Features for Multi-Target Multi-Camera Tracking and Re-Identification

Ergys Ristani Carlo Tomasi
Duke University
Durham, NC, USA
{ristani, tomasi}@cs.duke.edu

Abstract

Multi-Target Multi-Camera Tracking (MTMCT) tracks many people through video taken from several cameras. Person Re-Identification (Re-ID) retrieves from a gallery images of people similar to a person query image. We learn good features for both MTMCT and Re-ID with a convolutional neural network. Our contributions include an adaptive weighted triplet loss for training and a new technique for hard-identity mining. Our method outperforms the state of the art both on the DukeMTMC benchmarks for tracking, and on the Market-1501 and DukeMTMC-ReID benchmarks for Re-ID. We examine the correlation between good Re-ID and good MTMCT scores, and perform ablation studies to elucidate the contributions of the main components of our system. Code is available[1].

1. Introduction

Multi-Target Multi-Camera Tracking (MTMCT) aims to determine the position of every person at all times from video streams taken by multiple cameras. The resulting multi-camera trajectories enable applications including visual surveillance, suspicious activity and anomaly detection, sport player tracking, and crowd behavior analysis. In recent years, the number of cameras has increased dramatically in airports, train stations, and shopping centers, so it has become necessary to automate MTMC tracking.

MTMCT is a notoriously difficult problem: Cameras are often placed far apart to reduce costs, and their fields of view do not always overlap. This results in extended periods of occlusion and large changes of viewpoint and illumination across different fields of view. In addition, the number of people is typically not known in advance, and the amount of data to process is enormous.

Person re-identification (Re-ID) is closely related to MTMCT: Given a snapshot of a person (the query), a Re-ID system retrieves from a database a list of other snapshots of people, usually taken from different cameras and at different times, and ranks them by decreasing similarity to the query. The intent is that any snapshots in the database that are *co-identical* with (that is, depict the same person as) the person in the query are ranked highly.

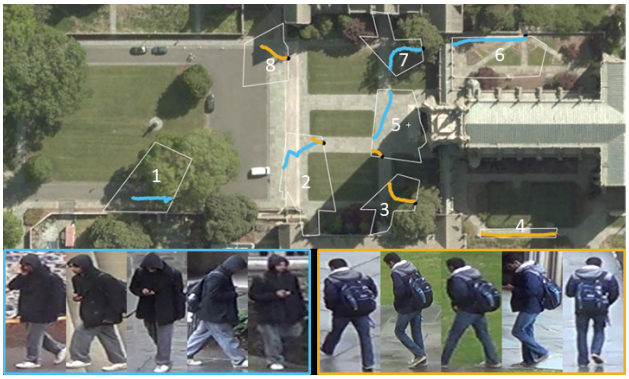

Figure 1. Two example multi-camera results from our tracker on the DukeMTMC dataset.

MTMCT and Re-ID differ subtly but fundamentally, because Re-ID *ranks* distances to a query while MTMCT *classifies* a pair of images as being co-identical or not, and their performance is consequently measured by different metrics: ranking performance for Re-ID, classification error rates for MTMCT. This difference would seem to suggest that appearance features used for the two problems must be learned with different loss functions. Ideally, the Re-ID loss ought to ensure that *for any query* a the largest distance between a and a feature that is co-identical to it is smaller than the smallest distance between a and a feature that is not co-identical to it. This would guarantee correct feature ranking for any given query. In contrast, the MTMCT loss ought to ensure that the largest distance between *any two* co-identical features is smaller that the smallest distance between *any two* non co-identical features, to guarantee a margin between within-identity and between-identity distances.

With these criteria, zero MTMCT loss would imply zero Re-ID loss, but not *vice versa*. However, training with a loss

This material is based upon work supported by the National Science Foundation under Grants No. IIS-1420894 and CCF-1513816.

[1] http://vision.cs.duke.edu/DukeMTMC/

of the MTMCT type is very expensive, because it would require using all pairs of features as input. More importantly, there would be a severe imbalance between the number of within-identity pairs and the much greater number of between-identity pairs. In this paper, we couple a triplet loss function of the Re-ID type with a training procedure based on hard-data mining and obtain high-performing features for both Re-ID and MTMCT. Our experiments also show that when tracking moderately crowded scenes, improving Re-ID rank accuracy beyond a certain point shows diminishing returns for MTMCT.

To use our features for MTMCT, we assemble a processing pipeline (Figure 2) that uses a state-of-the-art person detector and, at test time, a state-of-the art data association algorithm based on correlation clustering to group observations into identities. To reduce computational complexity, we also incorporate standard hierarchical reasoning and sliding temporal window techniques in our tracker. Some qualitative results from our method are shown in Figure 1.

We do *not* include correlation clustering when training. Instead, we make the conjecture that high-quality appearance features lead to good clustering solutions, and only train the features. Our state-of-the art experimental results on the DukeMTMCT benchmark bear out this conjecture.

In summary we make the following contributions:

- We propose an adaptive weighted triplet loss that, unlike fixed-weight variants, is both accurate and stable.
- We propose an inexpensive hard-identity mining scheme that helps learn better features.
- We provide new insights on the relation between tracking and ranking accuracy on existing benchmarks.
- We show experimentally that our features yield state-of-the-art results on both MTMCT and Re-ID tasks.

2. Related Work

We summarize work on different aspects of MTMCT.

Person Detection. MTMC trackers rely on person detection and some trackers assume that single-camera trajectories are given [11, 14, 19, 20, 21, 26, 27, 32, 38, 42, 48, 77]. The popular Deformable Parts Model detector [30] was used as the public detector for MOTChallenge sequences [43, 51, 57] and in labeling Re-ID datasets [63, 81]. Since the MOT17 challenge, trackers have shown increased accuracy by utilizing detectors that rely on deep learning. These include Faster R-CNN [56], SSD [47], KDNT [73], or pose-based detectors [17, 36]. We use OpenPose [17] which has shown good performance.

Data Association. Most existing formulations, with some exceptions [10, 52, 53], are special cases of the multidimensional assignment problem [25]: Input detections are arranged in a graph whose edges encode similarity and whose nodes are then partitioned into identities. Formulations with polynomial time solutions consider evidence along paths of time-consecutive edges [8, 16, 31, 37, 38, 39, 55, 75, 78] and some build on bipartite matching [12, 14, 20, 26, 42, 62, 71]. Methods that use all pairwise terms, not only time-consecutive ones, are significantly more accurate but NP-hard [18, 25, 27, 28, 41, 58, 61, 65, 66, 67]. Unary terms are sometimes added for completeness [28, 65]. Higher order terms have also been used [13, 70] but with sharply diminishing returns. Identities can be optimized jointly [28] or iteratively [74]. We choose correlation clustering [4, 57, 58] to trade off computational cost for simplicity of formulation and accuracy. This formulation considers evidence from all pairwise terms and optimizes identities jointly. An equivalent formulation is that of graph multicuts [66] which minimizes disagreement instead of maximizing agreement [29].

