversity lying in a bag of randomly sampled images. To demonstrate our claim, we test our proposal on Amur Tigers re-identification (ATRW), which was conceived as an Image-To-Image dataset. During comparisons: *i)* since other works do not conform to a unique backbone, here we opt for ResNet-101; *ii)* as common practice in this benchmark [21,23,52], we leverage re-ranking [57]. Table 7 compares VKD against the top scorers in the "Computer Vision for Wildlife Conservation 2019" competition. Importantly, the student ResVKD-101 improves over its teacher (1.5% on mAP and 2.9% on top$_5$) and places second behind [21], confirming its effectiveness in a challenging scenario. Moreover, we remark that the top-scorer requires additional annotations - such as body parts and pose information - which we do not exploit.

Distilling viewpoints *vs* time. Figure 4 shows results of distilling knowledge from multiple views against time (*i.e.* multiple frames from a tracklet). On one side, as multiple views hold more "*visual variety*", the student builds a more invariant representation for the identity. On the opposite, a student trained with tracklets still considerably outperforms the teacher. This shows that, albeit the visual variety is reduced, our distillation approach still successfully exploits it.

VKD reduces the camera bias. As pointed out in [43], the appearance encoded by a CNN is heavily affected by external factors surrounding the target

Table 8. Analysis on camera bias, in terms of viewpoint classification accuracy.

	MARS	Duke	VeRi-776
Prior Class.	0.19	0.14	0.06
ResNet-34	**0.61**	**0.73**	**0.55**
ResVKD-34	0.40	0.67	0.51
ResNet-101	**0.71**	**0.72**	**0.73**
ResVKD-101	0.51	0.70	0.68

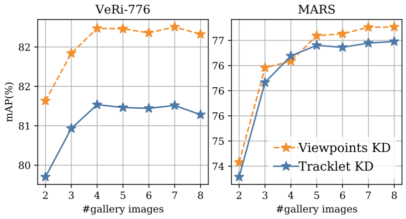


Fig. 4. Comparison between time and viewpoints distillation.

Table 9. Analysis on different modalities for training the teacher.

	Input Bags	MARS				Duke			
		I2V		V2V		I2V		V2V	
		cmc1	mAP	cmc1	mAP	cmc1	mAP	cmc1	mAP
ResNet-50	Viewpoints ($N = 2$)	80.05	71.16	84.70	76.99	77.21	75.19	89.17	87.70
ResNet-50	Tracklets ($N = 2$)	82.32	73.69	87.32	79.91	81.77	80.34	93.73	92.88
ResVKD-50	Viewpoints ($N = 2$)	**83.89**	**77.27**	**88.74**	**82.22**	**85.61**	**83.81**	**95.01**	**93.41**

object (*e.g.* different backgrounds, viewpoints, illumination . . .). In this respect, is our proposal effective for reducing such a bias? To investigate this aspect, we perform a camera classification test on both the teacher (*e.g.* ResNet-34) and the student network (*e.g.* ResVKD-34) by fitting a linear classifier on top of their features, with the aim of predicting the camera the picture is taken from. We freeze all backbone layers and train for 300 epochs (lr $= 10^{-3}$ and halved every 50 epochs). Table 8 reports performance on the gallery set for different teachers and students. To provide a better understanding, we include a baseline that computes predictions by sampling from the cameras prior distribution. As expected: *i)* the teacher outperforms the baseline, suggesting it is in fact biased towards background conditions; *ii)* the student consistently reduces the bias, confirming VKD encourages the student to focus on identities features and drops viewpoint-specific information. Finally, it is noted that time-based distillation does not yield the bias reduction we observe for VKD (see supplementary materials).

Can performance of the student be obtained without distillation? To highlight the advantages of the two-stage procedure above discussed, we here consider a teacher (ResNet-50) trained straightly using few frames ($N = 2$) only. First two rows of Table 9 show the performance achieved by this baseline (using tracklets and views respectively). Results show that major improvements come from the teacher-student paradigm we devise (third row), instead of simply reducing the number of input images available to the teacher.

