Learning Local Feature Descriptors for Multiple Object Tracking

Dmytro Mykheievskyi[1], Dmytro Borysenko[1], and Viktor Porokhonskyy[1⋆]

Samsung R&D Institute Ukraine (SRK),
57 L'va Tolstogo Str., Kyiv 01032, Ukraine
{d.borysenko,d.mykheievsk,v.porokhonsk}@samsung.com

Abstract The present study aims at learning class-agnostic embedding, which is suitable for Multiple Object Tracking (MOT). We demonstrate that the learning of local feature descriptors could provide a sufficient level of generalization. Proposed embedding function exhibits on-par performance with its dedicated person re-identification counterparts in their target domain and outperforms them in others. Through its utilization, our solutions achieve state-of-the-art performance in a number of MOT benchmarks, which includes CVPR'19 Tracking Challenge.

1 Introduction

Multiple Object Tracking (MOT) problem has been receiving considerable attention from the computer vision community due to its significance to a number of practical tasks, such as scene understanding[1], activity recognition[2], behavior analysis[3], etc. Over the past few years, MOT solutions have gained considerable performance improvement[4–6] partially due to the advance in Object Detection. Namely, detectors rapidly progressed from Ada-Boost-based solutions, such as ACF detector[7], to CNN-based ones[8–12]. While the former are capable of returning bounding boxes with the fixed aspect ratio for a single category, the latter successfully deal with multiple categories and arbitrarily shaped objects. This progress was started by the invention of a two-stage detector[11]. It was further stimulated by the introduction of a single-stage paradigm[10] and consequent competition between the two[8, 9, 12]. The invention of Feature Pyramid Network (FPN)[13], backbone efficiency improvement[14, 15], Regression Cascades[16], Deformable Convolutions[17,18] also resulted in considerable accuracy gains. This progress was one of the main reasons why Detection-Based Tracking (DBT), which is also known as tracking-by-detection, has become the dominating paradigm in MOT domain[19]. It essentially breaks MOT problem into the following sub-tasks: detection, embedding function application to the image area corresponding to each detect, and data association. Making the detection stage independent of the remaining two, DBT facilitates the adoption of modern detectors. In its turn, any gain in Multiple Object Detection Accuracy (MODA)

⋆ Corresponding author.

strongly positively affects the main MOT metrics[19–21], such as Multiple Object Tracking Accuracy (MOTA), via False Positive (FP) and False Negative (FN) counts.

At the same time, contemporary DBT solutions do not live up to the full potential of their detection stage. In particular, most of them keep relying on embeddings learned with single category datasets, thus making the scope of the entire solution restricted to that single category. Moreover, the choice of categories themselves is quite limited as long as the public datasets[5, 6, 22–24] are concerned. The main two options are pedestrians and vehicles. Even though a few DBT solutions have managed to keep their scope up with the scope of the corresponding detector, so far all of them are associated with costly compromises. For example, SORT[20] became applicable to multiple categories via the sacrifice of appearance information. Relying solely on the intersection-over-union (IoU) between considered detects at the data association stage obviously reduces its accuracy and limits the applicability. Quite an opposite path was taken by Wang et al. [25], who proposed to Jointly learn the Detection and Embedding (JDE) model in a multi-task manner so that the detector and embedding could deal with the same set of categories. However, such simultaneous training requires ID annotation to be available throughout the corresponding dataset. Besides the batch formation procedure for embedding learning, in this case, is expected to be less flexible compared to a devoted one.

All in all, a wide application of contemporary DBT solutions in practice is somewhat impeded by either the necessity of specific dataset creation or their compromised accuracy and applicability. Due to the former factor, the tracking of arbitrary objects (e.g. animals, robots, biological cells, etc.) becomes prohibitively expensive for the vast majority of applications.

Perhaps, the most radical solution to the problem of restricted DBT scope would be the utilization of class-agnostic embedding functions. Among tentative candidates for this role, one could mention Learned Local Features Descriptors (LLFD)[26–29]. While corresponding embedding functions shall obviously be sensitive to object appearance, to the best of our knowledge their utilization for object representation has not been reported. Indeed, there are a few factors that make such endeavor questionable. In the first place, the representation of objects in the case of LLFD could hardly correspond to the conventional one. Namely, the objective of typical DBT embedding function is to produce compact and separable object manifolds, when the loss formulation is adopted from metric learning[30]. The same loss combined with a different sampling approach in the case of LLFD is expected to result in extended and non-separable manifolds. At the same time, one could argue that visually similar samples, even if attributed to different manifolds, would tend to get mapped close to each other in the metric space. Along with the gradual evolution of object appearance observed in typical MOT setting, such property could potentially serve as a trade-off for lost manifold compactness and separability. Next, susceptibility to background variation and/or occlusions is another point of concern. Finally, the discriminative capability of LLFD may suffer from a rather low resolution of its input.

As the nature of listed above concerns calls for an empirical approach, in this study we report our rather successful results of LLFD utilization for Multiple Object Tracking. In addition, we discuss the essential features of object representation, which corresponds to the employed embedding. Finally, the aspects, which turn out to have a positive impact on its efficiency, are indicated.

The paper is organized in the following way: the second section is dedicated to related work. Section 3 is devoted to the preparation of our DBT MOT solution, which takes into account the object representation expected from LLFD. In particular, we consider the necessary adjustments, which ameliorate embedding function performance. Among them are the resolution of input patches, preserving color information, etc. For the sake of reproducibility, the details on the detector and the association stage are provided as well. In the following section, our MOT solution is evaluated and compared with other methods. In Section 5, the properties of the proposed embedding function are discussed.

2 Related Work

The two approaches to embedding learning, which are relevant to our study, are the following: person re-identification (Re-ID) and learned local feature descriptors. While being similar in some aspects, they exhibit a conceptual difference regarding the criteria, according to which given two samples are treated as such that belong to the same category or, in other words, could form a positive sample pair during training. In the case of person Re-ID, image patches representing a given category have to depict the same object. The difference in object appearance in these patches remains disregarded during the label assignment. In the case of LLFD, however, the appearance of samples affects label assignment in a more profound manner. Namely, it is required that the patches ascribed to a given category depict the same pattern. At the same time, it could be viewed from different perspectives and/or under various lighting conditions[26,31]. Typically each such pattern represents the vicinity of a particular local keypoint. Due to their nature, the patterns normally do not possess any specific boundary. As a consequence, the patterns extend right to the borders of corresponding patches. This is another important difference from the case of person Re-ID, where the image patches serving as training samples depict objects with well-defined boundaries. And these boundaries rarely extend to the patch borders. Some other aspects of these two approaches are summarized below.

