CycAs: Self-supervised Cycle Association for Learning Re-identifiable Descriptions

Zhongdao Wang[1], Jingwei Zhang[1], Liang Zheng[2], Yixuan Liu[1], Yifan Sun[3], Yali Li[1], and Shengjin Wang[1]

[1] Department of Electronic Engineering, Tsinghua University
wcd17@mails.tsinghua.edu.cn, {liyali13,wgsgj}@tsinghua.edu.cn
[2] Australian National University liang.zheng@anu.edu.au
[3] MEGVII Technology peter@megvii.com

Abstract. This paper proposes a self-supervised learning method for the person re-identification (re-ID) problem, where existing unsupervised methods usually rely on pseudo labels, such as those from video tracklets or clustering. A potential drawback of using pseudo labels is that errors may accumulate and it is challenging to estimate the number of pseudo IDs. We introduce a different unsupervised method that allows us to learn pedestrian embeddings from raw videos, without resorting to pseudo labels. The goal is to construct a self-supervised pretext task that matches the person re-ID objective. Inspired by the *data association* concept in multi-object tracking, we propose the **Cyc**le **As**sociation (**CycAs**) task: after performing data association between a pair of video frames forward and then backward, a pedestrian instance is supposed to be associated to itself. To fulfill this goal, the model must learn a meaningful representation that can well describe correspondences between instances in frame pairs. We adapt the discrete association process to a differentiable form, such that end-to-end training becomes feasible. Experiments are conducted in two aspects: We first compare our method with existing unsupervised re-ID methods on seven benchmarks and demonstrate CycAs' superiority. Then, to further validate the practical value of CycAs in real-world applications, we perform training on self-collected videos and report promising performance on standard test sets.

Keywords: self-supervised, cycle consistency, person re-ID

1 Introduction

Self-supervised learning is a recent solution to the lack of labeled data in various computer vision areas like optical flow estimation [17,45], disparity/depth estimation [7,20,37], pixel/object tracking [29,10,26] and universal representation learning [4,6,28,19,40,1]. As a branch of unsupervised learning, the idea of self-supervised learning is to construct a *pretext task*. It is supposed that free supervision signals of the pretext task can be generated directly from the data, and the challenges lie in the design of the pretext task, so that the learned representation matches the task objective.

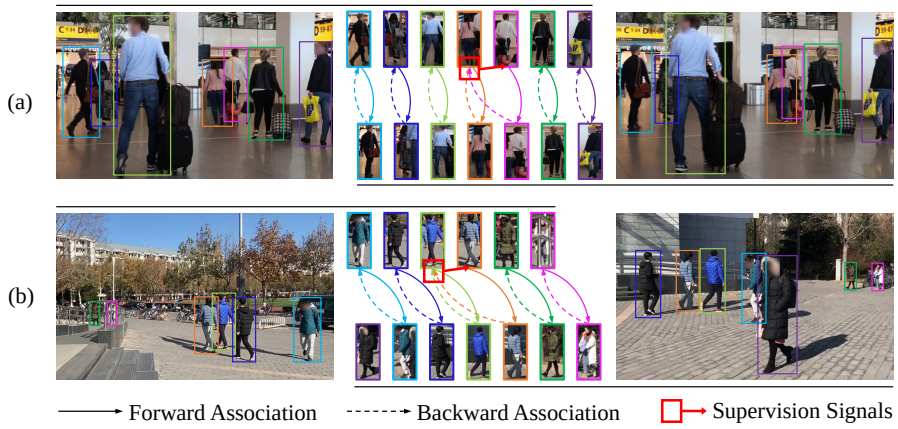


Fig. 1. Cycle association between (a). Temporal consecutive frames in a single video; (b). Temporal-aligned frames from two cameras that shares an overlapped visual field. In these examples, (a) shows a symmetric case (two frames contain the same group of identities) and (b) shows an asymmetric case.

Self-supervised / unsupervised learning finds critical significance in the person re-identification (re-ID) area, because of the high annotation cost. The goal of the re-ID task is to search for cross-camera bounding boxes containing the same person with a query: the cross-camera requirement increases the burden of data annotation. Existing unsupervised methods usually rely on pseudo labels that can be obtained from video tracklets [34,13] or clustering [5,16]. This strategy achieves descent accuracy, but its potential drawback consists of the error accumulation and the challenge in estimating the number of pseudo identities.

We are interested in finding a pretext task for re-ID, such that pedestrian descriptors can be learned in a self-supervised way. We are motivated by the cycle association between a pair of video frames that contain multiple persons. Considering two temporal-consecutive frames from such a video, because of the short time interval between them, they usually share the same group of identities (Figure 1 (a)). With perfect person representations, if we apply data association[4] between the two frames, we can find accurate correspondences between the two sets of identities. Further, if we perform forward data association and then backward, an instance is supposed to be associated to itself. Based on this motivation, we construct the **Cy**cle **As**sociation (CycAs) pretext task: apply data association in a cycle, *i.e.*, forward then backward, and use the inconsistency in the cycle association matrix as supervision signals. To maximize the cycle consistency or minimize inconsistency, the model inclines to learn a meaningful representation that can well-describe correspondences between instances [29,10].

[4] In Multi-Object Tracking (MOT) [11], data association means matching observations in a new frame to a set of tracked trajectories. In our case, we simplify the concept to matching observations between a frame pair.

In the above pretext task, we face a dilemma, *i.e.,* appearance diversity vs. accurate association. To learn robust descriptors, we need a pedestrian to exhibit sufficient diversity between two frames, *e.g.,* people from frames with long temporal interval. It means that persons from the two frames do not come from the exact same group of identities, creating asymmetry and leading to inaccurate association. On the other hand, if we use consecutive frames from a single video to ensure accurate association, the appearance variation would be small, compromising the descriptor robustness. To address this dilemma, we modify the optimization objective of CycAs by a relaxation to allow tolerance to a moderate level of asymmetry. Moreover, we adopt a two-stage training procedure: first train the pretext task using consecutive video frames, and then for fine-tuning add frame pairs from different cameras with overlapping field of view (FOV) (see Figure 1 (b)). This training process allows learning correspondences across cameras and benefits feature robustness to large appearance variation, an essential requirement of re-ID.

