re-OBJ: Jointly Learning the Foreground and Background for Object Instance Re-identification

Vaibhav Bansal[1,3(✉)], Stuart James[2], and Alessio Del Bue[1]

[1] Visual Geometry and Modelling (VGM),
Istituto Italiano di Tecnologia, Genoa, Italy
vaibhav.bansal@iit.it
[2] Center for Cultural Heritage Technology (CCHT),
Istituto Italiano di Tecnologia, Genoa, Italy
[3] Università degli studi di Genova, Genoa, Italy

Abstract. Conventional approaches to object instance re-identification rely on matching appearances of the target objects among a set of frames. However, learning appearances of the objects alone might fail when there are multiple objects with similar appearance or multiple instances of same object class present in the scene. This paper proposes that partial observations of the background can be utilized to aid in the object re-identification task for a rigid scene, especially a rigid environment with a lot of reoccurring identical models of objects. Using an extension to the Mask R-CNN architecture, we learn to encode the important and distinct information in the background jointly with the foreground relevant to rigid real-world scenarios such as an indoor environment where objects are static and the camera moves around the scene. We demonstrate the effectiveness of our joint visual feature in the re-identification of objects in the ScanNet dataset and show a relative improvement of around 28.25% in the rank-1 accuracy over the deepSort method.

Keywords: Re-identification · Object detection · Multi-view · Triplet loss

1 Introduction

Multiple object matching and association are classical problems in many important tasks such as video surveillance, semantic scene understanding and also, Simultaneous Localization And Mapping (SLAM). Given an indoor scene, where the environment is frequently cluttered with several near-identical objects, it is challenging to identify and track a particular instance of an object among a number of objects present in the scene, e.g. see Fig. 1. The problem is even more challenging when there is a wide baseline among multiple views (or temporally disjoint). It is complex to re-identify a vast variety of objects based on appearance only. There are many challenges for the association problem i.e. occlusions, motion blur, mis-detections, etc. Conventional methods use two major

E. Ricci et al. (Eds.): ICIAP 2019, LNCS 11752, pp. 402–413, 2019.
https://doi.org/10.1007/978-3-030-30645-8_37

approaches to build a re-ID system - appearance-based and motion-based. Most methods use an appearance-based approach because motion prediction based systems try to localize each object instance based on a motion model, however, due to the possibility of huge unpredictable trajectories across the frames, these methods tend to fail when the same object instance reappear after a long time.

Fig. 1. Similar looking objects in rigid, indoor scenes from ScanNet dataset. Multiple instances of the same object class, chair, in this case, are hard to differentiate with each other. In such cases, background can be highly useful to re-identify a particular instance in multiple views.

Many previous studies focus on *person* re-identification where the goal is to assign a correct ID of an instance of a specific class (i.e. a pedestrian) across multiple-views obtained from cameras with possibly non-overlapping views. In general, these methods try to learn discriminative features based on person's face [18], clothing [14] or symmetry-driven local features [9] to re-ID people. In contrast, the problem of associating an unique ID to instances of objects is often solved as the association of multiple unknown objects between views [16]. This problem is closely related to person re-ID and often evaluated in the pedestrian (person) scenario with early work on PET2009 [5].

However, the specific task of re-identifying multiple near-identical objects in a rigid scene presents a different challenge, we refer to as re-OBJ, a specific case of re-ID. In this paper, we consider a static indoor video dataset where large displacement in the camera motion is unlikely and so the background of an instance cannot undergo a sudden drastic change. Therefore, we propose to jointly learn the foreground and the background to build a robust object re-identification system at the instance level. We propose not only to learn the appearance of an object but also the background that can provide a lot of useful information regarding the surroundings of an instance which is unique to that instance at any given viewpoint. Consider a scene of an office room with multiple chairs and tables present. To re-identify a particular object instance across multiple images, it is important to be able to distinguish it from other instances of the

same object class. Intuitively, if we can observe and encode the surroundings of that particular instance within a stream of images, we can be confident to an extent that the object instance in consideration has been seen before and it is different from other instances of the same class because the environment around it is unique at any given point of time even when other instances have similar appearance (see Fig. 1).