2.1 Embedding Learning: Person Re-ID

Let us consider some person Re-ID aspects, which are relevant to MOT problem. While several approaches are being used in this domain the solutions applied to MOT task usually rely on metric learning. The employed backbone topologies vary from those adopted from classification domain to the specifically designed ones, as e.g. OSNet[32]. ResNet-50[14] became de facto the standard option in the former case. Out of several representation options, global features turn out

to be the most popular in DBT paradigm[19]. Successful solution tend to benefit from attention utilization[30,33,34]. Multi-scale feature learning is another topic of active research, which is approached explicitly by e.g. OSNet[32].

As far as training data are concerned, it is worth mentioning the following factors. There is a number of widely used image and video datasets, see Ref. [30] for an extensive list. Some of them are restricted to cropped patches, while others contain entire scenes. The annotation quality depends on whether a manual or automatic method was employed. The latter may include the application of Object Detectors (OD), trackers, and Re-ID solutions. However, even in the former case, the noise rate remains non-negligible[30]. Occlusions are quite frequent due to the scene and objects nature.

2.2 Embedding Learning: Local Feature Descriptors

Due to this task nature, it is desirable to start with the description of public datasets. Among the most influential options in this area are UBC PhotoTour[31] and HPatches[26]. Despite being introduced in 2007, UBC PhotoTour provides a sufficient amount of data for CNN training. Being a patch-based dataset it contains more than 10^6 grayscale patches extracted from three scenes. The patches represent image areas around local keypoints. The positive samples correspond to the matching keypoints. Recently introduced HPatches[26] addresses several limitations inherent in UBC PhotoTour, such as scarce diversity with respect to data, tasks, and the type of features. In particular, it consists of 116 sequences containing 6 RGB images each. These sequences form two groups: 59 sequences with significant viewpoint change and 57 with illumination change. In addition to the extracted grayscale patches, the dataset includes the set of original images in RGB format as well as the homography matrices representing the transformation between the images belonging to the same sequence.

The necessity to process numerous features per image imposes certain limitations on the network topologies. For this reason, specific lightweight CNN topologies find use in this domain. In recent years fully convolutional 7 layer CNN design of L2-Net[28] has been enjoying broad utilization[27,35–38].

Since the training task is usually formulated as a metric learning problem, a lot of effort is being invested into the following two directions: formulation of more efficient loss function; and rising the proportion of so-called hard training samples. The former activity is mainly related to modifying the triplet loss. The latter deals with the fact that a fair portion of randomly formed triplets tends to satisfy the objective function right away. Balntas *et al.* [29] approach this problem by means of the anchor swap, also known as hard-negative sampling within the triplet. In HardNet[27] the hardest-in batch sampling strategy was proposed. In this case, triplet loss reads

$$L_t = \frac{1}{N}\sum_{i=1}^{N} \max\left\{ \left\| x_i - x_i^+ \right\|_2 + 1 - \min_{j \neq i} \| x_i - x_j \|_2, 0 \right\}, \tag{1}$$

where $(x_i, x_i^+)_{i=1...N}$ designate descriptors corresponding to a set of N positive pairs. Each pair originates from the same local keypoint. This method gained broad recognition[35, 36] due to its straightforward implementation and lack of side-effects. As examples of triplet loss modification, it is worthwhile to mention the following. Zhang *et al.*[39] proposed a regularization term named Global Orthogonal Regularization in order to achieve a more uniform distribution of embedding vectors over the unit hypersphere. The authors of SOS-Net[36] put forward an additional term named Second-Order Similarity aimed at improving cluster compactness in the metric space. In terms of Eq. 1, the net loss function, in this case, assumes the next form

$$L = L_t + \frac{1}{N}\sum_{i=1}^{N}\sqrt{\sum_{j\neq i, j\in c_i}^{N}\left(\left\|x_i - x_j\right\|_2 - \left\|x_i^+ - x_j^+\right\|_2\right)^2}, \tag{2}$$

where

$$c_j = \left\{x_i \in KNN\left(x_j\right) \vee x_i^+ \in KNN\left(x_j^+\right)\right\} \tag{3}$$

with $KNN\left(x_i\right)$ being a set of k nearest neighbors in the euclidean space.

3 ODESA-based Tracker

In order to assess the applicability of LLFD for MOT problem we prepare a DBT solution, which accounts for the peculiarities in corresponding object representation, and test its performance with relevant benchmarks. Regarding these peculiarities, we make the following assumptions. Gradual appearance evolution of objects observed between adjacent frames in a typical MOT environment being combined with local descriptor capability to relate similar patterns is expected to produce object manifolds that evolve in the metric space in a non-abrupt manner. At longer time scales, certain objects may exhibit drastic changes in appearance. Thus corresponding manifolds could likely become quite extended. Periodic motion potentially could produce closed "trajectories" in the metric space. Also, the embedding derived from LLFD is assumed to exhibit better discrimination properties, if color information is preserved. This assumption is based on much lower probability to encounter lighting condition variations in a MOT environment compared to e.g. the case of image retrieval, where LLFD are typically find use. We also assume them to benefit from higher resolution of input patches. To distinguish the embeddings that were modified according to our assumptions from their LLFD origin, we will refer to them, hereafter, as Object DEscriptor that is Smooth Appearance-wise (ODESA).

The rest of this section is devoted to the description of our DBT solution, which in many aspects turns out to be similar to DeepSORT[40]. The subsections devoted to its components are ordered in accordance with DBT processing flow.

3.1 Object Detector

The main objective for detector design was to obtain a competitive solution constructed from the building blocks reported elsewhere[11, 13–16, 18, 41]. In

Table 1: Ablation analysis of object detector components. R101, X101, Cascade, DCN, GCB, Libra stand for ResNet-101, ResNeXt-101, Cascade R-CNN, Deformable ConvNets v2, Global Context Block, and specific Libra R-CNN components, respectively. mAP(Mod) represents the mean Average Precision value for Moderate KITTI split.