Experiments are conducted in two aspects: First, we compare with existing unsupervised re-ID methods on a wide range of public datasets under the same train / test protocol. Second, to further validate the practical value of the proposed method in real-world applications, we train CycAs with self-collected videos and conduct cross-domain evaluation on Market-1501 [42] and DukeMTMC-ReID [21]. Very promising results are shown compared with some direct transfer baselines. Our strengths are summarized below,

(1) We propose CycAs, a self-supervised pretext task for person re-ID. Strong features can be learned from associating persons between videos frames.
(2) We design a relaxed optimization objective for CycAs, allowing leveraging frames with large appearance variation. It significantly improves the discriminative ability of learned representations.
(3) We showcase the strength of CycAs on public benchmarks. We further validate the practical value of CycAs using self-collected videos as training data.

2 Related Work

Unsupervised Person re-ID. Most existing deep learning based unsupervised person re-ID approaches in literature can be categories into there paradigms:

(1). Domain adaptation methods [43,3,25,44]. These methods start with a supervised learned model which is pre-trained using the source domain data, and then transfer knowledge from the unlabeled target domain data.
(2). Clustering-based methods [5,16,33]. These methods usually adopt the iterative clustering-and-training strategy. Unlabeled data are grouped with clustering algorithms and assigned pseudo labels, then the model is fine-tuned with these pseudo labels. Such procedure repeats until convergence.
(3). Tracklet-based methods [13,14,34]. These methods label different trackelts, from a specific camera, as different identities, and train multiple classification tasks for multiple cameras in a parallel manner. Cross-camera matching is usually modeled as metric learning with pseudo labels.

Domain adaptation methods need supervised learned pre-train models so they are not totally unsupervised in essence. Clustering/Tracklet-based methods all rely on pseudo labels. In contrast, the proposed CycAs is unsupervised and does not require pseudo labels.

Self-supervised Learning. As a form of unsupervised learning, self-supervised learning seeks to learn from unlabeled data by constructing pretext tasks. For instance, image-level pretext tasks such as predicting context [4], rotation [6] and color [40] are useful for learning universal visual representations.

Video-level pretext tasks [28,29,10] are recently prevalent due to the large amount of available web videos and the fertile information they contain. Our work is closely related to a line of video-based self-supervised methods that utilize cycle consistency as free supervision [29,10,26]. While these methods usually focus on learning fine-grained correspondences between pixels [29,10], thus mainly tackle the tracking problem, ours focus more on learning high-level semantic correspondences and is more adaptive to the re-ID problem. To the best of our knowledge, we are the first to provide a self-supervised solution to learn re-identifiable object descriptions.

3 Proposed Cycle Association Task

3.1 Overview

Our goal is to learn a discriminative pedestrian embedding function Φ by learning correspondences between two sets of person images $\boldsymbol{I}_1$ and $\boldsymbol{I}_2$. Specifically, $\boldsymbol{I}_1$ and $\boldsymbol{I}_2$ are detected pedestrian bounding boxes in a pair of frames. In this paper, we design two strategies to sample the frame pairs as follows (Figure 2).

- **Intra-sampling.** Frame pairs are sampled from the same video within a short temporal interval, *i.e.*, 2 seconds.
- **Inter-sampling.** Each frame pair is sampled at the same timestamp from two different cameras that capture an overlapped FOV.

In both strategies, with proper selection of temporal interval or deployment of cameras, which should not be too difficult to control, a reasonable identity overlap between $\boldsymbol{I}_1$ and $\boldsymbol{I}_2$ can be guaranteed. We define $\tau = \frac{\#\ overlapped\ IDs}{max\{|\boldsymbol{I}_1|,|\boldsymbol{I}_2|\}}$ as the *symmetry* between $\boldsymbol{I}_1$ and $\boldsymbol{I}_2$, and its impact on our system is investigated in Section 4.1. For ease of illustration, let us begin with the absolute symmetric case $\tau = 1$, *i.e.*, for any instance in $\boldsymbol{I}_1$ there is a correspondence in $\boldsymbol{I}_2$, and vice versa. Then we introduce how we deal with asymmetry in Section 3.3. The training procedure is presented in Figure 3.

3.2 Association Between Symmetric Pairs

Consider all the images in $\boldsymbol{I}_1 \cup \boldsymbol{I}_2$ forming a minibatch. Suppose the size of the two sets $|\boldsymbol{I}_1| = |\boldsymbol{I}_2| = K$ ($\tau = 1$, *i.e.,* absolute symmetry). The bounding boxes are mapped to the embedding space by Φ, such that $\boldsymbol{X}_1 = \Phi(\boldsymbol{I}_1)$ and $\boldsymbol{X}_2 = \Phi(\boldsymbol{I}_2)$,

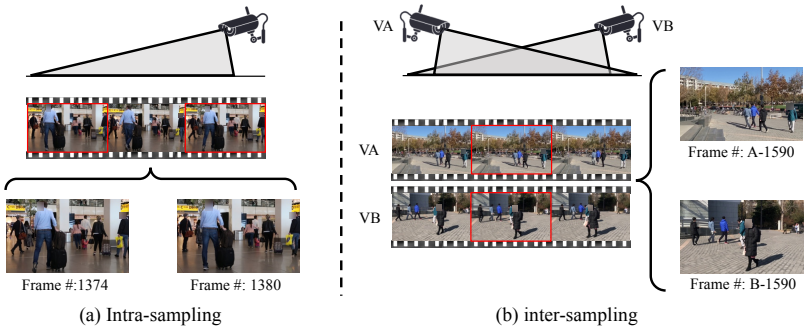


Fig. 2. Illustration on the two frame pair sampling methods. (a) Intra-sampling, frame pairs are drawn from a single video within a short temporal interval. (b) Inter-sampling, frame pairs are drawn from cameras capturing an overlapped FOV, at the same time.