2 Related Work

There is a vast literature for object re-identification that is mostly focused on person re-identification. The ability to re-identify objects in the images heavily relies on finding a similar set of images for a given image of the target object, possibly with multiple instances, using visual search to retrieve similar images to the given query image. Some works in the literature like [2,9] exploit the knowledge that the same individual is been detected in consecutive frames and then learning an appearance-based transfer function for a robust re-identification system. Additionally, in [9], they extract features from three different complementary modalities: the chromatic content, spatial arrangement of colors and local motifs derived from different parts of the human body to accumulate local features. Other deep learning models learn the category-level similarity [20] that mainly involves semantic similarity. The study highlights the effect of significant visual variability within a category although the semantic and visual similarities are generally quite consistent across different categories. Thus, applications that involve the computation of image similarity like re-identification, image retrieval, search-by-example require learning a fine-grained image similarity that can also distinguish the differences between different images of the same category. Relative attribute [17] learns image attribute ranking among the images with the same attributes. OASIS [4] performs local distance learning [10] learn image similarity ranking models on top of the hand-crafted features. Such appearance-based approaches are good at distinguishing intra-class variation, in contrast, we focus on the objects' relationship to the background to jointly learn a foreground and background discriminative appearance feature.

Many image similarity models [3,4,20] simply extract features like Gabor filters, SIFT [15], HOG [7] features to learn similarity between images. However, the representation of the hand-crafted features limits the performance of these methods. Some deep learning-based models popular in image classification tasks [13] have shown great success in learning features from the images but these models cannot directly fit similar image ranking especially the fine-grained distinction between similar images. Thus, in order to learn the fine-grained image similarity deep ranking model has been proposed by [21]. Pairwise ranking model is a widely used learning-to-rank formulation. It is used to learn image ranking models in [4,10,17]. Generating good triplet samples is a crucial aspect of learning pairwise ranking model. FaceNet [18] showed that the triplet loss is a suitable loss function for the verification, recognition and clustering than the verification loss [19]. The difference is that the verification loss minimizes the $L2$-distance

between objects of the same identity and enforces a margin between the distance of objects of different identities whereas the triplet loss also encourages a relative distance constraint and thus, enhancing the ability to discriminate between dissimilar identities. In [4] and [17], the triplet sampling algorithms assume that we can load the whole dataset into memory, which is impractical for a large dataset. Our work is built upon the deep ranking model proposed by [21] with an efficient triplet sampling algorithm that does not require loading the whole dataset into the memory.

3 Object Instance Separation Encoding

For a robust object re-identification system for a rigid scenario, we hypothesize that the background information is useful in order to discriminate between multiple instances of the same semantic class and also the objects that have a similar appearance as shown intuitively in Fig. 1. To include the background information, the first step in our approach is to use an off-the-shelf object detector, i.e. Mask-RCNN (Sect. 3.1), and obtain foreground masks of the objects with the bounding boxes that are expanded (see Sect. 4) in order to include a substantial background around the object within the bounding boxes. Encodings from the separated masked foregrounds and the masked backgrounds are extracted using ResNet50 (Sect. 3.2), which are concatenated to obtain joint embeddings. These embeddings then are sampled into triplets $\{positive, negative, anchor\}$ and fed to a triplet-based network architecture consisting of three identical ConvNets (see Figs. 3 and 4) with the pairwise ranking model to learn image similarity for a triple-based ranking loss function.

3.1 Object Detection

Our approach relies on previous work, Mask-RCNN [11] which uses region-based object detector like Faster R-CNN to detect objects. It does not only provide a bounding box around an object but also performs image segmentation and provides a mask representing a set of pixels belonging to the same object. A Region Proposal Network (RPN) is used to generate a number of region proposals followed by a position-sensitive RoI pooling layer to warp them into a fixed dimension. Finally, it is fed into fully-connected layers to produce class scores and the bounding box predictions. A parallel branch of two additional fully-connected layers provides the mask. Using the output from the Mask-RCNN, we extract each bounding box including masks as separate images and resize them into images of a fixed size in order to train our network to learn a visual encoding of the objects' mask and the background surrounding them within the bounding boxes (see first column, Fig. 2).