Appearance. Human appearance has been described by color [14, 19, 20, 21, 27, 32, 38, 39, 42, 77, 78] and texture descriptors [14, 20, 26, 42, 77, 78]. Lighting variations are addressed through color normalization [14], exemplar-based approaches [20], or brightness transfer functions learned with [27, 38] or without supervision [19, 32, 77, 78]. Discriminative power is improved by *saliency* information [50, 80] or by *learning* features specific to body parts [14, 20, 21, 26, 27, 39, 42], either in the image [6, 7, 24] or back-projected onto an articulated [2, 23] or monolithic [3] 3D body model. The current state of the art in person re-identification relies on deep learning [82, 84], hard negative mining [84], data augmentation [5, 85], special purpose layers [64] or branches [83], and specialized loss functions [35]. We use a residual network [34] and similar techniques to learn good features for MTMCT and Re-ID.

Multiple Cameras. Spatial relations between cameras are either explicitly mapped in 3D [19, 77], learned by tracking known identities [15, 38, 39], or obtained by comparing entry/exit rates across pairs of cameras [14, 42, 48]. Preprocessing methods may fuse data from partially overlapping views [78], while some systems rely on completely overlapping and unobstructed views [1, 8, 11, 33, 40]. People *entry and exit points* may be explicitly modeled on the ground [14, 19, 42, 48] or image plane [32, 39]. *Travel time* is also modeled, either parametrically [39, 77] or not [19, 32, 38, 42, 48]. We use time constraints to rule out unlikely inter-camera associations. Similarly to [57] we decay correlations to zero as the time distance between observations increases. Correlation decay ensures that time-distant observations are associated if there is a chain of positively-correlated observations that connect them. The idea is similar to lifted multicuts [67], although we employ no threshold or hard constraints.

Learning to Track. There have been several attempts to learn multi-target tracking data association in a supervised way, either through recurrent neural networks for end-to-

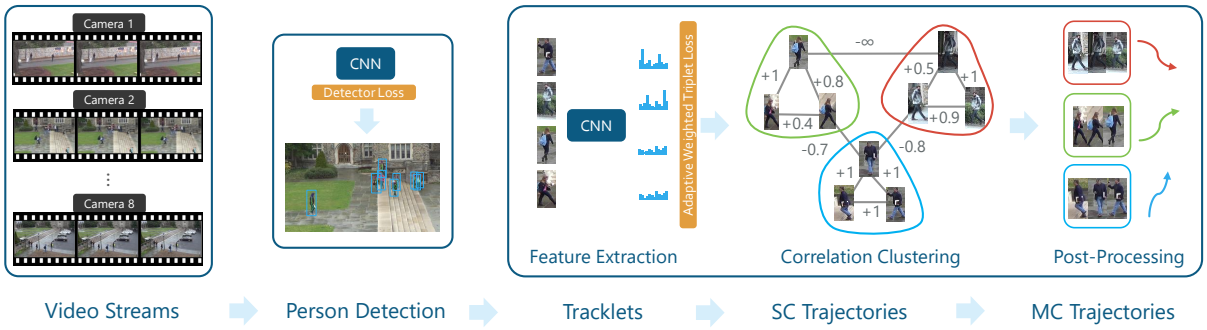

Figure 2. An illustration of our pipeline for Multi-Target Multi-Camera Tracking. Given video streams, a person detector extracts bounding box observations from video. For trajectory inference, a feature extractor extracts motion and appearance features from observations. These are in turn converted into correlations and labeled using correlation clustering optimization. Finally, post-processing interpolates missing detections and discards low confidence tracks. Multi-stage reasoning repeats trajectory inference for tracklets, single- and multi-camera trajectories. At train time the detector is trained independently, and the feature loss penalizes features that yield wrong correlations.

end prediction of trajectories [52] or by learning data association by back-propagating through a network-flow solution [60]. These methods have been pushing in the right direction even though they haven't yet topped single-camera tracking benchmarks. In our method we learn features for correlations without measuring trajectory quality through combinatorial optimization. Our argument is that if correlations are good, even greedy association suffices. This idea has been shown to work for person detection [17], and implicitly pursued in single-camera trackers [66, 67, 73] and Re-ID methods [35, 82, 83] that improve features to increase accuracy. Learning good correlations makes training simpler and less expensive, and we show that it achieves state-of-the-art performance for MTMCT.

3. Method

The input is a set of videos $V = \{V_1, \ldots, V_n\}$ from n different cameras, and the ground truth is a set of multi-camera trajectories $T = \{T_1, \ldots, T_\ell\}$. MTMCT could be cast as a supervised learning problem: Find the optimal parameters Θ^* of a function $f(\Theta, V)$ that estimates the true trajectories as well as possible:

$$\Theta^* = \arg\min_{\Theta} \mathcal{L}(f(\Theta, V), T) \tag{1}$$

where the loss function $\mathcal{L}$ could be derived from the multi-camera tracking accuracy measure IDF1 [57].

However, end-to-end training would require back-propagating the loss through a combinatorial optimization layer that performs data association, and this is expensive [60]. We avoid this complexity by noting that if the correlations were positive for co-identical pairs and negative for non co-identical pairs, then combinatorial optimization would be trivial. Thus, we aim to learn features that produce good correlations during training, while at test time we employ correlation clustering to maximize agreement between potentially erroneous correlations.

An additional source of difficulty during training is model depth, as weight updates can fail to propagate back to early layers responsible for person detection. If the network is monolithic and trained with a single loss, training becomes more difficult. We therefore separate detection and association as is customary in the literature (Figure 2). In the following we describe how we learn appearance features, and the different parts of the tracker.

3.1. Learning Appearance Features

Given a large collection of labeled person snapshots we learn appearance features using an adaptive weighted triplet loss. For an anchor sample x_a, positive samples $x_p \in P(a)$ and negative samples $x_n \in N(a)$, we re-write the triplet loss in its most general form as:

$$L_3 = \left[m + \sum_{x_p \in P(a)} w_p d(x_a, x_p) - \sum_{x_n \in N(a)} w_n d(x_a, x_n) \right]_+ \tag{2}$$

where m is the given inter-person separation margin, d denotes distance of appearance, and $[\cdot]_+ = max(0, \cdot)$. This reformulation has two advantages. First it avoids the combinatorial process of triplet generation by using all the samples rather than a selection. Instead, the challenge of learning good features is to assign larger weights to difficult positive and negative samples. Second, the positive/negative class imbalance is easily handled by reflecting it in the weight distribution.

Hermans *et al* [35] and Mischuk *et al* [54] have proposed the batch-hard triplet loss with built-in hard sample mining. The batch-hard loss weights for Equation 2 are binary in their approach, as the loss considers only the most difficult positive and negative sample:

$$w_p = \left[x_p == \arg\max_{x \in P(a)} d(x_a, x) \right] \tag{3}$$

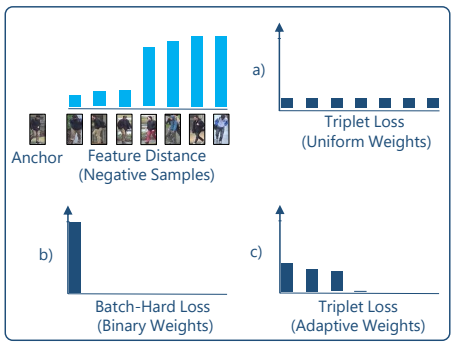

Figure 3. Triplet loss weighing schemes.