R101[14]	X101[15]	Cascade[16]	DCN[18]	GCB[41]	Libra[44]		FP ↓	FN ↓	MODA ↑	mAP(Mod) ↑
✓						Car	**685**	2500	86.77	89.88
	✓					Car	967	2113	87.20	**90.05**
	✓	✓				Car	857	2021	88.04	90.00
	✓	✓	✓	✓		Car	876	1876	88.50	89.92
	✓	✓	✓	✓	✓	Car	774	**1671**	**89.84**	89.87
✓						Pedestrian	1162	2723	65.03	82.91
	✓					Pedestrian	**787**	3019	65.74	78.40
	✓	✓				Pedestrian	843	2809	67.13	78.83
	✓	✓	✓	✓		Pedestrian	934	2658	67.67	**85.69**
	✓	✓	✓	✓	✓	Pedestrian	1018	**2447**	**68.81**	85.61

particular, the component implementations of MMDetection repository[42] were reused. KITTI Tracking Benchmark[4] was considered as the main target.

The two-stage detector based on Faster R-CNN[11] was taken as the starting point. ResNet-101[14] and ResNeXt-101[15] pre-trained on ImageNet[43] were opted for the backbone. The following components were utilized as extension options: FPN[13], Regression Cascades[16], the deformable convolution[18], and Global Context Blocks (GCB)[41]. Also, IoU-balanced Sampling, Balanced Feature Pyramid, and Balanced L1 Loss of Libra R-CNN[44] were considered.

For experiments conducted on KITTI dataset[22], separate detectors were prepared for *Car* and *Pedestrian* targets. 7481 training images for the detection task were divided into a *trainset* of 4418 images and a *local validation* set of 3063 images so that the former does not contain any samples from 21 training sequences for the tracking task. 4418 KITTI images were supplemented with a sub-set from BDD[45] dataset, where 4975 suitable images were selected. Each detector was configured to output two categories. These were either *Car* and *Van* or *Pedestrian* and *Cyclist*. Such a configuration helps to get better accuracy.

The models were trained using SGD optimizer with momentum 0.9 and weight decay 10^{-4}. The learning rate was set to linearly increase from 3×10^{-4} to 10^{-2} over the initial 500 warm-up training steps. The resulting rate was used for 15 epochs. Finally, it was set to 10^{-3} for the last 10 epochs. The training was performed with a batch size of 16 on 8 GPUs. As augmentation, we applied multi-scale training and random horizontal flips. During evaluation and inference, the input images were resized to bring the shorter side to 700 pixels with preserved aspect ratio. The results of detector evaluation on 21 KITTI tracking train sequences are presented in Table 1.

3.2 ODESA Embedding

The topology of CNN adopted for ODESA models, which is depicted in Fig. 1, was derived from L2-Net[28]. The modifications are related to adopting multi-

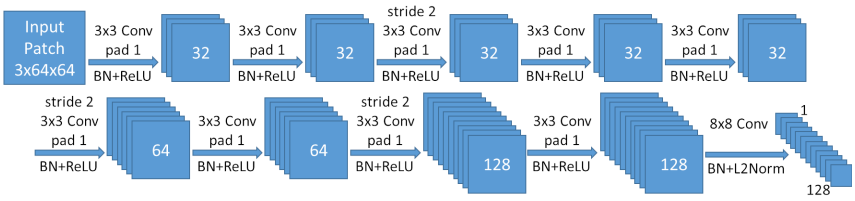

Fig. 1. CNN topology of ODESA(HSV64/128) model. The last seven convolutions correspond to L2-Net. The three layers introduced at the beginning reduce the lateral dimension of feature maps to 32x32, which is the standard size of LLFD input patch.

channel input and the increase of input patch resolution.

For training, we utilize exclusively the data from HPatches dataset[26]. In particular, the original set of RGB images and corresponding homography matrices were used as the starting point for patch extraction. By conducting this procedure by ourselves we obtained a set of patches with preserved color information. The original set is accessible in the grayscale format only. And to the best of our knowledge, the corresponding keypoint parameters are not publicly available. As for the patch extraction routine itself, we followed the procedure of Ref. [26]. The random transformations of keypoints with respect to their position, scale, and orientation were reproduced as well. Unless explicitly stated otherwise, they were applied before projecting keypoints into related images in order to simulate the detection noise. As it will be discussed later these transformations significantly affect embedding properties. Single intentional modification of the patch extraction routine concerned with the choice of local keypoint detectors. Namely, as an alternative to the original option, i.e. a combination of Difference of Gaussian, Hessian-Hessian, and Harris-Laplace[1] the following learnable Keypoint Detectors were considered: LF-Net[47] and D2-Net[48]. Hereafter, the former will be referred to as HandCrafted Keypoint Detectors (HCKD).

The loss function of Eq. 2 was used with the default settings from Ref. [36]. SGD optimizer with the momentum value of 0.9 and the weight decay of 10^{-4} was used. The learning rate was configured to decrease linearly from the initial value of 0.1 to zero over 20 epochs. Xavier[49] weight initialization method was employed. The batch size was set to 512. Random horizontal flips were used.

The embedding model for our tracking solution was selected via the validation procedure described in Ref. [25]. Essentially this routine estimates True Positive Rate at a given value of False Acceptance Rate (TPR@FAR) for a retrieval task set on a combination of person Re-ID datasets[24,50,51]. The results of such validation are presented in Table 2. As long as the range of FAR values from 10^{-4} to 10^{-1} is concerned, about every next entry in Table 2 shows a gradual improvement of TPR. In particular, the increase of input patch resolution from 32x32 to 64x64 results in higher performance. The patch sizes beyond 64x64 do not bring any considerable improvement. This observation indicates

[1] Corresponding VLFeat[46] implementations were employed similarly to Ref. [26].

Table 2: Embedding validation results obtained according to the routine described in Ref. [25] on a set of public person Re-ID datasets. The model name in the first column refers to the color space and the size of square input patch in pixels. A number of person Re-ID and LLFD models were added to serve as a reference.