where $\boldsymbol{X}_1 = [\boldsymbol{x}_1^1, \boldsymbol{x}_1^2, ..., \boldsymbol{x}_1^K] \in \mathbb{R}^{D\times K}$ and $\boldsymbol{X}_2 = [\boldsymbol{x}_2^1, \boldsymbol{x}_2^2, ..., \boldsymbol{x}_2^K] \in \mathbb{R}^{D\times K}$ are embedding matrices composed of K embedding vectors of dimension D. All the embedding vectors are ℓ_2-normalized. To capture similarity between instances, we compute an affinity matrix between all instances in $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ by calculating the pairwise cosine similarities,

$$\boldsymbol{S} = \boldsymbol{X}_1^\top \boldsymbol{X}_2 \in \mathbb{R}^{K\times K}. \tag{1}$$

We take $\boldsymbol{S}$ as input to perform data association, which aims to predict correspondences in $\boldsymbol{X}_2$ for each instance in $\boldsymbol{X}_1$. Formally, the goal is to obtain an assignment matrix,

$$\boldsymbol{A} = \boldsymbol{\psi}(\boldsymbol{S}) \in \{0,1\}^{K\times K}, \tag{2}$$

where 1 indicates correspondence. In MOT, it is usually modeled as a linear assignment problem, and the solution $\boldsymbol{\psi}$ can be found by the Hungarian algorithm (examples can be found in many MOT algorithms [32,38,30]).

Suppose the embedding function Φ is perfect, *i.e.*, the cosine similarity between vectors of the same identity equals 1, while the cosine similarity between vectors from different identities equals -1. The Hungarian algorithm can output the optimal assignment $\boldsymbol{A}^* = \frac{\boldsymbol{S}+1}{2}$ for the forward association process $\boldsymbol{X}_1 \rightarrow \boldsymbol{X}_2$. The backward association process $\boldsymbol{X}_2 \rightarrow \boldsymbol{X}_1$ is similar, and the optimal assignment matrix $\boldsymbol{A}'^* = \frac{\boldsymbol{S}^\top+1}{2}$. A Cycle Association pass is then defined as a forward association pass plus a backward association pass,

$$\boldsymbol{A}^{\texttt{cycle}} = \boldsymbol{A}\boldsymbol{A}'. \tag{3}$$

Intuitively, if $\boldsymbol{A} = \boldsymbol{A}^*$ and $\boldsymbol{A}' = \boldsymbol{A}'^*$, an instance will be associated to itself. In other words, the cycle association matrix underpinning perfect association $\boldsymbol{A}^{\texttt{cycle}}$ should equal the identity matrix $\boldsymbol{I}$. Accordingly, the difference between $\boldsymbol{A}^{\texttt{cycle}}$ and $\boldsymbol{I}$ can be used as signals to implicitly supervise the model to learn correspondences between $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$.

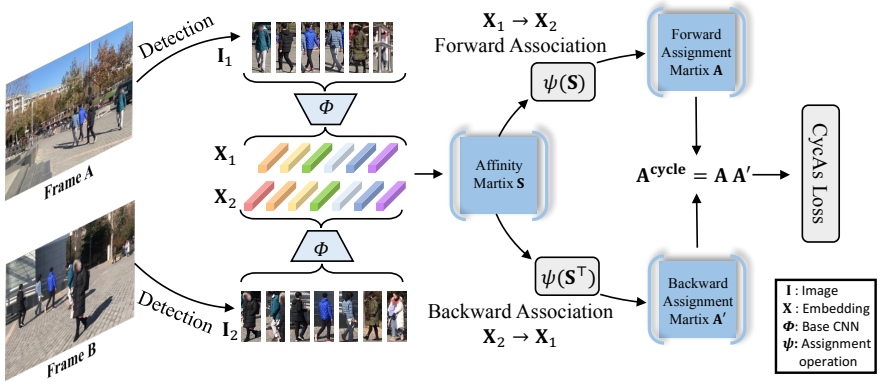


Fig. 3. An overview of the proposed cycle association task. First, two sets of detected pedestrians are mapped to embeddings via the base CNN Φ. Pairwise affinity matrix $\boldsymbol{S}$ is computed from the two sets of embeedings, then forward/backward assignment matrix $\boldsymbol{A}$ and $\boldsymbol{A'}$ if computed from $\boldsymbol{S}$. Finally the inconsistency in the cycle association matrix $\boldsymbol{A}^{\texttt{cycle}} = \boldsymbol{A}\boldsymbol{A'}$ is used as supervision signals.

The whole process needs to be differentiable for end-to-end training. However, the assignment operation ψ (Hungarian algorithm) is not differentiable. This motivates us to design a differentiable ψ. We notice that if the one-to-one correspondence constraint is removed, ψ can be approximated by the row-wise `argmax` function. Considering `argmax` is not differentiable either, we further soften this operation by the row-wise softmax function. Now, the assignment matrix is computed as,

$$\boldsymbol{A}_{i,j} = \psi_{i,j}(\boldsymbol{S}) = \frac{e^{T\boldsymbol{S}_{i,j}}}{\sum_{j'}^{K} e^{T\boldsymbol{S}_{i,j'}}}, \tag{4}$$

where $\boldsymbol{A}_{i,j}$ is the element of $\boldsymbol{A}$ in the i-th row and the j-th colomn, and T is the temperature of the softmax operation. The backward association pass has a different temperature T'. T and T' are designed to be adaptive to the size of $\boldsymbol{A}$ and $\boldsymbol{A'}$, and more details will be described in Section 3.4.

Combing Eq. 1, Eq. 4 and Eq. 3, a cycle association matrix $\boldsymbol{A}^{\texttt{cycle}}$ can be computed with all operations therein being differentiable. Finally, the loss function is defined as the mean ℓ_1 error between $\boldsymbol{A}^{\texttt{cycle}}$ and $\boldsymbol{I}$,

$$\mathcal{L}_{\texttt{symmetric}} = \frac{1}{K^2}\|\boldsymbol{A}^{\texttt{cycle}} - \boldsymbol{I}\|_1. \tag{5}$$

Discussion. Theoretically, cycle consistency is a *necessary* but not *sufficient* requirement for discriminative embeddings. In Figure 4 we present a trivial solution: what the embedding model has learned is matching an identity to the next identity, and the last identity is matched to the first. Such a trivial solution

Fig. 4. A non-trivial solution (**Left**) *v.s.* a trivial solution (**Right**) of Eq. 5. Solid and dotted lines means forward and backward correspondences, respectively. We argue the trivial solutions are very unlikely to be learned.

requires the model to learn strong correlations between random identity pairs, which share very limited, if any, similar visual patterns. Therefore, we reasonably argue that by optimizing the cycle association loss, it is very unlikely for the model to converge to such trivial solutions and that it is much easier to converge to non-trivial solutions, *i.e.*, the discriminative embeddings. Actually, this argument is proved by our experiment: all the converged solutions are disciminative embeddings, and trivial solutions like Figure 4 never emerge.