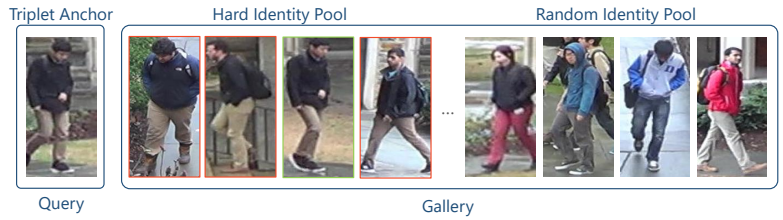

Figure 4. Hard identity mining: For each anchor identity, half of the $P-1$ identities in the batch are sampled from the hard identity pool, the other half from the random identity pool. Hard-negative identities (correct matches) are outlined in red (green).

$$w_n = \Big[x_n == \arg\min_{x \in N(a)} d(x_a, x)\Big] \tag{4}$$

where $[\cdot]$ denotes the Iverson bracket. This loss gives better results than the original triplet loss with uniform weights [59] because the latter washes out the contribution of hard samples and is driven to worse local minima by easy samples. On the other hand, the uniformly weighted loss is more robust to outliers because they cannot affect the weights.

Can we define weights such that L_3 converges to parameters at least as good as the batch-hard loss, yet remains robust to outliers? Our first improvement pertains to weights that achieve high accuracy and training stability *simultaneously*. Equations 3-4 assign full weight to the hardest positive/negative sample for each anchor while ignoring the remaining positive and negative samples. Instead, we assign adaptive weights using the softmax/min weight distributions as follows (see Figure 3):

$$w_p = \frac{e^{d(x_a,x_p)}}{\sum\limits_{x \in P(a)} e^{d(x_a,x)}}, \quad w_n = \frac{e^{-d(x_a,x_n)}}{\sum\limits_{x \in N(a)} e^{-d(x_a,x)}}. \tag{5}$$

The adaptive weights in Equation 5 give little importance to easy samples and emphasize the most difficult ones. When several difficult samples appear in a batch, they all get their fair share of the weight. This differs from the hard weight assignments of Equations 3-4 which give importance to the *single* most difficult sample. Adaptive weights are useful when the most difficult sample in a batch is an outlier, yet there exist other difficult samples to learn from. Experiments in such cases demonstrate the favorable properties of adaptive weights.

For batch construction during training we leverage the idea of PK batches also introduced by Hermans *et al* [35]. In each batch there are K sample images for each of P identities. This approach has shown very good performance in similarity-based ranking and avoids the need to generate a combinatorial number of triplets. During a training epoch each identity is selected in its batch in turn, and the remaining $P-1$ batch identities are sampled at random. K samples for each identity are then also selected at random.

Our second improvement is on the procedure that selects difficult identities. As the size of the training set increases, sampling $P-1$ identities at random rarely picks the hardest negatives, thereby moderating batch difficulty. This effect can also be observed in the last few epochs of training, when many triplets within a batch exhibit zero loss.

To increase the chances of seeing hard negatives, we construct two sets to sample identities from. An example is shown in Figure 4. The hard identity pool consists of the H most difficult identities given the anchor, and the random identity pool consists of the remaining identities. Then in a PK batch of an anchor identity we sample the remaining $P-1$ identities from the hard or random identity pool with equal probability. This technique samples hard negatives more frequently and yet the batch partially preserves dataset statistics by drawing random identities. The pools can be constructed either after training the network for few epochs, or computed from a pre-trained network. We demonstrate the benefit of using the hard-identity mining scheme in the experiments section.

3.2. MTMC Tracker

Given an array O_d of k-dimensional detections as input, the tracker outputs a $(k+1) \times o_t$ array $O_t = f_t(\Theta_t, O_d)$ of o_t detections. The added dimension is the identity label assignment to the input observations. In our design, the tracker first computes features for all o_d input observations, then estimates correlation between all pairs of features, and finally solves a correlation clustering problem to assign identities to observations. Two post-processing steps, interpolation and pruning, interpolate detections to fill gaps and remove trajectories with low confidence. For this reason, the number o_t of output detections can differ from the number o_d of input detections.

Detector. We use the off-the-shelf OpenPose person detector which achieves good performance [17]. This detector learns part affinity fields to capture the relation be-

tween body parts and applies greedy parsing to combine part affinities into bounding boxes. During training it is supervised directly on part affinities rather than bounding box accuracy.

Appearance Features. We use the ResNet50 model pretrained on ImageNet and follow its *pool5* layer by a dense layer with 1024 units, batch normalization, and ReLU. Another dense layer yields 128-dimensional appearance features. We train the model with the adaptive weighted triplet loss, data augmentation, and hard-identity mining.

We define the appearance correlation between two detections as $w_{ij} = \frac{t_a - d(x_i, x_j)}{t_a}$ (a number ≤ 1) where the threshold $t_a = \frac{1}{2}(\mu_p + \mu_n)$ separates the means of positive and negative distances μ_p and μ_n of all training pairs.

Data Augmentation. We augment the training images online with crops and horizontal flips to compensate for detector localization errors and to gain some degree of viewpoint/pose invariance. For illumination invariance we additionally apply contrast normalization, grayscale and color multiplication effects on the image. For resolution invariance we apply Gaussian blur of varying σ. For additional viewpoint/pose invariance we apply perspective transformations and small distortions. We additionally hide small rectangular image patches to simulate occlusion.

Motion Correlation. We use a linear motion model to predict motion correlation. As the forward-backward prediction error $e_m = e_f + e_b$ is non-negative, we use the trajectories from the training set to learn a threshold t_m that separates positive and negative evidence, and a scaling factor α to convert errors to correlations: $w_m = \alpha(t_m - e_m)$. Impossible associations receive correlation $w_m = -\infty$.

Optimization. A matrix $W = (W_a + W_m) \odot D$ collects appearance and motion correlations, and the matrix D specifies discounts $= e^{-\beta \Delta t} \in [0, 1]$ that decay correlations to zero as the time lag Δt between observation increases. D ensures association of time-distant trajectories only if there is a chain of associations with positive net correlation that connects them. Parameters t_m, α, β are chosen to maximize tracking accuracy over small subsets of the training set.

We establish co-identity by correlation clustering. Given a weighted graph $G = (V, E, W)$, two nodes v_i and v_j are co-identical if the binary incidence variable $x_{ij} = 1$ in the solution. Correlation clustering is defined as:

$$X^* = \arg\max_{\{x_{ij}\}} \sum_{(i,j) \in E} w_{ij} x_{ij} \tag{6}$$

$$\text{subject to:} \quad x_{ij} + x_{jk} \leq 1 + x_{ik} \quad \forall i, j, k \in V \tag{7}$$

Equation 6 maximizes positive (negative) correlation within (between) clusters and the constraints in Equation 7 enforce transitivity in the solution.

Multi-Level Reasoning. Our method reduces the computational burden by reasoning hierarchically over three levels. The first level computes one-second long tracklets, the second associates tracklets into single-camera trajectories, and the third associates single-camera trajectories into multi-camera identities.