Model	**Keypoint Detector**	**Embedding Size**	**TPR@FAR**, % ↑						
			10e-6	10e-5	10e-4	0.001	0.01	0.05	0.1
GRAY32	HCKD	128	7.08	11.30	18.48	31.32	52.47	70.38	78.67
GRAY32	D2-Net[48]	128	7.18	11.46	18.80	31.82	52.60	71.10	79.59
GRAY64	D2-Net[48]	128	**7.79**	**12.35**	19.74	32.75	53.72	72.24	80.38
RGB64	LF-Net[47]	128	5.35	10.91	20.14	35.13	60.29	79.25	86.59
RGB64	D2-Net[48]	128	4.79	10.65	19.45	35.85	61.62	80.21	87.33
HSV32	D2-Net[48]	128	5.06	10.93	20.51	36.78	61.78	80.27	87.30
HSV64	D2-Net[48]	128	4.76	12.01	**21.81**	**37.34**	62.37	80.87	87.87
HSV64	D2-Net[48]	512	4.65	10.59	20.70	36.37	**63.67**	**81.35**	**87.95**
SOSNet[36]	HPatches[26]	128	7.31	10.79	16.38	27.05	47.88	66.92	75.60
OSNet[32]	-	512	3.23	5.98	11.35	22.46	44.76	70.01	81.45
HACNN[33]	-	1024	2.96	11.10	20.21	33.46	54.34	74.43	83.93

that the upscaling of training patches from their original size of 65x65 pixels could be the limiting factor here. The datasets obtained by means of the learned keypoint detectors[47,48] permit to learn models, which consistently outperform the combination of hand-crafted keypoint detectors. The color information appears to be beneficial for this particular dataset. As will be shown later, the same conclusion holds for MOT datasets as well. We assume that this observation is related to limited changes in lighting conditions in typical person Re-ID or MOT scenes. HSV color space consistently outperforms other checked options, such as LAB, YUV, HLS, etc. Finally, the embedding dimensionality increase beyond 128 brings only moderate performance gains. The extended version of Table 2 is provided in the supplementary material.

3.3 Data Association Stage

The Hungarian algorithm[52] was employed to perform the association between incoming detects and known tracks. The appearance of each detect was reflected by a single embedding vector x_j of dimensionality k. The manifolds $\mathcal{H}_i = \{x_t\}_{t=1}^{D}$ containing up to D_{max} most recent embedding vectors, which were linked to each other in the past, represented the known tracks. Keeping in mind our assumptions about the non-compact nature of ODESA object representation and gradual appearance evolution, in the most general case we calculate the assignment matrix in the following way

$$c_{i,j} = (1-\lambda)\, min\left\{ \frac{1}{2}\sqrt{\sum_k \left(x_t^k - x_j^k\right)^2} \mid x_t \in \mathcal{H}_i \right\} + \lambda\left(1 - IoU\left(i,j\right)\right). \quad (4)$$

Here each matrix element is represented by the normalized weighted sum of two terms. The first one reflects the distance in the metric space between j-th detect

and its closest element of i-th manifold. The second one depends on the IoU value between the same detect and the most recent element from the considered manifold. In other words, such cost formulation promotes association with the closest manifold. Due to the second term, the association with the most recent manifold element may get an additional boost. λ values were kept up to 0.1 in all cases, where IoU was used in combination with the appearance information.

In order to incorporate motion information, we used the Mahalanobis gate. It was based on possible object location predicted by the Kalman filter[53, 54] as described in DeepSORT[40]. The association was regarded as permitted, i.e. $b_{i,j} = 1$, on condition that Mahalanobis gating got passed and its final cost turned out to be smaller than the threshold value τ:

$$b_{i,j} = \mathbb{1}\left[c_{i,j} \leq \tau\right]. \tag{5}$$

Comparing our data association stage with those reported elsewhere[19, 20, 40], one could notice that it does not contain any unique components. The single requirement to it, which follows from the assumed object representation, is related to the criterion of visual similarity estimation. The expectations regarding considerable manifold extent make the minimal distance the simplest and, at the same time, quite a reasonable choice. By applying this or similar criterion, about any DBT solution could be adjusted to the utilization of LLDF or ODESA-like embeddings. On condition that the same ID restoration after object re-appearance is of interest, the utilization of larger D_{max} values makes an obvious sense. In our experiments, we kept $D_{max} \leq 100$, as further increases were not related to any accuracy gains. Such settings are also quite usual in DBT solutions[19, 40].

4 Results

The evaluation results for our tracking solution on the testset of KITTI Tracking Benchmark[4] are presented in Table 3 along with the top-performing submissions. In this particular case, the following settings were applied: the Kalman filter was switched off, D_{max} was set to 1, and λ in Eq. 4 was equal to 0.1. HSV64/128 model was employed as the embedding function. It is evident from this table that our trackers achieve state-of-the-art performance in both leaderboards. They outperform other methods according to MODA metric while demonstrating reasonable ID Switch (IDs) counts. In order to make direct comparison between ODESA and a number of embeddings that often find use in MOT domain our DBT solution was adjusted to accept the latter as alternative options. In this case, the influence of the detection and data association stages becomes separated. The results are summarized in Table 4. Here each entry corresponds to the case, where the model listed in the first column was employed in our DBT solution as the embedding function. All person Re-ID models were trained on Market-1501[58] dataset, which is a common practice

[2] The leaderboard state corresponds to July 7, 2020.

Table 3: Top-performing solutions from KITTI Tracking Benchmark leaderboards.[2] MOTA is being the main evaluation criterion.

Target	Method	TP ↑	FP ↓	FN ↓	MODA ↑	IDs ↓	FRAG ↓	MOTA ↑
Car	TuSimple[55]	34'322	705	3'602	87.48	293	501	86.62
	IWNCC	34'146	571	3'819	87.24	130	521	86.86
	EagerMOT	32'858	1'209	3'173	87.26	31	472	87.17
	FG-3DMOT	34'052	611	3'491	88.07	**20**	**117**	88.01
	RE3T	34'991	785	3'005	88.98	31	193	88.89
	CenterTrack	**36'562**	849	**2'666**	89.78	116	334	89.44
	ODESA (Our)	36'258	451	2'887	**90.29**	90	501	**90.03**
Pedestrian	3D-TLSR	13'767	**942**	9'606	54.44	100	835	54.00
	CenterTrack	15'351	2'196	8'047	55.75	**95**	**751**	55.34
	VV_team	15'640	2'366	7'757	56.27	201	1'131	55.40
	Quasi-Dense[56]	14'925	1'284	8'460	57.91	254	1'121	56.81
	TuSimple[55]	14'936	1'192	8'359	58.74	138	818	58.15
	HWFD	17'296	1'302	6'159	67.77	116	918	67.27
	ODESA (Our)	**17'516**	991	**5'791**	**70.70**	191	1'070	**69.88**

Table 4: Comparison with state-of-the-art person Re-ID embeddings on 21 KITTI tracking training sequences for *Car* and *Pedestrian* targets. The association is restricted solely to visual similarity, i.e. $\lambda = 0$ in Eq. 4.