3.2 Object Visual Encoding

For each object of the input images, we create two sets of images $F = \{I_f, I_b\}$. Using the detections obtained from Mask-RCNN, one set is created by extract-

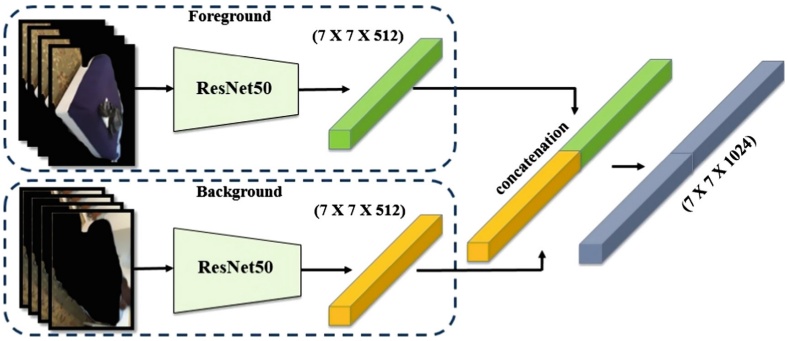

Fig. 2. As input, our network takes expanded bounding boxes (see Sect. 3.2) which construct a pair of images for masked foreground and masked background (seen on left of image). Each of the pair of images is passed through a ResNet50 where we take an intermediary representation $7 \times 7 \times 512$ providing spatial information, which is concatenated to provide a joint representation of $7 \times 7 \times 1024$.

ing masks representing objects in the foreground (I_f). The other set only contains the background with the subtracted foreground (I_b). As shown in Fig. 2, a pair of images is taken from each set to pass through two identical streams to learn an encoding between the masked foreground and the background. Each of the images, the masked background and the masked foreground is input to a ResNet50 [12] deep model pre-trained on ImageNet [8] dataset to extract the features. We take from an intermediary layer of the network providing $I_{(.)} \in \mathbb{R}^{7\times7\times512}$ representation of the two images retaining spatial context, the tensors are then concatenated to provide an embedding $F \in \mathbb{R}^{7\times7\times1024}$.

3.3 Triplet Loss

An effective algorithm for object instance re-identification should be able to distinguish not only between the images of different objects but also between different instances of the same object class. Especially, in the indoor scenes where multiple instances of the same object category are present, i.e. an office with multiple tables and chairs; it is highly challenging to re-identify a particular object instance amongst others.

A triplet of images has three kinds of images: an *anchor* which acts like a query template, a *positive* and a *negative* image. In order to ensure an effective re-identification at the instance level, it is important to also consider the intra-class variations and different instances of the same object as negative examples. For example, a backpack and a chair are definitely an example of *anchor-negative* pairs but two different instances of the same chair (with a different background) should also be considered an *anchor-negative* pair. We use a triplet-based network architecture with the pairwise ranking model to learn image similarity for the triple-based ranking loss function, inspired from [21]. If we have a set of $F = f_1, ...f_F$ images and $s_{i,j} = s(f_i, f_j)$ that gives the pairwise similarity score

between the images f_i and f_j. The score s is higher for more similar images and is lower for more dissimilar images. If we have a triplet $t_i = (f_{iA}, f_{iP}, f_{iN})$ where f_{iA}, f_{iP} and f_{iN} are the anchor, positive and negative images, respectively. The goal of the training is to learn an embedding function such that:

$$D(f_{iA}, f_{iP}) < D(f_{iA}, f_{iN}), s(f_{iA}, f_{iP}) > s(f_{iA}, f_{iN}) \quad (1)$$

where $D(.)$ is the squared Euclidean distance in the embeddings space. A triplet incorporates a relative ranking based on the similarity between the anchor, positive and the negative images.

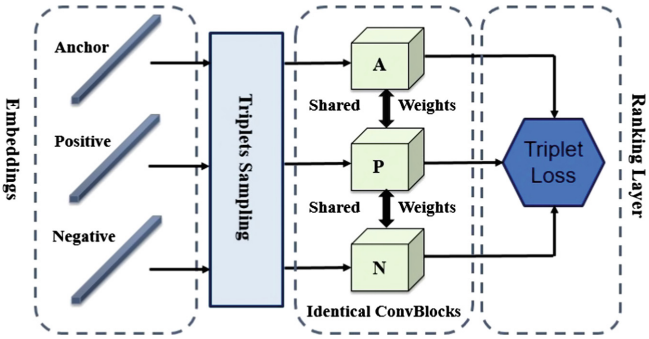

Fig. 3. Triplet of input tensors corresponding to images. Each tensor contains an embeddings of the anchor image A, positive image P and a negative Image N which are fed into three identical deep neural networks independently with shared weights where the triplet loss is optimized.