Tracklets are found in disjoint, one-second long windows. Trajectories are computed online in a sliding temporal window that overlaps 50% with the previous window. All trajectories that have at least one detection in the window are re-considered for association. We set the window width for single-camera trajectories to 10 seconds, and 1.5 minutes for multi-camera trajectories.

4. Experiments

We run the following experiments on recent benchmarks for MTMCT and Re-ID: (a) Measure overall MTMCT performance, (b) measure the impact of improved detector and features during tracking, (c) study the relation between measures of accuracy for ranking and tracking, (d) demonstrate the usefulness of the adaptive weighted triplet loss and hard negative mining, and (e) analyze tracker failures.

4.1. Benchmarks

DukeMTMC [57] is a large-scale tracking dataset recorded on the Duke University campus featuring 2.8k identities, of which 1.8k belong to the training/validation set. The dataset was recorded by 8 cameras with 1080p 60fps image quality and the evaluation is done on disjoint fields of view. The video duration of each camera is 1 hour and 25 minutes. We benchmark our method on the 25 minute long *test-easy* sequence and 15 minute long *test-hard* sequence hosted on MOTChallenge [43]. *test-hard* features a group of 50 people traveling through 4 cameras. We use the 17 minute long validation sequence for ablation experiments.

DukeMTMC-reID [57, 84] is a subset of the DukeMTMC tracking dataset [57] for image-based person re-identification. It features 1,404 identities appearing in more than two cameras and 408 identities who appear in only one camera are used as distractors. 702 identities are reserved for training and 702 for testing.

Market-1501 [81] is a large-scale person re-identification dataset with 1,501 identities observed by 6 near-synchronized cameras. The dataset was collected in the campus of Tsinghua University. It features 32,668 bounding boxes obtained using the deformable parts model detector. The dataset is challenging as the boxes are often misaligned and viewpoints can differ significantly. 751 identities are reserved for training and the remaining 750 for testing.

4.2. Evaluation

For MTMCT evaluation we use ID measures of performance [57] which indicate how well a tracker identifies who is where regardless of where or why mistakes occur. IDP

	Multi-Camera Easy			Multi-Camera Hard			Single-Camera Easy				Single-Camera Hard			
	IDF1	IDP	IDR	IDF1	IDP	IDR	IDF1	IDP	IDR	MOTA	IDF1	IDP	IDR	MOTA
BIPCC [57]	56.2	67.0	48.4	47.3	59.6	39.2	70.1	83.6	60.4	59.4	64.5	81.2	53.5	54.6
lx_b [45]	58.0	72.6	48.2	48.3	60.6	40.2	70.3	88.1	58.5	61.3	64.2	80.4	53.4	53.6
PT_BIPCC [49]	-	-	-	-	-	-	71.2	84.8	61.4	59.3	65.0	81.8	54.0	54.4
MTMC_CDSC [68]	60.0	68.3	53.5	50.9	63.2	42.6	77.0	87.6	68.6	70.9	65.5	81.4	54.7	59.6
MYTRACKER [72]	64.8	70.8	59.8	47.3	55.6	41.2	80.0	87.5	73.8	77.7	63.4	74.5	55.2	59.0
MTMC_ReID [79]†	78.3	82.6	74.3	67.7	**78.6**	59.4	86.3	91.2	82.0	83.6	77.6	**90.1**	68.1	69.6
DeepCC	**82.0**	**84.3**	**79.8**	**68.5**	75.8	**62.4**	**89.2**	**91.7**	**86.7**	**87.5**	**79.0**	87.4	**72.0**	**70.0**

Table 1. DukeMTMCT results. Methods in † are unrefereed submissions.

(IDR) is the fraction of computed (true) detections that are correctly identified. IDF1 is the ratio of correctly identified detections over the average number of true and computed detections. IDF1 is used as the principal measure for ranking MTMC trackers. ID measures first compute a 1-1 mapping between true and computed identities that maximizes true positives, and then compute the ID scores.

For single-camera evaluation we also report MOTA, which counts mistakes by how often, not how long, incorrect decisions are made. MOTA is based on the CLEAR-MOT mapping [9] which under-reports multi-camera errors, therefore we report it only in single camera experiments.

For person re-identification experiments we report rank accuracy as well as mean average precision (mAP) [81].

4.3. Model Training

For training we set $P = 18$, $K = 4$, $m = 1$, resolution 256×128. The learning rate is $3 \cdot 10^{-4}$ for the first 15000 iterations, and decays to 10^{-7} at iteration 25000. In experiments with hard identity mining we construct the hard and random pools once with features obtained at iteration 5000, then sample identities from these pools until the last iteration. The hard identity pool size H is set to 50 and we found that similar scores were obtained with 30-100 identities (4%-15% of all training identities). Extreme sizes yield little gain: A size of 1 contains a single hard identity which can be an outlier, a large HN pool nears random sampling.

5. Results

We discuss results for MTMC tracking, where our proposed method outperforms previous and concurrent work in IDF1 score and identity recall IDR; study the influence of different components; and analyze common tracking failures. We also present results on person re-identification datasets, where our learned appearance features achieve competitive results.

5.1. Impact of Learning

We evaluate how detector and feature choice impact multi-camera IDF1 on the DukeMTMC validation set. Results are shown in Table 2.

	IDF1	IDP	IDR
BIPCC (DPM + HSV) [57]	54.98	62.67	48.97
DeepCC (OpenPose + HSV)	58.24	60.60	56.06
DeepCC (DPM + ResNet)	65.68	74.87	58.50
DeepCC (OpenPose + ResNet)	**80.26**	**83.50**	**77.25**

Table 2. Impact of improving detector and features on multi-camera performance for the validation sequence.

First we compare the behavior of our baseline method BIPCC with and without deep features. BIPCC uses part based color histograms as appearance features. Our learned features play an important role in improving IDF1 by 10.7 points (third row) in multi-camera performance.

Second we measure the impact of the deep learned detector. We substituted the baseline's DPM detections (first row) with those obtained from OpenPose [17] (second row). Although single-camera IDF1 increases from 75.0 to 85.5, multi-camera IDF1 increases by only 3.26 points (from 54.98 to 58.24%). This indicates that the detector plays an important role in single-camera tracking by reducing false negatives, but in multi-camera tracking weak features take little advantage of better single-camera trajectories.

These results imply that good features are crucial for MTMC tracking, and that a good detector is most useful for improving single-camera performance. The best MTMCT performance is achieved by combining both.

5.2. MTMC Tracking

Overall results are presented in Tables 1 and 3. Our method DeepCC improves the multi-camera IDF1 accuracy w.r.t to the previous state of the art MTMC_CDSC [68] by 22 and 17.6 points for the *test-easy* and *test-hard* sequences, respectively. For the single-camera easy and hard sequences, the IDF1 improvement is 12.2 and 13.5 points, and MOTA improves by 16.6 and 10.4 points.

Compared to unrefereed submissions, we perform slightly worse on IDP on the hard sequence. This could be due to a choice of detector that works better for crowded scenarios, a detector that is more conservative, and/or more conservative association. We nonetheless outperform all methods on IDF1, IDR and MOTA.