Method	Car				Pedestrian			
	MODA ↑	IDs ↓	FRAG ↓	MOTA ↑	MODA ↑	IDs ↓	FRAG ↓	MOTA ↑
SOSNet[36]	89.84	47	304	89.65	68.81	98	540	67.93
HardNet++[27]		43	298	89.66		93	537	67.97
OSNet[32]		65	317	89.57		**78**	**523**	**68.11**
MLFN[57]		57	312	89.61		88	532	68.02
HACNN[33]		50	305	89.63		81	525	68.08
AGW[30]		51	308	89.63		81	522	68.08
ODESA(GRAY64/128)		**31**	**288**	**89.71**		88	530	68.02
ODESA(HSV64/128)		35	292	89.70		84	527	68.05
ODESA(HSV64/512)		40	296	89.69		81	525	68.08

for MOT solutions. The remaining models were trained on HPatches[26]. The detects employed for these experiments correspond to the detector entries, which achieve the highest MODA values for a given target in Table 1. Let us consider IDs count, which appears to be more discriminative compared to Fragmentation (FRAG). Our most universal target-wise model, which is HSV64/128, with 84 IDs is about 8% behind OSNet[32]. The latter with 78 switches turns out to be the best-performing embedding for *Pedestrian* target. At the same time HACNN[33], which has the lowest IDs count for *Car* target, is outperformed by the same ODESA model by 43%. Also for both targets, our models show consistently lower IDs counts compared to the contemporary LLFD models represented by HardNet++[27] and SOSNet[36]. These results are also consistent with the retrieval experiments summarized in Table 2. ODESA achieves such accuracy level while operating with smaller input patches, shorter embedding vector, and exhibiting faster inference time compared to the person Re-Id models.[3]

[3] See the processing time comparison in the supplementary material.

Table 5: The results of the top-performing solutions from CVPR19 Tracking Challenge Leaderboard. MOTA serves as the main evaluation criterion.

Method	MOTA ↑	IDF1 ↑	FP ↓	FN ↓	IDs ↓	FRAG ↓
IITB_trk[2]	45.5	43.6	23'931	278'042	3'002	5'478
Aaron[2]	46.5	46.6	40'676	256'671	**2'315**	**2'968**
V_IOU[61]	46.7	46.0	33'776	261'964	2'589	4'354
DD_TAMA19[62]	47.6	48.7	38'194	252'934	2'437	3'887
TracktorCV[60]	51.3	47.6	**16'263**	253'680	2'584	4'824
ODESA[1] (Our)	**54.9**	**52.7**	24'609	225'292	2'614	4'322
ODESA[2] (Our)	54.8	52.2	33'814	**215'572**	3'750	5'493

For further validation, our ODESA-based tracking solution was adjusted to the conditions of CVPR'2019 Tracking Challenge[59]. The modifications accounted for the requirement of relying exclusively on the set of public detects. In particular, we prepared a detector based on ResNet-101[14], FPN[13], and Regression Cascade[16] using the training protocol described in Section 3.1. RPN[11] block was then stripped from the detector. Its output was replaced with either the items from the public set of detects, which correspond to earlier frames, or the detects derived from them in the manner described in Ref. [60]. However, no data association was performed at this point. HSV64/128 was employed as the embedding function. The history depth D_{max} was set to 100. The matching was performed in two stages. At the first one λ in Eq. 4 was set to zero. It accounted for about 97% of the associations. At the second one λ was set to 1 and the Kalman filter was turned off. The results of our two submissions, which differ solely by the rejection routine settings at the detect refinement stage, are presented in Table 5. Both show comparable performance, while the last table entry turned out to be the challenge winner.

The videos representing the output of our tracker for the benchmarks mentioned above have been made available for visual inspection.

5 Discussion

5.1 ODESA Object Representation

Let us check the assumptions which were made at the beginning of Section 3 about the object representation expected from ODESA. For this purpose, we utilize ALOI[63] dataset. It contains the images for 1000 objects taken on a rotating stage every 5° against a uniform background. Due to such restricted angle increment, each set of images could be regarded as representing gradual appearance evolution. Uniform background and absent occlusions also help to model a simplified version of MOT environment. A few samples from ALOI dataset are depicted in the top row of Fig. 2. For each object, a set of 72 embedding vectors was extracted by HSV64/128 model. t-SNE projections for corresponding manifolds are shown in Fig. 2(a), where the original ALOI object number to color correspondence is provided as well. Our assumption about the non-compact nature of ODESA object representation can be readily checked via the estimation

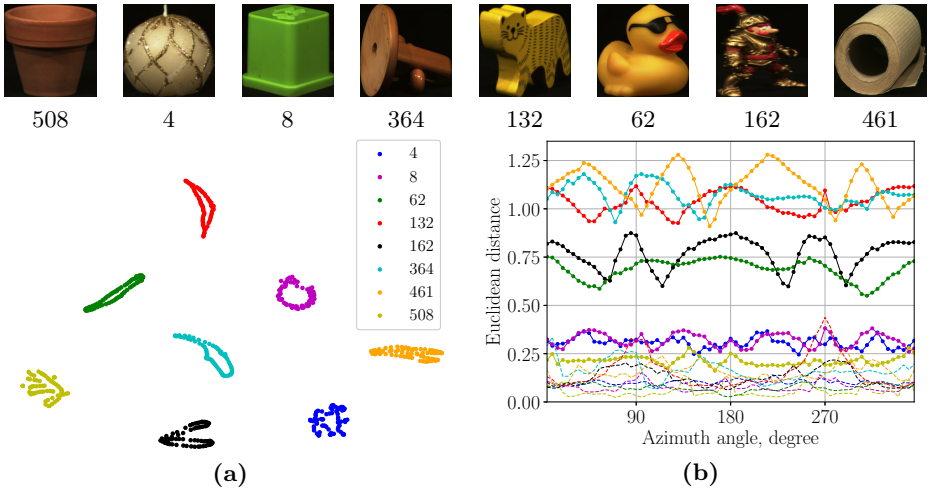

Fig. 2. Object mapping into the metric space for ODESA(HSV64/128) model. (a) t-SNE projection for all embedding vectors corresponding to the objects depicted at the top. (b) (*Solid*) and (*dashed lines*) represent the distances from a given sample embedding vector to its furthest element of the same manifold and its two closest angle-wise neighbors halved, respectively. Best viewed in color.