Student explanation. To further assess the differences between teachers and students, we leverage GradCam [39] to highlight the input regions that have been considered paramount for predicting the identity. Figure 5 depicts the impact of

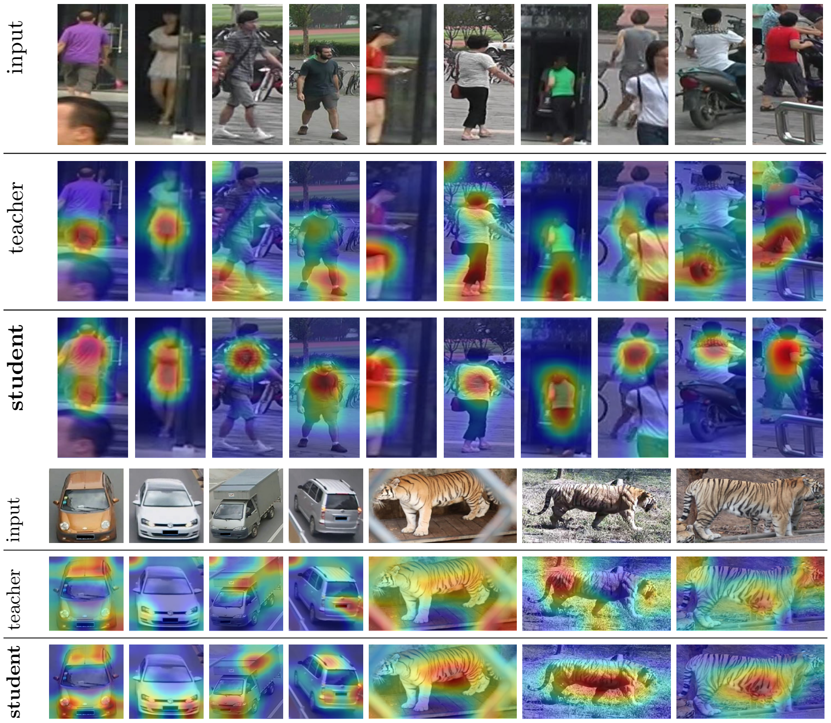


Fig. 5. Model explanation via GradCam[39] on ResNet-50 (teacher) and ResVKD-50 (student). The student favours visual details characterising the target, discarding external and uninformative patterns.

VKD for various examples from MARS, VeRi-776 and ATRW. In general, the student network pays more attention to the subject of interest compared to its teacher. For person and animal Re-ID, background features are suppressed (third and last columns) while attention tends to spread to the whole subject (first and fourth columns). When dealing with vehicle Re-ID, one can appreciate how the attention becomes equally distributed on symmetric parts, such as front and rear lights (second, seventh and last columns). Please see supplementary materials for more examples, as well as a qualitative analysis of some of our model errors.

Cross-Distillation. Differently from other approaches [5,10], VKD is not confined to self-distillation, but instead allows the knowledge transfer from a complex architecture (e.g. ResNet-101) into a simpler one, such as MobileNet-V2 or ResNet-34 (*cross-distillation*). Here, drawing inspirations from the model compression area, we attempt to reduce the network complexity but, at the same time, increase the profit we already achieve through self-distillation. In this respect, Table 11 shows results of cross-distillation, for various combinations of a teacher and a student. It appears that *better the teacher, better the student*: as

Table 10. Ablation study questioning the impact of each loss term.

					MARS				Duke				VeRi-776			
	$\mathcal{L}_{\text{CE}}$	$\mathcal{L}_{\text{TR}}$	$\mathcal{L}_{\text{KL}}$	$\mathcal{L}_{\text{DP}}$	I2V		V2V		I2V		V2V		I2I		I2V	
					cmc1	mAP	cmc1	mAP	cmc1	mAP	cmc1	mAP	cmc1	mAP	cmc1	mAP
ResNet-50 (teacher)					82.22	73.38	87.88	81.13	82.34	80.19	95.01	94.17	93.50	73.19	93.33	77.88
ResVKD-50 (students)	✓	✓	✗	✗	80.25	71.26	85.71	77.45	82.62	81.03	94.73	93.29	92.61	70.06	92.31	74.82
	✗	✗	✓	✓	84.09	**77.37**	88.33	82.06	84.90	83.56	95.30	93.79	95.29	**79.35**	**95.29**	**82.26**
	✓	✓	✓	✗	83.54	75.18	88.43	80.77	83.90	82.34	94.30	92.97	**95.41**	78.01	95.17	81.32
	✓	✓	✗	✓	**84.29**	76.82	88.69	81.82	85.33	83.45	**95.44**	**93.90**	94.40	77.41	94.87	80.93
	✓	✓	✓	✓	83.89	77.27	**88.74**	**82.22**	**85.61**	**83.81**	95.01	93.41	95.23	79.17	95.17	82.16