	BIPCC [57]	PT_BIPCC [49]	MTMC_CDSC [68]	MYTRACKER [72]	MTMC_ReID [79]†	**DeepCC**	BIPCC [57]	PT_BIPCC [49]	MTMC_CDSC [68]	MYTRACKER [72]	MTMC_ReID [79]†	**DeepCC**	BIPCC [57]	PT_BIPCC [49]	MTMC_CDSC [68]	MYTRACKER [72]	MTMC_ReID [79]†	**DeepCC**	BIPCC [57]	PT_BIPCC [49]	MTMC_CDSC [68]	MYTRACKER [72]	MTMC_ReID [79]†	**DeepCC**
	MOTA						IDP						IDR						IDF1					
Easy-all	59.4	59.3	70.9	77.7	83.6	**87.5**	83.6	84.8	87.6	87.5	91.2	**91.7**	60.4	61.4	68.6	73.8	82.0	**86.7**	70.1	71.2	77.0	80.0	86.3	**89.2**
Cam1	43.0	42.9	69.9	84.9	87.4	**93.3**	91.2	91.9	89.1	89.7	91.1	**95.6**	41.8	42.2	67.7	79.6	86.2	**93.0**	57.3	57.8	76.9	84.3	88.6	**94.3**
Cam2	44.8	44.7	71.5	78.4	84.2	**87.1**	69.3	70.4	90.9	88.9	92.4	**93.6**	67.1	68.0	73.4	75.9	82.9	**87.4**	68.2	69.2	81.2	81.9	87.4	**90.4**
Cam3	57.8	57.8	67.4	65.7	**82.4**	79.7	78.9	78.2	76.3	76.2	**87.8**	86.2	48.8	48.4	56.0	63.5	**79.7**	77.7	60.3	59.8	64.6	69.3	**83.6**	81.8
Cam4	63.2	63.2	76.8	79.8	**91.9**	91.8	88.7	91.7	91.2	84.1	**97.7**	96.3	62.8	64.9	79.0	77.6	93.1	**94.4**	73.5	76.0	84.7	80.7	**95.4**	95.3
Cam5	72.8	72.6	68.9	76.6	80.8	**86.2**	83.0	83.0	76.1	81.4	**87.2**	83.6	65.4	65.6	61.9	67.3	75.8	**77.7**	73.2	73.3	68.3	73.7	**81.1**	80.6
Cam6	73.4	73.4	77.0	82.8	83.1	**88.7**	87.5	91.7	91.6	88.9	91.7	**93.4**	69.1	72.4	75.3	78.8	82.5	**92.2**	77.2	80.9	82.7	83.5	86.9	**92.8**
Cam7	71.4	71.4	73.8	77.0	80.8	**82.2**	93.6	93.6	**94.0**	91.4	92.8	93.7	70.6	70.6	72.5	73.5	80.1	**83.7**	80.5	80.5	81.8	81.5	86.0	**88.5**
Cam8	60.7	60.9	63.4	71.6	79.9	**85.0**	**92.2**	**92.2**	89.1	90.8	91.1	89.4	59.6	60.0	61.8	71.3	78.6	**82.4**	72.4	72.7	73.0	79.9	84.4	**85.8**
Hard-all	54.6	54.4	59.6	59.0	69.6	**70.0**	81.2	81.8	81.4	74.5	**90.1**	87.4	53.5	54.0	54.7	55.2	68.1	**72.0**	64.5	65.0	65.5	63.4	77.6	**79.0**
Cam1	37.8	37.4	63.2	61.1	74.4	**79.6**	92.5	91.9	83.0	72.2	92.3	**94.7**	36.8	36.7	56.4	58.4	76.1	**80.1**	52.7	52.5	67.1	64.6	83.4	**86.8**
Cam2	47.3	46.6	54.8	50.4	**70.9**	57.9	65.7	66.0	78.8	61.2	**89.1**	77.5	56.1	56.7	53.1	52.6	66.7	**67.3**	60.6	61.0	63.4	56.6	**76.3**	72.0
Cam3	46.7	46.7	68.8	70.3	**87.1**	84.2	**96.1**	96.1	91.1	86.9	94.9	90.8	46.5	46.5	73.7	74.1	**89.2**	87.1	62.7	62.7	81.5	80.0	**91.9**	88.9
Cam4	85.3	85.5	75.6	81.2	**95.0**	90.3	86.0	93.6	87.1	84.4	**97.3**	93.0	82.7	91.0	78.1	82.2	**97.7**	97.0	84.3	92.3	82.3	83.3	**97.5**	94.9
Cam5	78.3	78.3	78.6	81.9	77.2	**86.0**	90.1	90.1	91.5	**93.3**	88.4	90.9	75.1	75.1	75.7	79.2	75.3	**85.5**	81.9	81.9	82.8	85.7	81.3	**88.1**
Cam6	59.4	59.4	53.3	56.1	58.4	**63.3**	81.7	82.4	71.2	70.0	86.3	**87.0**	52.7	53.3	42.3	44.9	55.4	**62.2**	64.1	64.7	53.1	54.7	67.5	**72.5**
Cam7	50.8	50.6	50.8	49.8	60.3	**61.4**	81.2	81.4	84.7	74.7	**91.4**	85.2	47.1	47.2	47.1	44.4	59.7	**61.3**	59.6	59.8	60.6	55.7	**72.2**	71.3
Cam8	73.0	73.0	70.0	71.5	**85.6**	85.0	**94.9**	94.9	90.3	93.5	92.2	92.3	72.8	72.8	73.9	70.5	83.7	**87.7**	82.4	82.4	81.3	80.4	87.7	**89.9**

Table 3. Detailed DukeMTMCT single-camera tracking results for the *test-easy* and *test-hard* sequences. Methods in † are unrefereed submissions.

It is worth noting that our method achieves the highest identity recall IDR on all scenarios, and on nearly all single-camera sequences. Identity recall is Achille's heel for modern multi-target trackers, as they commonly fail to re-identify targets after occlusions [44]. We believe that this improvement is a combination of better detections, joint optimization, and a discriminative feature embedding.

5.3. Impact of Loss and Hard Negative Mining

Our Re-ID results for similarity-based ranking are shown in Tables 4 and 5. Scores are averages of five repetitions and no test-time augmentation is used. (a) Our Adaptive Weighted Triplet Loss (AWTL) consistently improves over the batch-hard loss [35, 54]. (b) When training with square Euclidean distance to emphasize sensitivity to outliers our loss is robust in all scenarios, whereas the batch-hard loss shows to be unstable on the Duke dataset. (c) The proposed hard identity mining scheme (HNM) is also beneficial, and our adaptive weighted loss is both accurate and stable with difficult batches. (d) We also compare against a recent method that combines two network streams for better performance [22]. When employing a similar technique (2-stream ensemble) we improve our ranking accuracy further.

5.4. Accuracy of Tracking vs. Ranking

As more and more re-identification methods are being applied to multi-target tracking, we study the relation between ID measures for MTMC tracking and rank measures for Re-ID. In this experiment, we freeze ground truth single-camera trajectories and perform across-camera tracking with features at different times during training, resulting in different levels of ranking accuracy. Appearance features are learned from scratch using the 461 DukeMTMC-reID training IDs that do not appear in the validation sequence. Tracking accuracy is evaluated on the DukeMTMC validation sequence (241 IDs), and rank-1 accuracy on both DukeMTMC-reID test (702 IDs) and DukeMTMC validation. Results are shown in Figures 5-6.