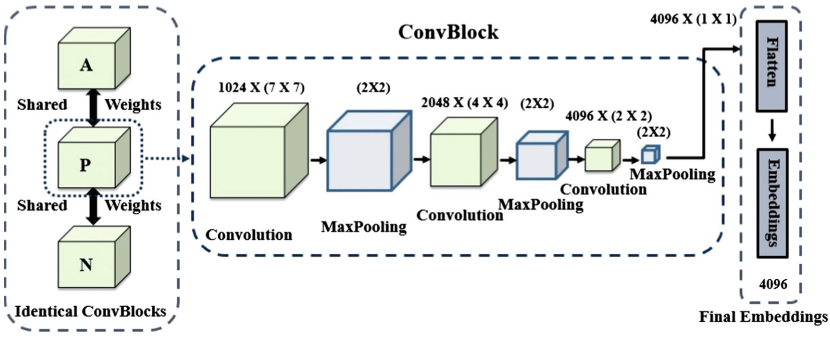

Fig. 4. Our ConvBlock takes in the encoding from Fig. 2. The ConvBlock consists of a network of convolutional and maxpooling layers, which pool the spatial information and merge the foreground and background encodings to obtain final embeddings.

The triplet ranking loss function is given as:

$$l(f_{iA}, f_{iP}, f_{iN}) = max\{0, M + D(f_{iA}, f_{iP}) - D(f_{iA}, f_{iN})\} \quad (2)$$

where M is a parameter called *margin* that regulates the gap between the pairwise distance: (f_{iA}, f_{iP}) and (f_{iA}, f_{iN}). The model learns to minimize the distance between more similar images and maximize the distance between the dissimilar ones. Our model is based on the work proposed in [21] with the difference that the input image triplets we use are the concatenated embeddings of the masked foregrounds and backgrounds.

4 Experiments

Training Data. We use ScanNet dataset [6] for our experiments which consists of 1500 indoor RGBD scans annotated with 3D camera poses, surface reconstructions, and mesh segmentation related to several object categories. These annotations allowed us to evaluate the accuracy of Mask-RCNN on the ScanNet images to be used in the proposed pipeline. To generate our training data, we ran Mask-RCNN over a subset of 863 scenes randomly selected from the whole ScanNet dataset. In total, the Mask-RCNN provided 646, 156 object detections with masks belonging to 29 object classes (see Table 1). Since not all the objects in the dataset are annotated, we computed the bounding box overlap ratio between the ground truth (GT) bounding boxes and the detections provided by Mask-RCNN to select only the *valid* detections. If the overlap ratio was higher than 60% and the label of the detected object matches with the GT label, it was considered a *valid* detection.

After mapping each detection obtained from the Mask-RCNN with the corresponding 2D ground truth (GT), we found 9.11% of the total, i.e. around 58876 detections to be considered fit for the experiments. The regions indicated by the bounding boxes were extended by an additional 10 pixels-wide border in order to allow loosely-fitted bounding boxes around the objects and thus, allowing a more significant background around each object's mask within the bounding boxes. These regions were then extracted out of the full images, resized to 224×224 and categorically stored based on the object's class and it's observed instances. Finally, for each object image, the foreground masks and the background masks were extracted and stored as separate images. The data is split into a 3-fold cross-validation manner with 39250 images for training and 19626 images for test over 1701 instances of objects.

We performed our experiments in three different setups. In all the experimental setups, we used pre-trained ResNet50 [12] on the ImageNet [8] dataset as the backbone model to extract features from the images of the objects. **no-train:** In this setup, the features extracted from full images were matched against each other by using an $L2$ distance-based metric, without any training. **full:** In another setup, our model is trained on the embeddings obtained using the full images without extracting separate foreground and background masks. **concat:** The third type of experimental setup is the approach proposed in this paper

Table 1. Number of views after mapping with GT for *valid* detections, selected based on object's label and the bounding box overlap ratio and the number of unique instances for each object category.