of the corresponding manifold extent. Due to ODESA embedding vector normalization, i.e. projection to the unit hypersphere, the maximal extent is limited to the value of two. It can be deduced from the solid lines in Fig. 2(b), which represent the distance from the embedding vector corresponding to a given azimuth angle to its most distant member of the same manifold. These curves indicate that for one of the examined objects the maximal extent exceeds the value of 1.25. At the same time, rather compact manifolds could also be observed, on condition that the appearance of objects does not vary much within considered set of images. This is the case of objects #508, 4, and 8. The distribution of the object manifold extent across the entire ALOI dataset is shown in Fig. 4 for HardNet++ and HSV64/128 models. This figure confirms that for the majority of objects the extent is significant, thus validating our initial assumption. It is also worthwhile to note that while t-SNE projection does not reflect the manifold extent, together with the solid lines in Fig. 2(b) it is indicative of the symmetry exhibited by the objects #8, 132, 162, and 461.

To estimate the manifold continuity, the average distance from each embedding vector to its two closest angle-wise neighbors was calculated. It is shown in Fig. 2(b) by the dashed lines as a function of azimuth angle. We assume that rotation by 5° shall not affect the object appearance significantly. Therefore, any discontinuity is expected to manifest itself as a sharp peak in such a curve. Our data indicate that the distance to neighbors is rather uniformly distributed and scales with the manifold extent. The former conclusion is also supported by the t-SNE projection. These observations are rather supportive of our assump-

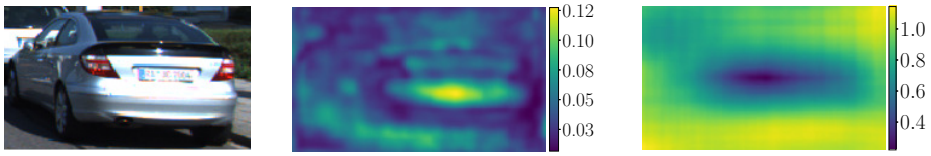

Fig. 3. The influence of occlusions and the bounding box misalignment on embedding vector: (a) image patch, (b) D_o, and (c) D_m as the function of δx and δy. The occluder size s amounts 10% of the patch height h. Best viewed in color.

tion about the non-abrupt nature of ODESA mapping. The utilization of color information tends to scale down the manifold extent, as shown in Fig. 4, and the distance between the angle-wise neighbors. The evolution of these properties from LLFD to ODESA models is discussed in the supplementary material.

5.2 Sensitivity to Occlusions, Background and Detection Noise

In Section 1, we mentioned the factors which cast doubts on LLFD utilization as global object features. A few of them are related to the nature of object bounding boxes. Namely, they usually contain a certain amount of background, could be occluded and could deviate from the ground truth shape and position due to the detection noise. The first two factors do not take place in LLFD domain. The sensitivity of ODESA models to them could be estimated by examining the influence of an occlusion applied in the sliding window manner. As a measure of such influence, one could use the distance between the embedding vector corresponding to the original patch $P(x, y, h, w)$ and those originating from occluded ones. Such distance could be calculated as the function of the occluder position

$$D_o(\delta x, \delta y) = \left\| f\left(P(x, y, h, w)\right) - f\left(P(x, y, h, w) \odot T_{\delta x, \delta y}\left(O(x, y, s, s)\right)\right)\right\|_2, \tag{6}$$

where $f(\cdot)$ - embedding function, $\odot$ designate pixel-wise replacement operation, $T_{\delta x, \delta y}$ is a translation operation, $O(x, y, s, s)$ - the square occluding patch of size s filled with noise. An example of such estimation is shown in Fig. 3, where (a) depicts the image area within the ground truth bounding box, (b) contains $D_o(\delta x, \delta y)$. In this case, the occluder size s amounted as much as 10% of the patch height h. Figure 3(b) indicates that the modification of the patch periphery does not result in any considerable displacement of embedding vectors in the metric space. At the same time, the central part of the patch plays a rather significant role. Figure 5 represents our attempt to generalize these conclusions. It shows D_o averaged over 198 and 149 detects from KITTI dataset belonging to pedestrian and car categories, respectively. The detects of the highest resolution with unique IDs were selected for this purpose. To eliminate the aspect ratio, scale, and D_o magnitude differences, at first, relative D_o for individual detects was calculated. Then the patch area was split into 10 bands. Each band was limited to the area between adjacent rectangles resulted from the bounding box

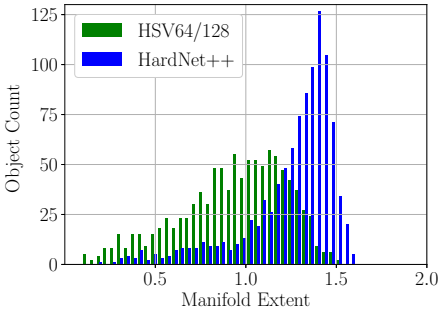

Fig. 4. Distribution of manifold extent across the entire ALOI dataset for HSV64/128 and HardNet++ models.