We observe the following: Figure 5: Rank-1 accuracy for DukeMTMC-reID test and DukeMTMC validation correlate, even if the former is more difficult than the latter due to 3x as many identities. Figure 6: (a) Features with modest rank-1 performance can still do well in MTMCT because of more limited and diverse identities to compare between, and because tracking is also helped by motion information. (b) MTMCT IDF1 performance improves with rank-1 accuracy. However, after a point, further improvement in rank-1 accuracy yields diminishing returns in IDF1.

Our interpretation for this saturation effect is as follows. Initially, the Re-ID model learns to separate positive and negative samples, and tracking performance increases linearly with rank-1 performance. Once enough correlations have the correct sign, correlation clustering can infer the remaining missing agreements by enforcing transitivity (inequality 7). Therefore, correcting the sign of the remaining correlations has a smaller effect on IDF1. Even beyond that point, the Re-ID model tries to satisfy the separation margin of L_3 by further pulling co-identical samples together and non co-identical ones apart. These changes do not affect the correlation signs and have little influence on IDF1.

5.5. Weakness Analysis

We analyze the one-to-one ID mapping between true and computed trajectories to understand failures in the DukeMTMC validation sequence. During evaluation, each true trajectory that is mapped to an actual computed trajectory (not a false positive) has its own ID recall, as some of its detections could be missed by the tracker. Similarly, the computed trajectories have their own precision, as they can contain false positive detections.

	Euclidean		SqEuclidean	
	mAP	rank-1	mAP	rank-1
BoW+KISSME [81]	12.17	25.13	-	-
LOMO+XQDA [46]	17.04	30.75	-	-
Baseline [82]	44.99	65.22	-	-
PAN [83]	51.51	71.59	-	-
SVDNet [64]	56.80	76.70	-	-
TriHard [35]	54.60	73.24	0.28	0.89
AWTL	**54.97**	**74.23**	**52.37**	**71.45**
TriHard (+Aug)	56.65	74.91	0.48	1.25
AWTL (+Aug)	**57.28**	**75.31**	**55.94**	**75.04**
TriHard (+Aug+HNM)	54.90	74.23	0.30	0.94
AWTL (+Aug+HNM)	**58.74**	**77.69**	**57.84**	**76.21**
DPFL (1-stream) [22]	48.90	70.10	-	-
DPFL (2-stream) [22]	60.60	79.20	-	-
AWTL (2-stream)	**63.40**	**79.80**	**63.27**	**79.08**

Table 4. Re-ID results on DukeMTMC-ReID

	Euclidean		SqEuclidean	
	mAP	rank-1	mAP	rank-1
DNS [76]	29.87	55.43	-	-
GatedSiamese [69]	39.55	65.88	-	-
PointSet [86]	44.27	70.72	-	-
SomaNet [5]	47.89	73.87	-	-
PAN [83]	63.35	82.81	-	-
TriHard [35]	66.63	82.99	64.47	82.01
AWTL	**68.03**	**84.20**	**65.95**	**82.16**
TriHard (+Aug)	69.57	85.14	68.92	84.12
AWTL (+Aug)	**70.83**	**86.11**	**69.64**	**84.71**
TriHard (+Aug+HNM)	71.13	86.40	0.16	0.36
AWTL (+Aug+HNM)	**71.76**	**86.94**	**70.19**	**85.39**
DPFL (1-stream) [22]	66.50	85.70	-	-
DPFL (2-stream) [22]	72.60	88.06	-	-
AWTL (2-stream)	**75.67**	**89.46**	**74.81**	**87.92**

Table 5. Re-ID results on Market-1501

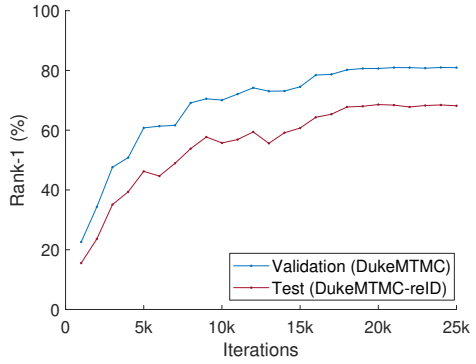

Figure 5. Relation between validation and test sets.

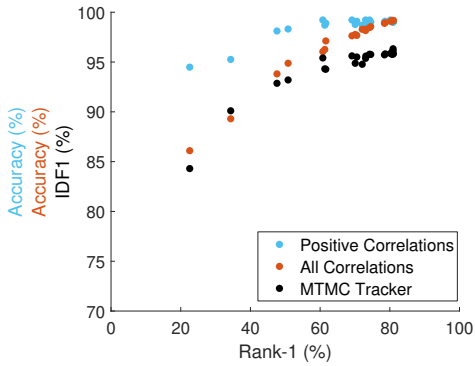

Figure 6. Relation of tracking, correlation, and rank accuracy.

We rank computed trajectories by ID precision and true trajectories by ID recall, then inspect the trajectories with the lowest scores. This helps clarify which situations are difficult in single- and multi-camera scenarios. Single- and multi-camera scenarios are analyzed separately because their ID mapping is different.

Two failure cases are illustrated in Figure 7. In single-camera tracking, correlations are poor when there is significant pose change, significant occlusion, and/or abrupt motion, resulting in low identity recall (left in the Figure). In

Figure 7. *Left*: A multi-camera trajectory with low identity precision. *Right*: Example ground truth trajectory with poor identity recall in single camera tracking. Red indicates failure.

multi-camera tracking, fragmentation is mostly caused by delays in blind spots and unpredictable motion. Merge errors happen in cases where people dress similarly and their inter-camera motion is plausible.

The example in Figure 7 (right) highlights one of the most difficult situations in the validation sequence, where several construction workers share similar appearance. They enter and exit the field of view a few times, and both appearance and motion correlations are weak, resulting in poor identity recall during tracking.

6. Conclusion

We showed that a new triplet loss with real-valued, adaptive weights, coupled with a new hard-identity mining technique that mixes difficult and random identities, yields appearance features that achieve state-of-the art performance in both MTMCT and Re-ID, whether measured by IDF1, MOTA, or rank-1 scores.

Our experiments also elucidate the relation between changes in rank-1 Re-ID score and changes in IDF1 tracking accuracy. The two performance measures relate linearly with each other at first, but the dependency saturates once rank-1 scores are good enough to yield data association correlations with the correct signs.

We hope that new large-scale data sets will be introduced to further validate our ideas.