3.3 Relaxation on asymmetric pairs

In practice, asymmetric always arises along with large appearance diversity. The main reasons are multi-folds: First, pedestrians enter and leave the visual field of the camera; second, real-world videos usually contain high-velocity motions, severe occlusions, and low-quality persons, so the detector may fail sometimes; for inter-sampled data, it's also impossible to ensure FOVs overlap exactly. For better leveraging such appearance-diverse data, we make the following efforts to reduce the negative impact of the asymmetry.

First, consider descriptor matrices $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ with the number of descriptors being K_1 and K_2, respectively. According to Eq. 3, the resulting $\boldsymbol{A}^{\texttt{cycle}}$ is of size $K_1 \times K_1$. If $K_1 > K_2$, there will exist at least $K_1 - K_2$ instances that cannot be associated back to themselves. This will introduce ambiguity. Therefore, we swap $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ in such cases to ensure $K_1 \leq K_2$ always holds.

Second, the learning objective is modified. In the asymmetric scenario, the loss function $\mathcal{L}_{\texttt{symmetric}}$ is sub-optimal, because some instances may lose correspondences in cycle association, and thus the corresponding diagonal elements in $\boldsymbol{A}^{\texttt{cycle}}$ are not supposed to equal 1. Simply changing the supervision of these lost instances from 1 to 0 is not feasible, because there are no annotations and we do not know which instances are lost. To address this, our solution is to relax the learning objective. More specifically, we expect a diagonal element $\boldsymbol{A}^{\texttt{cycle}}_{i,i}$

to be greater than all the other elements along the same row and column, by a given margin m. The loss function is formulated as,

$$\mathcal{L}_{\texttt{asymmetric}} = \frac{1}{K_1}\sum_{i=1}^{K_1}\left[\left(\max_{j\neq i} \boldsymbol{A}^{\texttt{cycle}}_{i,j} - \boldsymbol{A}^{\texttt{cycle}}_{i,i} + m\right)_+ + \left(\max_{k\neq i} \boldsymbol{A}^{\texttt{cycle}}_{k,i} - \boldsymbol{A}^{\texttt{cycle}}_{i,i} + m\right)_+\right], \quad (6)$$

which has a similar form as the triplet loss [22]. The margin m is a hyper-parameter ranging in $(0, 1)$ with smaller values indicating softer constrains. We set $m = 0.5$ in all the experiment if not specified.

We will show through experiment that the relaxation of the loss function benefits learning in both the asymmetric and symmetric cases. In the experiment, we use $\mathcal{L}_{\texttt{asymmetric}}$ by default unless specified.

3.4 Adapt softmax temperature to varying sizes

Cosider two vectors with different sizes, $\boldsymbol{v} = (1, 0.5)^\top$ and $\boldsymbol{u} = (1, 0.5, 0.5)^\top$. Let $\boldsymbol{\sigma}$ be the softmax operation, then $\boldsymbol{\sigma}(\boldsymbol{v}) = (0.62, 0.38)^\top$ and $\boldsymbol{\sigma}(\boldsymbol{u}) = (0.45, 0.27, 0.27)^\top$. The *Soft-Max* operation, as we observe, has different levels of softening ability on inputs with different sizes. The Max value in a longer vector is less highlighted, or maxed, and vice versa. To alleviate this problem and stabelize the training, we let the softmax temperature be adaptive to the varying input size, so that for input vectors of different sizes, the max values in them are equally highlighted.

To fulfill this goal, we let the temprature $T = \frac{1}{\epsilon}\log\left[\frac{\delta(K-1)+1}{1-\delta}\right]$, where ϵ and δ are two hyper-parameters ranging from 0 to 1. In fact, the only hyper-parameter that matters is ϵ, and δ can be simply set to 0.5. This leads to the final form $T = \frac{1}{\epsilon}\log(K+1)$. Detailed derivation and discussion can be found in the supplementary material.

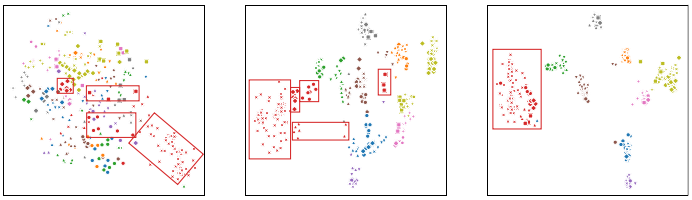
3.5 Two-stage Training

The two types of sampling strategies have their respective advantages and drawbacks. To get the best of both worlds, we design a two-stage training procedure that initializes from a model pretrained on ImageNet [2].

In Stage I, we train the model with intra-sampled data only, and the temporal interval for sampling is set to rather small, *e.g.*, 2 seconds, so that the appearance variation between two person sets is small, or the symmetric τ is high (> 0.9).

After convergence of Stage I, we start Stage II. In this stage, we train the model using both the inter-sampled data and intra-sampled data in a multi-task manner with $1:1$ loss weights. The inter-sampled data have much higher appearance variations but a lower τ (around 0.6).

Discussions. This training strategy is carefully designed so as to converge well. Directly starting from Stage II converges with a slower speed *w.r.t.* our progressive training strategy, while starting from inter-sampled data only fails in converging. We will quantitatively demonstrate the effectiveness of this training strategy in Section 4.2.



(a) Initial embedding space (b) Stage I convergence (c) Stage II convergence

Fig. 5. Illustration on the evolution of the embedding space from (a) Initial model to (b) Stage I convergence then to (c) Stage II convergence. Different colors indicate different identities. Different markers indicate different cameras. Visualized via Barnes-Hut t-SNE [24].