Table 11. Measuring the benefit of VKD for cross-architecture transfer.

		MARS		Duke		VeRi-776	
Student (#params)	Teacher (#params)	I2V		I2V		I2V	
		cmc1	mAP	cmc1	mAP	cmc1	mAP
ResNet-34 (21.2M)	ResNet-34 (21.2M)	82.17	73.68	83.33	80.60	94.76	79.02
	ResNet-50 (23.5M)	83.08	75.45	84.05	82.61	**95.05**	80.05
	ResNet-101 (42.5M)	**83.43**	**75.47**	**85.75**	**83.65**	94.87	**80.41**
ResNet-50 (23.5M)	ResNet-50 (23.5M)	83.89	77.27	85.61	83.81	95.17	82.16
	ResNet-101 (42.5M)	**84.49**	**77.47**	**85.90**	**84.34**	**95.41**	**82.99**
MobileNet-V2 (2.2M)	MobileNet-V2 (2.2M)	83.33	73.95	**83.76**	80.83	92.61	75.27
	ResNet-101 (42.5M)	**83.38**	**74.72**	**83.76**	**81.36**	**93.03**	**76.38**

an example, ResVKD-34 gains an additional 3% mAP on Duke when educated by ResNet-101 rather than "itself".

On the impact of loss terms. We perform a thorough ablation study (Table 10) on the student loss (Eq. 5). It is noted that leveraging ground truth solely (second row) hurts performance. Differently, best performance for both metrics are obtained exploiting teacher signal (from the third row onward), with particular emphasis to $\mathcal{L}_{\text{DP}}$, which proves to be a fundamental component.

5 Conclusions

An effective Re-ID method requires visual descriptors robust to changes in both background appearances and viewpoints. Moreover, its effectiveness should be ensured even for queries composed of a single image. To accomplish these, we proposed Views Knowledge Distillation (VKD), a teacher-student approach where the student observes only a small subset of input views. This strategy encourages the student to discover better representations: as a result, it outperforms its teacher at the end of the training. Importantly, VKD shows robustness on diverse domains (person, vehicle and animal), surpassing by a wide margin the state of the art in I2V. Thanks to extensive analysis, we highlight that the student presents stronger focus on the target and reduces the camera bias.

Acknowledgement The authors would like to acknowledge Farm4Trade for its financial and technical support.