No. of views and unique instances per object class					
Class	No. of views	No. of instances	Class	No. of views	No. of instances
Bicycle	110	6	Toilet	1755	103
Bench	27	4	Tv	562	46
Backpack	1563	117	Laptop	600	41
Handbag	486	32	Mouse	59	6
Suitcase	377	30	Keyboard	1879	67
Sports ball	379	21	Microwave	667	61
Bottle	903	27	Oven	72	6
Cup	278	25	Toaster	11	4
Chair	38203	508	Sink	2694	157
Couch	1371	75	Refrigerator	60	11
Potted plant	1294	55	Book	3124	65
Bed	83	17	Clock	25	6
Bowl	121	8	Person	260	8
Dining table	1853	185	Teddy bear	47	8
Vase	13	2	-	–	–

where the model is trained on the embeddings obtained by concatenating the features from masked foregrounds and the backgrounds. In *concat* setup, the model learns to minimize the difference between the anchor f_{iA} and the positive f_{iP} images while also learning to maximize the difference between the anchor f_{iA} and the negative f_{iN} images by employing the triplet-loss based training.

Evaluation Metrics. Most re-ID algorithms use Cumulative Matching Characteristic (CMC) curve as a standard metric to measure their performance which compares the identification rate vs rank. The proportions of good matches of the probe image with the set of images in rank-1 would indicate a good or bad performance of the algorithm. A CMC curve is computed for all these individual ranks. In our evaluation procedure, however, we compare with the deepSort [22] tracking algorithm which is used here as a rank-1 re-ID method, which is why we cannot compare with a CMC curve. Also, it will not be fair to compare recall and precision values between the deepSort and our method. Thus, we compute the rank-1 accuracy by measuring the percentage of correctly identified objects.

Analysis. Evaluated using the aforementioned experimental setup, the proposed method achieves the best performance on the ScanNet dataset in regards to both the rank-1 accuracy as shown in Table 2. Figure 5 shows that the proposed method, *concat* was able to find the best match with the probe image. In the bottom row, *no-train* and *full* tried to match with an image which either had an object of the same color or the shape. However, the proposed method, *concat* could not always correctly identify the images and was performing occasionally poor as can be seen in Fig. 6. Overall, the results from Table 2 show that the *concat* method was able to improve the rank-1 accuracy by 22.19% and 17.1% against *no-train* and *full*, respectively.

Table 2. Scores on our ScanNet validation data split with Rank-1, -5, -20 and -50 accuracy values. The best performing type of setups is highlighted in bold.

Type	Rank-1 (%)	Rank-5 (%)	Rank-20 (%)	Rank-50 (%)
no-train	55.66	66.67	77.46	89.67
full	60.75	69.61	80.90	95.21
concat	**77.85**	**91.55**	**98.36**	**99.80**

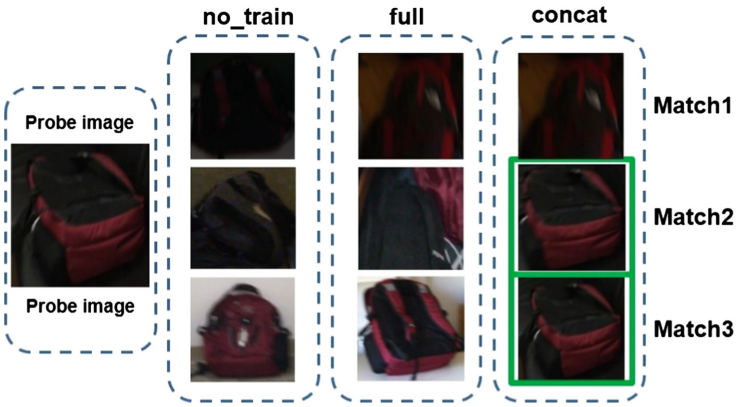

Fig. 5. The visualizations show some examples of the matches found in *no-train*, *full* and *concat* setups. The right matches with the probe image are highlighted in green color. (Color figure online)

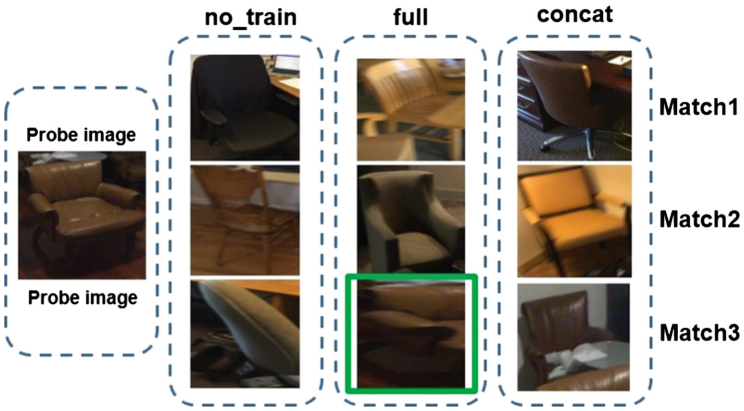

Fig. 6. Examples of the matches found in *no-train*, *full* and *concat* setups. The right matches with the probe image are highlighted in green color. (Color figure online)