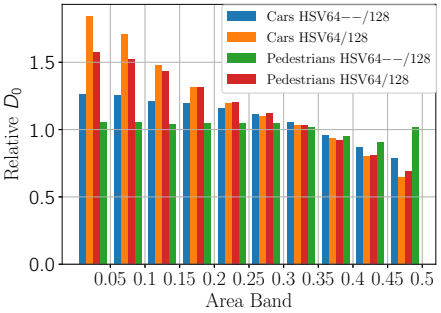

Fig. 5. Relative $D_o(\delta x, \delta y)$ averaged over a number of detects from KITTI dataset for HSV64/128 and HSV64--/128 models.

scaling with a decrement of 5% height- and width-wise. Finally, the relative intensity was averaged over each band across all detects and shown in Fig. 5. This procedure was performed for HSV64/128 and HSV64--/128 models. The latter was learned while the random patch transformations and flips were both turned off. The data from Fig. 5 indicate that these options contribute to making the embedding models considerably less susceptible to the patch periphery.

Finally, the influence of the detection noise could be estimated by examining the embedding vector displacement due to the bounding box misalignment with its ground truth position

$$D_m(\delta x, \delta y) = \|f\left(P\left(x, y, h, w\right)\right) - f\left(P\left(x + \delta x, y + \delta y, h, w\right)\right)\|_2. \tag{7}$$

An example of $D_m(\delta x, \delta y)$ produced by means of HSV64/128 model is shown in Fig. 3(c). The corresponding distribution of distances is quite shallow for $(\delta x, \delta y)$ values up to about 20% of the bounding box size. The displacement values tend to grow as the misalignment increases. Such a picture appears to be typical for the majority of examined samples. Fig. 3(c) indicates that ODESA models, as well as LLFD ones, could put up with a certain level of the detection noise. In the supplementary material, we indicate that the random patch transformations applied during dataset preparation[26] have a profound effect on $D_m(\delta x, \delta y)$. By controlling their strength, one could either get a model with a better generalization or higher discriminative capability.

6 Conclusions

Our study shows that starting with LLFD it is possible to derive a class-agnostic embedding function, which being deployed as a part of DBT solution is capable of achieving competitive results in MOT domain. It produces meaningful object manifolds with predictable properties. Corresponding object representation turns out to be compatible with the association stages of contemporary DBT solutions. For this reason, ODESA could be readily deployed with about any DBT solution.

References

1. Grant, J.M., Flynn, P.J.: Crowd scene understanding from video: A survey. ACM Trans. Multimedia Comput. Commun. Appl. **13** (2017)
2. Choi, W., Savarese, S.: A unified framework for multi-target tracking and collective activity recognition. In Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C., eds.: Computer Vision – ECCV 2012, Springer Berlin Heidelberg (2012) 215–230
3. Weiming Hu, Tieniu Tan, Liang Wang, Maybank, S.: A survey on visual surveillance of object motion and behaviors. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) **34** (2004) 334–352
4. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The KITTI vision benchmark suite. In: Conference on Computer Vision and Pattern Recognition (CVPR). (2012)
5. Milan, A., Leal-Taixé, L., Reid, I., Roth, S., Schindler, K.: MOT16: A benchmark for multi-object tracking. arXiv preprint arXiv:1603.00831 (2016)
6. Wen, L., Du, D., Cai, Z., Lei, Z., Chang, M., Qi, H., Lim, J., Yang, M., Lyu, S.: UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking. arXiv preprint arXiv:1511.04136 (2015)
7. Dollár, P., Appel, R., Belongie, S., Perona, P.: Fast feature pyramids for object detection. Pattern Analysis and Machine Intelligence, IEEE Transactions on **36** (2014) 1532–1545
8. He, K., Gkioxari, G., Dollár, P., Girshick, R.B.: Mask R-CNN. arXiv preprint arXiv:1703.06870 (2017)
9. Li, Y., Qi, H., Dai, J., Ji, X., Wei, Y.: Fully convolutional instance-aware semantic segmentation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017)
10. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S.E., Fu, C., Berg, A.C.: SSD: single shot multibox detector. arXiv preprint arXiv:1512.02325 (2015)
11. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. arXiv preprint arXiv:1506.01497 (2015)
12. Tan, M., Pang, R., Le, Q.: EfficientDet: Scalable and efficient object detection. arXiv preprint arXiv:1911.09070 (2019)
13. Lin, T., Dollár, P., Girshick, R.B., He, K., Hariharan, B., Belongie, S.J.: Feature Pyramid Networks for Object Detection. arXiv preprint arXiv:1612.03144 (2016)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385 (2015)
15. Xie, S., Girshick, R.B., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. arXiv preprint arXiv:1611.05431 (2016)
16. Cai, Z., Vasconcelos, N.: Cascade R-CNN: high quality object detection and instance segmentation. arXiv preprint arXiv:1906.09756 (2019)
17. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. arXiv preprint arXiv:1703.06211 (2017)
18. Zhu, X., Hu, H., Lin, S., Dai, J.: Deformable ConvNets v2: More deformable, better results. arXiv preprint arXiv:1811.11168 (2018)
19. Luo, W., Xing, J., Milan, A., Zhang, X., Liu, W., Zhao, X., Kim, T.K.: Multiple object tracking: A literature review. arXiv preprint arXiv:1409.7618v4 (2017)
20. Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B.: Simple online and realtime tracking. In: 2016 IEEE International Conference on Image Processing (ICIP). (2016) 3464–3468