Alternatively, we give an intuitive illustration by visualizations on the embedding space at each training state, shown in Figure 5. In the initial embedding space before training, , most embeddings from different identities are not separable Figure 5 (a). In Stage I training, the model learns to find correspondence between persons within the same camera, so in the resulting embedding space embeddings of the same identity from the same cameras are grouped together, while embeddings of the same identity from different cameras are still separate (see red rectangles in Figure 5 (b)). Finally, in Stage II training, the model learns to associate across different cameras, in which case the appearance variations are large. Therefore, the resulting embedding can handle large appearance variation, thus is camera-invariant(Figure 5 (c)). To summarize, training Stage I functions as a "warm-up" process for Stage II, while in Stage II the model learns meaningful camera-invariant representation for the re-ID task.

4 Experiment

The experiments are organized as follows. First, we adopt standard train / test protocols on seven video- / image-based person re-ID datasets and compare our method with existing unsupervised methods (Section 4.1). Then we investigate the impact of different components and hyper-parameters (Section 4.2). Finally, to be more practical, we perform experiment using self-collected pedestrian videos as training data and compare with direct transfer (supervised) models (Section 4.3). We use ResNet-50 [8] as the backbone network in all experiments.

4.1 Experiment with Standard Datasets.

Setup. We test the proposed CycAs on both video-based (MARS [41], iLIDS [27], PRID2011 [9]) and image-based (Market-1501 [42], CUHK03 [15], DukeMTMC-ReID [21], MSMT17 [31]) person re-id datasets. All the datasets provide camera annotations and the video-based datasets additionally provide tracklets. Following existing practice [13,14,34], for image-based datasets, we assume all images

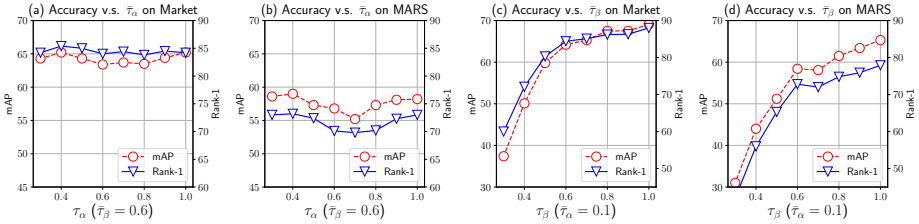


Fig. 6. Robustness against different levels of (a-b) intra-sampled data symmetry τ_α and (c-d) inter-sampled data symmetry τ_β. For evaluating τ_α, in each mini-batch, we fix τ_β and draw a random τ_α from $\mathcal{N}(\bar{\tau}_\alpha, 0.01)$, and plot the performance curve *w.r.t.* different mean $\bar{\tau}_\alpha$. The impact of τ_β is evaluated in a similar way.

per ID per camera are drawn from a single tracklet. Consider a mini-batch with batch size B, to mimic the intra- / inter-sampling, we first randomly sample $B/2$ identities. For intra-sampling, a tracklet is sampled for each of these identities; then, two bounding boxes are sampled within each tracklet. For inter-sampling, two tracklets from different cameras are sampled for each of the $B/2$ identities; then, one image is is sampled from each tracklet. Results on image- / video-based re-ID datasets are shown in Table 1 and Table 2, respectively.

Performance upper bound analysis. According to above sampling strategy, the data are absolutely symmetric, *i.e.*, $\tau = 1$. The performance under this setting can be seen as an upper bound of the proposed method, denoted as CycAs$^{\texttt{sym}}$. For comparison, we implement a supervised baseline (IDE [42]). We observe that the performance of CycAs$^{\texttt{sym}}$ is consistently competitive on all the datasets. Compared with IDE, the rank-1 accuracy of CycAs$^{\texttt{sym}}$ is lower by only 1.1%, 0.3% and 1.6% on Market-1501, DukeMTMC-ReID and MSMT17, respectively. On CUHK03, CycAs$^{\texttt{sym}}$ even surpasses IDE by +2.2%. This result partially prove the good alignment between the CycAs task and the objective of re-ID. We also list state-of-the-art supervised methods [23,12] for comparison and observe that the performance gap between CycAs$^{\texttt{sym}}$ and these methods is not too large. These results suggest the potential of CycAs in the re-ID task.

Robustness against different levels of data symmetry τ. To investigate the robustness of CycAs against different levels of symmetry, we introduce asymmetry to the sampled data and observe how the ReID accuracy changes. Specifically, we control intra-sampling symmmetry τ_α and inter-sampling asymmetry τ_β by replacing a portion of images in $\boldsymbol{I}_2$ with randomly sampled images from irrelevant identities. To evaluate the impact of τ_α, in each mini-batch, we fix τ_β and draw τ_α from a gaussian distribution $\mathcal{N}(\bar{\tau}_\alpha, 0.01)$, truncate the value in range $(0, 1)$, and plot the model performance against the mean $\bar{\tau}_\alpha$. The impact of τ_β is evaluated in a similar way. We report results in Fig. 6.

Two observations can be made from the curves. First, with a moderate fixed value of τ_β, *i.e.*, 0.6 in our case, the model accuracy is robust to a wide range of $\bar{\tau}_\alpha$. For example, in Fig. 6 (a), the rank-1 accuracy is both 84.2% when $\bar{\tau}_\alpha$ is set