References

1. Alfasly, S.A.S., Hu, Y., Liang, T., Jin, X., Zhao, Q., Liu, B.: Variational representation learning for vehicle re-identification. In: IEEE International Conference on Image Processing (2019)
2. Bagherinezhad, H., Horton, M., Rastegari, M., Farhadi, A.: Label refinery: Improving imagenet classification through label progression. arXiv preprint arXiv:1805.02641 (2018)
3. Bao, L., Ma, B., Chang, H., Chen, X.: Masked graph attention network for person re-identification. In: IEEE International Conference on Computer Vision and Pattern Recognition Workshops (2019)
4. Bergamini, L., Porrello, A., Dondona, A.C., Del Negro, E., Mattioli, M., D'alterio, N., Calderara, S.: Multi-views embedding for cattle re-identification. In: IEEE International Conference on Signal-Image Technology & Internet-Based Systems (2018)
5. Bhardwaj, S., Srinivasan, M., Khapra, M.M.: Efficient video classification using fewer frames. In: IEEE International Conference on Computer Vision and Pattern Recognition (2019)
6. Chen, D., Li, H., Xiao, T., Yi, S., Wang, X.: Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In: IEEE International Conference on Computer Vision and Pattern Recognition (2018)
7. Chu, R., Sun, Y., Li, Y., Liu, Z., Zhang, C., Wei, Y.: Vehicle re-identification with viewpoint-aware metric learning. In: IEEE International Conference on Computer Vision (2019)
8. Fu, Y., Wang, X., Wei, Y., Huang, T.: Sta: Spatial-temporal attention for large-scale video-based person re-identification. In: AAAI Conference on Artificial Intelligence (2019)
9. Furlanello, T., Lipton, Z.C., Tschannen, M., Itti, L., Anandkumar, A.: Born again neural networks. International Conference on Machine Learning (2018)
10. Gu, X., Ma, B., Chang, H., Shan, S., Chen, X.: Temporal knowledge propagation for image-to-video person re-identification. In: IEEE International Conference on Computer Vision (2019)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE International Conference on Computer Vision and Pattern Recognition (2016)
12. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)
13. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. In: NeurIPS Deep Learning and Representation Learning Workshop (2015)
14. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: IEEE International Conference on Computer Vision and Pattern Recognition (2017)
15. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning (2015)
16. Khan, S.D., Ullah, H.: A survey of advances in vision-based vehicle re-identification. Computer Vision and Image Understanding (2019)
17. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. International Conference on Learning Representations (2015)

18. Li, S., Li, J., Lin, W., Tang, H.: Amur tiger re-identification in the wild. arXiv preprint arXiv:1906.05586 (2019)
19. Li, Z., Hoiem, D.: Learning without forgetting. In: European Conference on Computer Vision (2016)
20. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: IEEE International Conference on Computer Vision and Pattern Recognition (2015)
21. Liu, C., Zhang, R., Guo, L.: Part-pose guided amur tiger re-identification. In: IEEE International Conference on Computer Vision Workshops (2019)
22. 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: British Machine Vision Conference (2019)
23. Liu, N., Zhao, Q., Zhang, N., Cheng, X., Zhu, J.: Pose-guided complementary features learning for amur tiger re-identification. In: IEEE International Conference on Computer Vision Workshops (2019)
24. Liu, X., Zhang, S., Huang, Q., Gao, W.: Ram: a region-aware deep model for vehicle re-identification. In: IEEE International Conference on Multimedia and Expo (ICME) (2018)
25. Liu, X., Liu, W., Mei, T., Ma, H.: A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In: European Conference on Computer Vision (2016)
26. Liu, X., Liu, W., Mei, T., Ma, H.: Provid: Progressive and multimodal vehicle reidentification for large-scale urban surveillance. IEEE Transactions on Multimedia (2017)
27. Liu, Y., Junjie, Y., Ouyang, W.: Quality aware network for set to set recognition. In: IEEE International Conference on Computer Vision (2017)
28. Luo, H., Gu, Y., Liao, X., Lai, S., Jiang, W.: Bag of tricks and a strong baseline for deep person re-identification. In: IEEE International Conference on Computer Vision and Pattern Recognition Workshops (2019)
29. Matiyali, N., Sharma, G.: Video person re-identification using learned clip similarity aggregation. In: The IEEE Winter Conference on Applications of Computer Vision (2020)
30. Nguyen, T.B., Le, T.L., Nguyen, D.D., Pham, D.T.: A reliable image-to-video person re-identification based on feature fusion. In: Asian conference on intelligent information and database systems (2018)
31. Park, J., Woo, S., Lee, J., Kweon, I.S.: BAM: bottleneck attention module. In: British Machine Vision Conference (2018)
32. Qian, J., Jiang, W., Luo, H., Yu, H.: Stripe-based and attribute-aware network: A two-branch deep model for vehicle re-identification. arXiv preprint arXiv:1910.05549 (2019)
33. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: European Conference on Computer Vision (2016)
34. Ristani, E., Tomasi, C.: Features for multi-target multi-camera tracking and reidentification. In: IEEE International Conference on Computer Vision and Pattern Recognition (2018)
35. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. International Conference on Learning Representations (2015)

36. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: IEEE International Conference on Computer Vision and Pattern Recognition (2018)
37. Schneider, S., Taylor, G.W., Linquist, S., Kremer, S.C.: Past, present and future approaches using computer vision for animal re-identification from camera trap data. Methods in Ecology and Evolution (2019)
38. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: IEEE International Conference on Computer Vision and Pattern Recognition (2015)
39. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: IEEE International Conference on Computer Vision and Pattern Recognition (2017)
40. Si, J., Zhang, H., Li, C.G., Kuen, J., Kong, X., Kot, A.C., Wang, G.: Dual attention matching network for context-aware feature sequence based person reidentification. In: IEEE International Conference on Computer Vision and Pattern Recognition (2018)
41. Sohn, K.: Improved deep metric learning with multi-class n-pair loss objective. In: Neural Information Processing Systems (2016)
42. Tang, Z., Naphade, M., Birchfield, S., Tremblay, J., Hodge, W., Kumar, R., Wang, S., Yang, X.: Pamtri: Pose-aware multi-task learning for vehicle re-identification using highly randomized synthetic data. In: IEEE International Conference on Computer Vision (2019)
43. Tian, M., Yi, S., Li, H., Li, S., Zhang, X., Shi, J., Yan, J., Wang, X.: Eliminating background-bias for robust person re-identification. In: IEEE International Conference on Computer Vision and Pattern Recognition (2018)
44. Tung, F., Mori, G.: Similarity-preserving knowledge distillation. In: IEEE International Conference on Computer Vision (2019)
45. Ustinova, E., Lempitsky, V.: Learning deep embeddings with histogram loss. In: Neural Information Processing Systems (2016)
46. Wang, G., Lai, J., Xie, X.: P2snet: can an image match a video for person reidentification in an end-to-end way? IEEE Transactions on Circuits and Systems for Video Technology (2017)
47. Wang, J., Zhou, F., Wen, S., Liu, X., Lin, Y.: Deep metric learning with angular loss. In: IEEE International Conference on Computer Vision and Pattern Recognition (2017)
48. Wang, Z., Tang, L., Liu, X., Yao, Z., Yi, S., Shao, J., Yan, J., Wang, S., Li, H., Wang, X.: Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification. In: IEEE International Conference on Computer Vision (2017)
49. 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: IEEE International Conference on Computer Vision and Pattern Recognition (2018)
50. Xie, Z., Li, L., Zhong, X., Zhong, L., Xiang, J.: Image-to-video person reidentification with cross-modal embeddings. Pattern Recognition Letters (2019)
51. Yang, C., Xie, L., Qiao, S., Yuille, A.: Knowledge distillation in generations: More tolerant teachers educate better students. arXiv preprint arXiv:1805.05551 (2018)
52. Yu, J., Su, H., Liu, J., Yang, Z., Zhang, Z., Zhu, Y., Yang, L., Jiao, B.: A strong baseline for tiger re-id and its bag of tricks. In: IEEE International Conference on Computer Vision Workshops (2019)

53. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In: International Conference on Learning Representations (2017)
54. Zhang, D., Wu, W., Cheng, H., Zhang, R., Dong, Z., Cai, Z.: Image-to-video person re-identification with temporally memorized similarity learning. IEEE Transactions on Circuits and Systems for Video Technology (2017)
55. 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: European Conference on Computer Vision (2016)
56. Zheng, L., Yang, Y., Hauptmann, A.G.: Person re-identification: Past, present and future. arXiv preprint arXiv:1610.02984 (2016)
57. Zhong, Z., Zheng, L., Cao, D., Li, S.: Re-ranking person re-identification with k-reciprocal encoding. In: IEEE International Conference on Computer Vision and Pattern Recognition (2017)
58. Zhou, Y., Liu, L., Shao, L.: Vehicle re-identification by deep hidden multi-view inference. IEEE Transactions on Image Processing (2018)
59. Zhou, Y., Shao, L.: Aware attentive multi-view inference for vehicle re-identification. In: IEEE International Conference on Computer Vision and Pattern Recognition (2018)