21. Yu, F., Li, W., Li, Q., Liu, Y., Shi, X., Yan, J.: POI: Multiple object tracking with high performance detection and appearance feature. In: IEEE European Conference on Computer Vision (ECCV). (2016)
22. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The KITTI dataset. International Journal of Robotics Research (IJRR) (2013)
23. Gray, D., Brennan, S., Tao, H.: Evaluating appearance models for recognition, reacquisition, and tracking. In: In IEEE International Workshop on Performance Evaluation for Tracking and Surveillance, Rio de Janeiro. (2007)
24. Zheng, L., Zhang, H., Sun, S., Chandraker, M., Tian, Q.: Person re-identification in the wild. arXiv preprint arXiv:1604.02531 (2016)
25. Wang, Z., Zheng, L., Liu, Y., Wang, S.: Towards real-time multi-object tracking. arXiv preprint arXiv:1909.12605v1 (2019)
26. Balntas, V., Lenc, K., Vedaldi, A., Mikolajczyk, K.: HPatches: A benchmark and evaluation of handcrafted and learned local descriptors. In: CVPR. (2017)
27. Mishchuk, A., Mishkin, D., Radenovic, F., Matas, J.: Working hard to know your neighbor's margins: Local descriptor learning loss. arXiv preprint arXiv:1705.10872 (2017)
28. Tian, Y., Fan, B., Wu, F.: L2-Net: Deep learning of discriminative patch descriptor in euclidean space. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017) 6128–6136
29. Vassileios Balntas, Edgar Riba, D.P., Mikolajczyk, K.: Learning local feature descriptors with triplets and shallow convolutional neural networks. In Richard C. Wilson, E.R.H., Smith, W.A.P., eds.: Proceedings of the British Machine Vision Conference (BMVC), BMVA Press (2016) 119.1–119.11
30. Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., Hoi, S.C.H.: Deep learning for person re-identification: A survey and outlook. arXiv preprint arXiv:2001.04193 (2020)
31. Winder, S., Brown, M.: Learning local image descriptors. In: CVPR. (2007)
32. Zhou, K., Yang, Y., Cavallaro, A., Xiang, T.: Omni-Scale feature learning for person re-identification. In: ICCV. (2019)
33. Li, W., Zhu, X., Gong, S.: Harmonious attention network for person re-identification. arXiv preprint arXiv:1802.08122 (2018)
34. Song, C., Huang, Y., Ouyang, W., Wang, L.: Mask-guided contrastive attention model for person re-identification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2018)
35. Keller, M., Chen, Z., Maffra, F., Schmuck, P., Chli, M.: Learning deep descriptors with scale-aware triplet networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2018)
36. Tian, Y., Yu, X., Fan, B., Wu, F., Heijnen, H., Balntas, V.: SOSNet: Second order similarity regularization for local descriptor learning. In: CVPR. (2019)
37. Zhang, L., Rusinkiewicz, S.: Learning local descriptors with a CDF-based dynamic soft margin. In: International Conference on Computer Vision (ICCV). (2019)
38. Zhang, X.Y., Zhang, L., Zheng, Z.Y., Liu, Y., Bian, J.W., Cheng, M.M.: AdaSample: Adaptive sampling of hard positives for descriptor learning. arXiv preprint arXiv:1911.12110 (2019)
39. Zhang, X., Yu, F.X., Kumar, S., Chang, S.F.: Learning spread-out local feature descriptors. arXiv preprint arXiv:1708.06320 (2017)
40. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: 2017 IEEE International Conference on Image Processing (ICIP), IEEE (2017) 3645–3649
41. Cao, Y., Xu, J., Lin, S., Wei, F., Hu, H.: GCNet: Non-local networks meet squeeze-excitation networks and beyond. arXiv preprint arXiv:1904.11492 (2019)

42. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., Dai, J., Wang, J., Shi, J., Ouyang, W., Loy, C.C., Lin, D.: MMDetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)
43. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M.S., Berg, A.C., Li, F.: ImageNet large scale visual recognition challenge. arXiv preprint arXiv:1409.0575 (2014)
44. Pang, J., Chen, K., Shi, J., Feng, H., Ouyang, W., Lin, D.: Libra R-CNN: towards balanced learning for object detection. arXiv preprint arXiv:1904.02701 (2019)
45. Yu, F., Xian, W., Chen, Y., Liu, F., Liao, M., Madhavan, V., Darrell, T.: BDD100K: A diverse driving video database with scalable annotation tooling. arXiv preprint arXiv:1805.04687 (2018)
46. Vedaldi, A., Fulkerson, B.: VLFeat: An open and portable library of computer vision algorithms (2008)
47. Ono, Y., Trulls, E., Fua, P., Yi, K.M.: LF-Net: Learning local features from images. arXiv preprint arXiv:1805.09662 (2018)
48. Dusmanu, M., Rocco, I., Pajdla, T., Pollefeys, M., Sivic, J., Torii, A., Sattler, T.: D2-Net: A trainable CNN for joint detection and description of local features. arXiv preprint arXiv:1905.03561 (2019)
49. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In Teh, Y.W., Titterington, D.M., eds.: AISTATS. Volume 9 of JMLR Proceedings., JMLR.org (2010) 249–256
50. Dollár, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: A benchmark. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2009)
51. Xiao, T., Li, S., Wang, B., Lin, L., Wang, X.: Joint detection and identification feature learning for person search. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017)
52. Kuhn, H.W.: The Hungarian Method for the assignment problem. Naval Research Logistics Quarterly **2** (1955) 83–97
53. Kalman, R.: A new approach to linear filtering and prediction problems. Journal of Basic Engineering **82** (1960) 35–45
54. Welch, G., Bishop, G.: An Introduction to the Kalman filter. University of North Carolina at Chapel Hill, Chapel Hill, NC (1995)
55. Choi, W.: Near-online multi-target tracking with aggregated local flow descriptor. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 3029–3037
56. Pang, J., Qiu, L., Chen, H., Li, Q., Darrell, T., Yu, F.: Quasi-dense instance similarity learning. arXiv:2006.06664 (2020)
57. Chang, X., Hospedales, T.M., Xiang, T.: Multi-level factorisation net for person re-identification. arXiv preprint arXiv:1803.09132 (2018)
58. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: Computer Vision, IEEE International Conference on. (2015)
59. Dendorfer, P., Rezatofighi, S.H., Milan, A., Shi, J., Cremers, D., Reid, I.D., Roth, S., Schindler, K., Leal-Taixé, L.: CVPR19 tracking and detection challenge: How crowded can it get? arXiv preprint arXiv:1906.04567 (2019)
60. Bergmann, P., Meinhardt, T., Leal-Taixé, L.: Tracking without bells and whistles. arXiv preprint arXiv:1903.05625 (2019)

61. Bochinski, E., Eiselein, V., Sikora, T.: High-speed tracking-by-detection without using image information. In: International Workshop on Traffic and Street Surveillance for Safety and Security at IEEE AVSS 2017, Lecce, Italy (2017)
62. Yoon, Y., Kim, D.Y., Yoon, K., Song, Y., Jeon, M.: Online multiple pedestrian tracking using deep temporal appearance matching association. arXiv preprint arXiv:1907.00831 (2019)
63. Geusebroek, J.M., Burghouts, G.J., Smeulders, A.W.M.: ALOI: Amsterdam library of object images. International Journal of Computer Vision **61(1), 103-112** (2005)
64. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person reidentification. arXiv preprint arXiv:1703.07737 (2017)