Method	Category	Require	Market [42]		Duke [21]		CUHK03 [15]		MSMT17 [31]	
			R1	mAP	R1	mAP	R1	mAP	R1	mAP
SPGAN [3]	UDA	Pretrain	51.5	22.8	41.1	22.3	-	-	-	-
SPGAN+LMP [3]	UDA	Pretrain	57.7	26.7	46.4	26.2	-	-	-	-
TJ-AIDL [25]	UDA	Pretrain	58.2	26.5	44.3	23.0	-	-	-	-
HHL [43]	UDA	Pretrain	62.2	31.4	46.9	27.2	-	-	-	-
ECN [44]	UDA	Pretrain	75.1	43.0	63.3	40.4	-	-	30.2	10.2
PUL [5]	Clustering	Pretrain	44.7	20.1	30.4	16.4	-	-	-	-
CAMEL [39]	Clustering	Pretrain	54.5	26.3	-	-	39.4	-	-	-
BUC [16]	Clustering	None	66.2	38.3	47.4	27.5	-	-	-	-
CDS [33]	Clustering	Pretrain	71.6	39.9	67.2	42.7	-	-	-	-
TAUDL [13]	Tracklet	MOT	63.7	41.2	61.7	43.5	44.7	31.2	28.4	12.5
UTAL [14]	Tracklet	MOT	69.2	46.2	62.3	44.6	**56.3**	**42.3**	31.4	13.1
UGA [34]	Tracklet	MOT	**87.2**	**70.3**	*75.0*	*53.3*	-	-	*49.5*	*21.7*
CycAs$^{\texttt{asy}}$	Self-Sup	None	*84.8*	*64.8*	**77.9**	**60.1**	*47.4*	*41.0*	**50.1**	**26.7**
CycAs$^{\texttt{sym}}$	-	None	88.1	71.8	79.7	62.7	56.4	49.6	61.8	36.2
IDE	Supervised	Label	89.2	73.9	80.0	63.1	54.2	47.2	60.2	33.4
PCB+RPP [23]	Supervised	Label	93.8	81.6	83.3	69.2	63.7	57.5	-	-

Table 1. Comparison with state-of-the-art methods on image-based re-ID datasets. Note all the methods starts from a ImageNet pretrained model. The requirement *Pretain* and *None* refers to whether pretraining on labeled re-ID datasets is needed. CycAs$^{\texttt{asy}}$ is our method, and CycAs$^{\texttt{sym}}$ refers to an upper bound of our method.

Method	Category	Require	MARS [41]		PRID [9]	iLIDS [15]
			R1	mAP	R1	R1
DGM+IDE [36]	Clustering	Pretrain	36.8	21.3	56.4	36.2
RACE [35]	Clustering	MOT	43.2	24.5	50.6	19.3
BUC [16]	Clustering	None	61.1	38.0	-	-
SMP [18]	Tracklet	MOT	23.9	10.5	80.9	41.7
TAUDL [13]	Tracklet	MOT	43.8	29.1	49.4	26.7
UTAL [14]	Tracklet	MOT	49.9	35.2	54.7	35.1
UGA [34]	Tracklet	MOT	58.1	39.3	80.9	57.3
CycAs$^{\texttt{asy}}$	Self-Sup	None	**72.8**	**58.4**	**86.5**	**73.3**
CycAs$^{\texttt{sym}}$	-	None	79.2	67.5	85.4	77.3
IDE	Supervised	Label	81.7	67.8	90.5	78.4
GLTR [12]	Supervised	Label	87.0	78.5	95.5	86.0

Table 2. Comparison with state-of-the-art methods on video-based re-ID datasets.

to 1 and 0.3, respectively. Second, we observe that the accuracy improves when $\bar{\tau}_\beta$ becomes larger. The main reason is explained in Section 3.5 and Figrure 5: Training Stage I (Training with intra-sampled data) only functions as a "warm-up" process , to provide a meaningful initial point for Stage II. The knowledge learned from intra-sampled data contributes less on the overall performance. Therefore the final accuracy is less sensitive to $\bar{\tau}_\alpha$. In contrast, learning from inter-sampled data aligns with the objective of re-ID task, therefore the final accuracy is more sensitive to $\bar{\tau}_\beta$.

Remarks. Note that in Fig. 6 (c-d), the curves drop very slowly when $\tau_\beta = 1$ decreases from 1 to 0.6. This suggests that CycAs has a good ability to handle data with reasonably asymmetry. Such a property is valuable, because in practice we can control $\bar{\tau}_\beta$ in a reasonable range (say from 0.6 to 0.9), by carefully placing the cameras. Comparing with manually annotating data, this requires less effort.

Comparison with the state of the art. For fair comparisons, we train CycAs under a practically reasonable asymmetric assumption. We fix $\tau_\alpha = 0.9$ and $\tau_\beta = 0.6$, and compare the results (denoted as CycAs$^{\texttt{asy}}$) with existing unsu-

Table 3. Comparison between two losses under different data symmetry settings on Market-1501. We see the asymmetric loss always outperforms the symmetric loss no matter with asymmetric or symmetric data.

Data	Loss	Rank-1	mAP
symmetric			
$\tau_\alpha, \tau_\beta =$	$\mathcal{L}_{\texttt{symmetric}}$	78.9	59.1
$(1.0, 1.0)$	$\mathcal{L}_{\texttt{asymmetric}}$	**88.1**	**71.8**
asymmetric			
$\tau_\alpha, \tau_\beta =$	$\mathcal{L}_{\texttt{symmetric}}$	67.1	47.1
$(0.9, 0.6)$	$\mathcal{L}_{\texttt{asymmetric}}$	**84.8**	**64.8**

Table 4. Impact of the intra-sampling temporal interval and different training strategy. Evaluated on Market-1501.

Training		R-1	mAP
Stage I Only	*Interval:*		
	2 sec.	**29.4**	**11.9**
	4 sec.	25.8	9.7
	8 sec.	27.2	10.4
Stage II Only	*Data:*		
	intra + inter	84.6	64.7
	inter	-	-
Stage I + Stage II	*Interval:*		
	2 sec.	**84.8**	**64.8**
	4 sec.	84.2	64.1
	8 sec.	84.0	53.3

pervised re-ID approaches. Three categories of existing methods are compared, *i.e.*, unsupervised domain adaptation (UDA) [3,25,43,44], clustering-based methods [5,39,16,33,36,35], and tracklet-based methods [13,14,34]. Beside re-ID accuracy, we also compare another dimension, *i.e.*, ease of use, by listing the requirements of each method in Table 1 and Table 2.

Under image-based unsupervised learning, CycAs$^{\texttt{asy}}$ achieves state-of-the-art accuracy on two larger datasets, *i.e.*, DukeMTMC and MSMT17. The mAP improvement over the second best method [34] is +6.8% and +5.0% on DukeMTMC and MSMT17, respectively. On Market-1501 and CUHK03, CycAs$^{\texttt{asy}}$ is very competitive to the best performing methods [34].