References

[1] M. Ayazoglu, B. Li, C. Dicle, M. Sznaier, and O. Camps. Dynamic subspace-based coordinated multicamera tracking. In *2011 IEEE International Conference on Computer Vision (ICCV)*, pages 2462–2469, Nov. 2011. 2

[2] D. Baltieri, R. Vezzani, and R. Cucchiara. Learning articulated body models for people re-identification. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, pages 557–560, New York, NY, USA, 2013. ACM. 2

[3] D. Baltieri, R. Vezzani, and R. Cucchiara. Mapping appearance descriptors on 3d body models for people re-identification. *International Journal of Computer Vision*, 111(3):345–364, 2015. 2

[4] N. Bansal, A. Blum, and S. Chawla. Correlation clustering. In *Foundations of Computer Science*, 2002. 2

[5] I. B. Barbosa, M. Cristani, B. Caputo, A. Rognhaugen, and T. Theoharis. Looking beyond appearances: Synthetic training data for deep cnns in re-identification. *arXiv preprint arXiv:1701.03153*, 2017. 2, 8

[6] A. Bedagkar-Gala and S. Shah. Multiple person re-identification using part based spatio-temporal color appearance model. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 1721–1728, Nov 2011. 2

[7] A. Bedagkar-Gala and S. K. Shah. Part-based spatio-temporal model for multi-person re-identification. *Pattern Recognition Letters*, 33(14):1908 – 1915, 2012. Novel Pattern Recognition-Based Methods for Re-identification in Biometric Context. 2

[8] J. Berclaz, F. Fleuret, E. Türetken, and P. Fua. Multiple object tracking using k-shortest paths optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011. 2

[9] K. Bernardin and R. Stiefelhagen. Evaluating multiple object tracking performance: the CLEAR MOT metrics. *EURASIP Journal on Image and Video Processing*, (246309):1–10, 2008. 6

[10] L. Beyer, S. Breuers, V. Kurin, and B. Leibe. Towards a principled integration of multi-camera re-identification and tracking through optimal bayes filters. *CVPRWS*, 2017. 2

[11] M. Bredereck, X. Jiang, M. Korner, and J. Denzler. Data association for multi-object Tracking-by-Detection in multi-camera networks. In *2012 Sixth International Conference on Distributed Smart Cameras (ICDSC)*, pages 1–6, Oct. 2012. 2

[12] W. Brendel, M. Amer, and S. Todorovic. Multiobject tracking as maximum weight independent set. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1273–1280. IEEE, 2011. 2

[13] A. A. Butt and R. T. Collins. Multiple target tracking using frame triplets. In *Computer Vision–ACCV 2012*, pages 163–176. Springer, 2013. 2

[14] Y. Cai and G. Medioni. Exploring context information for inter-camera multiple target tracking. In *2014 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 761–768, Mar. 2014. 2

[15] S. Calderara, R. Cucchiara, and A. Prati. Bayesian-competitive consistent labeling for people surveillance. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(2):354–360, Feb 2008. 2

[16] L. Cao, W. Chen, X. Chen, S. Zheng, and K. Huang. An equalised global graphical model-based approach for multi-camera object tracking. *ArXiv:11502.03532 [cs]*, Feb. 2015. 2

[17] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 2, 3, 4, 6

[18] V. Chari, S. Lacoste-Julien, I. Laptev, and J. Sivic. On pairwise costs for network flow multi-object tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5537–5545, 2015. 2

[19] K.-W. Chen, C.-C. Lai, P.-J. Lee, C.-S. Chen, and Y.-P. Hung. Adaptive Learning for Target Tracking and True Linking Discovering Across Multiple Non-Overlapping Cameras. *IEEE Transactions on Multimedia*, 13(4):625–638, Aug. 2011. 2

[20] X. Chen, L. An, and B. Bhanu. Multitarget Tracking in Nonoverlapping Cameras Using a Reference Set. *IEEE Sensors Journal*, 15(5):2692–2704, May 2015. 2

[21] X. Chen, K. Huang, and T. Tan. Direction-based stochastic matching for pedestrian recognition in non-overlapping cameras. In *2011 18th IEEE International Conference on Image Processing (ICIP)*, pages 2065–2068, Sept. 2011. 2

[22] Y. Chen, X. Zhu, and S. Gong. Person re-identification by deep learning multi-scale representations. 2017. 7, 8

[23] D. Cheng and M. Cristani. Person re-identification by articulated appearance matching. In S. Gong, M. Cristani, S. Yan, and C. C. Loy, editors, *Person Re-Identification*, Advances in Computer Vision and Pattern Recognition, pages 139–160. Springer London, 2014. 2

[24] D. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino. Custom pictorial structures for re-identification. In *Proceedings of the British Machine Vision Conference*, pages 68.1–68.11. BMVA Press, 2011. http://dx.doi.org/10.5244/C.25.68. 2

[25] R. T. Collins. Multitarget data association with higher-order motion models. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1744–1751. IEEE, 2012. 2

[26] S. Daliyot and N. S. Netanyahu. A Framework for Inter-camera Association of Multi-target Trajectories by Invariant Target Models. In J.-I. Park and J. Kim, editors, *Computer Vision - ACCV 2012 Workshops*, number 7729 in Lecture Notes in Computer Science, pages 372–386. Springer Berlin Heidelberg, 2013. 2

[27] A. Das, A. Chakraborty, and A. K. Roy-Chowdhury. Consistent re-identification in a camera network. In *Computer Vision–ECCV 2014*, pages 330–345. Springer, 2014. 2

[28] A. Dehghan, S. M. Assari, and M. Shah. Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking. In *CVPR*, volume 1, page 2, 2015. 2

[29] E. D. Demaine, D. Emanuel, A. Fiat, and N. Immorlica. Correlation clustering in general weighted graphs. *Theoretical Computer Science*, 361(2-3):172–187, 2006. 2

[30] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2010. 2

[31] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua. Multi-Camera People Tracking with a Probabilistic Occupancy Map. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):267–282, February 2008. 2

[32] A. Gilbert and R. Bowden. Tracking Objects Across Cameras by Incrementally Learning Inter-camera Colour Calibration and Patterns of Activity. In A. Leonardis, H. Bischof, and A. Pinz, editors, *Computer Vision ECCV 2006*, number 3952 in Lecture Notes in Computer Science, pages 125–136. Springer Berlin Heidelberg, 2006. 2

[33] R. Hamid, R. Kumar, M. Grundmann, K. Kim, I. Essa, and J. Hodgins. Player localization using multiple static cameras for sports visualization. In *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 731–738, June 2010. 2

[34] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[35] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 2, 3, 4, 7, 8

[36] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *European Conference on Computer Vision*, pages 34–50. Springer, 2016. 2

[37] H. Izadinia, I. Saleemi, W. Li, and M. Shah. Mp2t: Multiple people multiple parts tracker. In A. W. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, editors, *ECCV (6)*, volume 7577 of *Lecture Notes in Computer Science*, pages 100–114. Springer, 2012. 2

[38] O. Javed, K. Shafique, Z. Rasheed, and M. Shah. Modeling inter-camera spacetime and appearance relationships for tracking across non-overlapping views. *Computer Vision and Image Understanding*, 109(2):146–162, Feb. 2008. 2

[39] W. Jiuqing and L. Li. Distributed optimization for global data association in non-overlapping camera networks. In *2013 Seventh International Conference on Distributed Smart Cameras (ICDSC)*, pages 1–7, Oct. 2013. 2

[40] A. Kamal, J. Farrell, and A. Roy-Chowdhury. Information Consensus for Distributed Multi-target Tracking. In *2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2403–2410, June 2013. 2