Under video-based unsupervised learning, CycAs$^{\texttt{asy}}$ achieves state-of-the-art results on three datasets. The rank-1 accuracy improvement over the second best method is +14.7%, +5.6% and +16.0% on MARS, PRID and iLIDS, respectively.

Comparing with other unsupervised strategies, CycAs requires less external supervision. For example, UDA methods use a labeld source re-ID dataset, and most clustering-based methods need a pre-trained model for initialization, which also uses external labeled re-ID datasets. The tracklet-based methods do not require re-ID labels, but require a good tracker to provide good supervision signals.Training such a good tracker also requires external pedestrian labels. Note that ImageNet pretraining is needed by all the methods.CycAs learns person representations directly from videos and does not require any external annotation. Its requiring less supervision and competitive accuracy making it potentially a more practical solution for unsupervised re-ID.

4.2 Ablation Study

$\mathcal{L}_{\texttt{symmetric}}$ ***v.s.*** $\mathcal{L}_{\texttt{asymmetric}}$**.** To prove the proposed relaxed loss $\mathcal{L}_{\texttt{asymmetric}}$ is superior to the original loss $\mathcal{L}_{\texttt{symmetric}}$, we compare the two losses with both symmetric and asymmetric training data. Results on Market dataset is shown

in Table 3. We observe that with both symmetric/asymmetric training data, $\mathcal{L}_{\texttt{asymmetric}}$ consistently outperforms $\mathcal{L}_{\texttt{symmetric}}$. These results reveal a misunderstanding that $\mathcal{L}_{\texttt{symmetric}}$ is better for symmetric data and $\mathcal{L}_{\texttt{asymmetric}}$ is better for asymmetric data. In contrast, the results suggest $\mathcal{L}_{\texttt{asymmetric}}$ is essentially a more reasonable objective for the CycAs task, and it results in a more desirable embedding space. We speculate $\mathcal{L}_{\texttt{symmetric}}$ is too rigorous for the task, and the large magnitudes of losses from those well-associated cycles hamper the training.

Impact of the temporal interval for intra-sampling. We investigate the impact of the temporal interval for intra-sampling and show comparisons in Table 4. When only intra-sampled data are used for training (Stage I only), the sampling interval slightly affects the final performance. However, longer temporal interval, *e.g.*, 8 seconds, does not bring performance gain.We speculate the reason could be the large asymmetry brought by long temporal interval .

Two-stage training strategy. The effectiveness of the proposed two-stage training strategy is also investigated, and results are shown in Tabel 4. For comparison, we first train with Stage I only, *i.e.*, only intra-sampled data are used. It can be observed the final accuracy is quite poor. As discussed in previous section, the reason is training with intra-sampled data does not align with the objective of the re-ID task, *i.e.*, cross-camera retrieval. We also compare with training with Stage II only. If both intra- and inter-sampled data are used, the model converges to a decent solution. However, if we remove intra-sampled data from training, the model fails in converging. This suggest training with intra-sampled data is necessary for converging to a meaningful solution. Finally, if we use the proposed two-stage training, the results are as good as training with Stage II only, and the benefit is that the model converges with a faster speed.

4.3 Experiment Using Self-collected Videos as Training Data

To our knowledge, prior works in unsupervised re-ID usually evaluate their systems on standard benchmarks by simulating real-world scenarios. In this paper, to further assess the practical potential of CycAs, we report experimental results obtained by training with real-world videos. The videos are captured by hand-hold cameras in several scenes with high pedestrian density, such as the airport, shopping mall and campus. The total length of the videos is about 6 hours. Among these videos, about 5 hours are captured from a single view, which can only be used for intra-sampling; The rest 1 hour videos are captured from two different views, and thus can be used for inter-sampling.

We employ an open-source pedestrian detector [30] to detect persons in every 7 frames and crop the detected persons. For intra-sampling, we set the maximum temporal interval between two frames to 2 seconds. In every mini-batch, we sample 8 frame pairs to enlarge the batch size for high training efficiency. Training lasts for 10k iterations for Stage I and another 35k iterations for Stage II.

We evaluate the performance on Market-1501 [42] and DukeMTMC-ReID [21] test sets. Note that we do not use the training sets of Market-1501 and DukeMTMC-ReID. Since we use self-collected videos that are under completely different environments from Market-1501 and Duke-ReID, there is a large domain gap between

	Training data	Market-1501		DukeMTMC	
		mAP	R1	mAP	R1
Supervised.M	Market-1501 [42]	73.9	89.2	*16.6*	*33.4*
Supervised.D	DukeMTMC [21]	*21.4*	*48.1*	63.1	80.0
Supervised.C	CUHK03 [15]	19.8	43.2	13.1	26.3
BUC [16]	6-hour unlabelled videos	14.2	29.8	11.2	21.5
UGA [34]	6-hour unlabelled videos	17.8	37.2	15.4	25.6
CycAs	6-hour unlabelled videos	**23.3**	**50.8**	**19.2**	**34.6**

Table 5. Results with real-world videos as training data. Comparisons are made with supervised baselines and existing unsupervised methods.

our training data (self-collected video) and the test data. Results are presented in Table 5. We make two observations.

First, our method is significantly superior to unsupervised methods BUC [16] and UGA [34]. Both models are trained on our self-collected videos for fair comparison. For BUC we use the public code; For UGA we use our own implementation, and employ the JDE [30] tracker to generate the tracklets. CycAs outperforms UGA and BUC by +13.6%, and +21.0% in rank-1 accuracy on Market. It shows the promising potential of CycAs in real-world applications.

Second, our method is very competitive or slightly superior to supervised models. For example, when trained on Market-1501 and tested on DukeMTMC, the IDE model obtains an mAP of 16.6%. In comparison, we achieve 19.2% mAP on DukeMTMC, which is +2.6% higher. Similarly, our test performance on Market-1501 is 1.9% higher than IDE trained on DukeMTMC. Moreover, our results on Market-1501 and DukeMTMC are consistently higher than IDE trained on CUHK03. In this experiment, the strength of CycAs lies in two aspects: 1) We utilize more training data. 2) We utilize unlabeled data in a more effective way. We also note that IDE is significantly higher than our method when IDE is trained and tested on the same domain. We reasonably think that if our system can be trained in a similar environment to Market-1501 or DukeMTMC (with properly deployed cameras), we would have a much better accuracy with a smaller domain gap.

5 Conclusion

This paper presents CycAs, a self-supervised person ReID approach. We carefully design the pretext task—cycle association—to closely match the objective of re-ID. Objective function relaxations are made to allow end-to-end learning and introduce higher appearance variations in the training data. We train CycAs using public data and self-collected videos, and both settings validate the competitive performance of CycAs.

References

1. Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: ECCV (2018)
2. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)
3. Deng, W., Zheng, L., Ye, Q., Kang, G., Yang, Y., Jiao, J.: Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In: CVPR (2018)
4. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: ICCV (2015)
5. Fan, H., Zheng, L., Yan, C., Yang, Y.: Unsupervised person re-identification: Clustering and fine-tuning. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) **14**(4), 83 (2018)
6. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. In: ICLR (2018)
7. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: ICCV (2019)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
9. Hirzer, M., Beleznai, C., Roth, P.M., Bischof, H.: Person re-identification by descriptive and discriminative classification. In: Scandinavian conference on Image analysis. pp. 91–102. Springer (2011)
10. Lai, Z., Xie, W.: Self-supervised learning for video correspondence flow. In: BMVC (2019)
11. Leal-Taixé, L., Milan, A., Reid, I., Roth, S., Schindler, K.: Motchallenge 2015: Towards a benchmark for multi-target tracking. arXiv preprint arXiv:1504.01942 (2015)
12. Li, J., Wang, J., Tian, Q., Gao, W., Zhang, S.: Global-local temporal representations for video person re-identification. In: ICCV (2019)
13. Li, M., Zhu, X., Gong, S.: Unsupervised person re-identification by deep learning tracklet association. In: ECCV (2018)
14. Li, M., Zhu, X., Gong, S.: Unsupervised tracklet person re-identification. IEEE transactions on pattern analysis and machine intelligence (2019)
15. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: Deep filter pairing neural network for person re-identification. In: CVPR (2014)
16. Lin, Y., Dong, X., Zheng, L., Yan, Y., Yang, Y.: A bottom-up clustering approach to unsupervised person re-identification. In: AAAI (2019)
17. Liu, P., Lyu, M., King, I., Xu, J.: Selflow: Self-supervised learning of optical flow. In: CVPR (2019)
18. Liu, Z., Wang, D., Lu, H.: Stepwise metric promotion for unsupervised video person re-identification. In: ICCV (2017)
19. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: ECCV (2016)
20. Pillai, S., Ambruş, R., Gaidon, A.: Superdepth: Self-supervised, super-resolved monocular depth estimation. In: ICRA (2019)
21. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: ECCV (2016)
22. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: CVPR (2015)

23. 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)
24. Van Der Maaten, L.: Accelerating t-sne using tree-based algorithms. The Journal of Machine Learning Research **15**(1), 3221–3245 (2014)
25. Wang, J., Zhu, X., Gong, S., Li, W.: Transferable joint attribute-identity deep learning for unsupervised person re-identification. In: CVPR (2018)
26. Wang, N., Song, Y., Ma, C., Zhou, W., Liu, W., Li, H.: Unsupervised deep tracking. In: CVPR (2019)
27. Wang, T., Gong, S., Zhu, X., Wang, S.: Person re-identification by video ranking. In: ECCV (2014)
28. Wang, X., He, K., Gupta, A.: Transitive invariance for self-supervised visual representation learning. In: ICCV (2017)
29. Wang, X., Jabri, A., Efros, A.A.: Learning correspondence from the cycle-consistency of time. In: CVPR (2019)
30. Wang, Z., Zheng, L., Liu, Y., Wang, S.: Towards real-time multi-object tracking. arXiv preprint arXiv:1909.12605 (2019)
31. Wei, L., Zhang, S., Gao, W., Tian, Q.: Person transfer gan to bridge domain gap for person re-identification. In: CVPR (2018)
32. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: ICIP (2017)
33. Wu, J., Liao, S., Wang, X., Yang, Y., Li, S.Z., et al.: Clustering and dynamic sampling based unsupervised domain adaptation for person re-identification. In: ICME (2019)
34. Wu, J., Yang, Y., Liu, H., Liao, S., Lei, Z., Li, S.Z.: Unsupervised graph association for person re-identification. In: ICCV (2019)
35. Ye, M., Lan, X., Yuen, P.C.: Robust anchor embedding for unsupervised video person re-identification in the wild. In: ECCV (2018)
36. Ye, M., Ma, A.J., Zheng, L., Li, J., Yuen, P.C.: Dynamic label graph matching for unsupervised video re-identification. In: ICCV (2017)
37. Ye, M., Johns, E., Handa, A., Zhang, L., Pratt, P., Yang, G.Z.: Self-supervised siamese learning on stereo image pairs for depth estimation in robotic surgery. arXiv preprint arXiv:1705.08260 (2017)
38. Yu, F., Li, W., Li, Q., Liu, Y., Shi, X., Yan, J.: Poi: Multiple object tracking with high performance detection and appearance feature. In: ECCV workshop (2016)
39. Yu, H.X., Wu, A., Zheng, W.S.: Cross-view asymmetric metric learning for unsupervised person re-identification. In: ICCV. pp. 994–1002 (2017)
40. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: ECCV (2016)
41. 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)
42. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: CVPR (2015)
43. Zhong, Z., Zheng, L., Li, S., Yang, Y.: Generalizing a person retrieval model hetero- and homogeneously. In: ECCV (2018)
44. Zhong, Z., Zheng, L., Luo, Z., Li, S., Yang, Y.: Invariance matters: Exemplar memory for domain adaptive person re-identification. In: CVPR (2019)
45. Zhu, A.Z., Yuan, L., Chaney, K., Daniilidis, K.: Ev-flownet: Self-supervised optical flow estimation for event-based cameras. arXiv preprint arXiv:1802.06898 (2018)