[41] R. Kumar, G. Charpiat, and M. Thonnat. Multiple object tracking by efficient graph partitioning. In *Computer Vision–ACCV 2014*, pages 445–460. Springer, 2014. 2

[42] C.-H. Kuo, C. Huang, and R. Nevatia. Inter-camera Association of Multi-target Tracks by On-Line Learned Appearance Affinity Models. In K. Daniilidis, P. Maragos, and N. Paragios, editors, *Computer Vision ECCV 2010*, number 6311 in Lecture Notes in Computer Science, pages 383–396. Springer Berlin Heidelberg, 2010. 2

[43] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler. Motchallenge 2015: Towards a benchmark for multi-target tracking. *arXiv:1504.01942 [cs]*, Apr. 2015. arXiv: 1504.01942. 2, 5

[44] L. Leal-Taixé, A. Milan, K. Schindler, D. Cremers, I. Reid, and S. Roth. Tracking the trackers: An analysis of the state of the art in multiple object tracking. *arXiv preprint arXiv:1704.02781*, 2017. 7

[45] Y. Liang and Y. Zhou. Multi-camera tracking exploiting person re-id technique. In *International Conference on Neural Information Processing*, pages 397–404. Springer, 2017. 6

[46] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2197–2206, 2015. 8

[47] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 2

[48] D. Makris, T. Ellis, and J. Black. Bridging the gaps between cameras. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004*, volume 2, June 2004. 2

[49] A. Maksai, X. Wang, F. Fleuret, and P. Fua. Non-markovian globally consistent multi-object tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 6, 7

[50] N. Martinel, C. Micheloni, and G. L. Foresti. Saliency weighted features for person re-identification. In *Computer Vision-ECCV 2014 Workshops*, pages 191–208. Springer International Publishing, 2014. 2

[51] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016. 2

[52] A. Milan, S. H. Rezatofighi, A. Dick, I. Reid, and K. Schindler. Online multi-target tracking using recurrent neural networks. In *AAAI*, February 2017. 2, 3

[53] A. Milan, S. Roth, and K. Schindler. Continuous energy minimization for multitarget tracking. *IEEE TPAMI*, 36(1):58–72, 2014. 2

[54] A. Mishchuk, D. Mishkin, F. Radenovic, and J. Matas. Working hard to know your neighbor's margins: Local descriptor learning loss. In *Advances in Neural Information Processing Systems*, pages 4829–4840, 2017. 3, 7

[55] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes. Globally-optimal greedy algorithms for tracking a variable number of objects. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1201–1208. IEEE, 2011. 2

[56] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 2

[57] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision Workshops*, pages 17–35. Springer, 2016. 2, 3, 5, 6, 7

[58] E. Ristani and C. Tomasi. Tracking multiple people online and in real time. In *ACCV-12th Asian Conference on Computer Vision*. Springer, 2014. 2

[59] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 4

[60] S. Schulter, P. Vernaza, W. Choi, and M. Chandraker. Deep network flow for multi-object tracking. *CVPR*, 2017. 3

[61] K. Shafique and M. Shah. A noniterative greedy algorithm for multiframe point correspondence. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(1):51–65, 2005. 2

[62] G. Shu, A. Dehghan, O. Oreifej, E. Hand, and M. Shah. Part-based multiple-person tracking with partial occlusion handling. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1815–1821. IEEE, 2012. 2

[63] Springer. *MARS: A Video Benchmark for Large-Scale Person Re-identification*, 2016. 2

[64] Y. Sun, L. Zheng, W. Deng, and S. Wang. Svdnet for pedestrian retrieval. 2017. 2, 8

[65] S. Tang, B. Andres, M. Andriluka, and B. Schiele. Subgraph decomposition for multi-target tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5033–5041, 2015. 2

[66] S. Tang, B. Andres, M. Andriluka, and B. Schiele. Multi-person tracking by multicut and deep matching. In *European Conference on Computer Vision*, pages 100–111. Springer, 2016. 2, 3

[67] S. Tang, M. Andriluka, B. Andres, and B. Schiele. Multiple people tracking by lifted multicut and person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3539–3548, 2017. 2, 3

[68] Y. T. Tesfaye, E. Zemene, A. Prati, M. Pelillo, and M. Shah. Multi-target tracking in multiple non-overlapping cameras using constrained dominant sets. *arXiv preprint arXiv:1706.06196*, 2017. 6, 7

[69] R. R. Varior, M. Haloi, and G. Wang. Gated siamese convolutional neural network architecture for human re-identification. In *European Conference on Computer Vision*, pages 791–808. Springer, 2016. 8

[70] L. Wen, W. Li, J. Yan, Z. Lei, D. Yi, and S. Z. Li. Multiple target tracking based on undirected hierarchical relation hypergraph. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1282–1289. IEEE, 2014. 2

[71] B. Wu and R. Nevatia. Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. *International Journal of Computer Vision*, 75(2):247–266, 2007. 2

[72] K. Yoon, Y.-m. Song, and M. Jeon. A multiple hypothesis tracking algorithm for multi-target multi-camera tracking with disjoint views. 02 2018. 6, 7

[73] F. Yu, W. Li, Q. Li, Y. Liu, X. Shi, and J. Yan. Poi: Multiple object tracking with high performance detection and appearance feature. In *European Conference on Computer Vision*, pages 36–42. Springer, 2016. 2, 3

[74] A. Zamir, A. Dehghan, and M. Shah. Gmcp-tracker: Global multi-object tracking using generalized minimum clique graphs. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2012. 2

[75] L. Zhang, Y. Li, and R. Nevatia. Global data association for multi-object tracking using network flows. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008. 2

[76] L. Zhang, T. Xiang, and S. Gong. Learning a discriminative null space for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1239–1248, 2016. 8

[77] S. Zhang, E. Staudt, T. Faltemier, and A. Roy-Chowdhury. A Camera Network Tracking (CamNeT) Dataset and Performance Baseline. In *2015 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 365–372, Jan. 2015. 2

[78] S. Zhang, Y. Zhu, and A. Roy-Chowdhury. Tracking multiple interacting targets in a camera network. *Computer Vision and Image Understanding*, 134:64–73, May 2015. 2

[79] Z. Zhang, J. Wu, X. Zhang, and C. Zhang. Multi-target, multi-camera tracking by hierarchical clustering: Recent progress on dukemtmc project. *arXiv preprint arXiv:1712.09531*, 2017. 6, 7

[80] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. 2

[81] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *Computer Vision, IEEE International Conference on*, 2015. 2, 5, 6, 8

[82] L. Zheng, Y. Yang, and A. G. Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016. 2, 3, 8

[83] Z. Zheng, L. Zheng, and Y. Yang. Pedestrian alignment network for large-scale person re-identification. *arXiv preprint arXiv:1707.00408*, 2017. 2, 3, 8

[84] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 2, 5

[85] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. Random erasing data augmentation. *arXiv preprint arXiv:1708.04896*, 2017. 2

[86] S. Zhou, J. Wang, J. Wang, Y. Gong, and N. Zheng. Point to set similarity based deep feature learning for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 8