Comparison with deepSort. deepSort [22] is an open-source implementation of the original SORT [1] algorithm which employs deep appearance descriptors to improve the performance in multiple object tracking. deepSort learns discriminative feature embeddings offline in order to obtain a deep association metric for a person re-identification dataset in the original work. For our experiments, we provided two random sets of image pairs obtained from the ScanNet scenes to the algorithm to identify multiple objects ensuring that an image pair is not consisting of images from two different scenes. We computed the performance by measuring the percentage of matched object instances across all the image pairs. Figure 7 shows the possible problems that standard object matching or tracking algorithms might face in re-identifying objects. The figure shows that the deepSort was able to match an object (in yellow bounding box) in multiple frames but lost an object (in red bounding box) when the camera revisits a similar view later. deepSort achieved a rank-1 accuracy of 49.60% against the rank-1 accuracy of 77.85% obtained with our method.

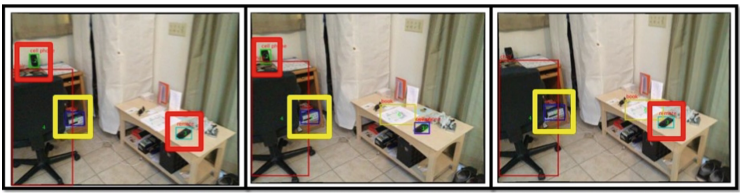

Fig. 7. An example object being matched by the deepSort algorithm inside the yellow bounding box and the lost object in the red bounding box. (Color figure online)

5 Conclusion

The contribution of this paper was to explore the intuition that the information obtained from the background surrounding the detected target objects in a rigid scene could be highly useful in discriminating two near-identical objects or two instances of the same object class. The discriminative features learned from the explicit concatenated foreground and background can be utilized to re-identify objects at the instance-level throughout the dataset. Our experiments have shown that the proposed method performs well even in the case of highly cluttered rigid environments like the indoor scenes obtained from ScanNet dataset. In future, we plan to explore if the temporal information obtained from multiple views in a video dataset can be integrated with our object instance re-identification system for a robust multiple object tracking algorithm in case of rigid and static scenes.

References

1. Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B.: Simple online and realtime tracking. In: Proceedings of IEEE International Conference on Image Processing, pp. 3464–3468, September 2016
2. Bhuiyan, A., Perina, A., Murino, V.: Exploiting multiple detections to learn robust brightness transfer functions in re-identification systems. In: Proceedings of IEEE International Conference on Image Processing, pp. 2329–2333, September 2015
3. Boureau, Y., Bach, F., LeCun, Y., Ponce, J.: Learning mid-level features for recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 2559–2566, June 2010
4. Chechik, G., Sharma, V., Shalit, U., Bengio, S.: Large scale online learning of image similarity through ranking. J. Mach. Learn. Res. **11**(Mar), 1109–1135 (2010)
5. Conte, D., Foggia, P., Percannella, G., Vento, M.: Performance evaluation of a people tracking system on pets2009 database. In: Proceedings of IEEE International Conference on Advanced Video and Signal Based Surveillance, pp. 119–126, August 2010
6. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3D reconstructions of indoor scenes. In: Proceedings of IEEE Computer Vision and Pattern Recognition, pp. 5828–5839 (2017)
7. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 886–893. IEEE (2005)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
9. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 2360–2367. IEEE (2010)
10. Frome, A., Singer, Y., Malik, J.: Image retrieval and classification using local distance functions. In: Advances in Neural Information Processing Systems, pp. 417–424 (2007)
11. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of IEEE International Conference on Computer Vision, pp. 2961–2969 (2017)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of Ieee Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
13. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
14. Li, A., Liu, L., Wang, K., Liu, S., Yan, S.: Clothing attributes assisted person reidentification. IEEE Trans. Circuits Syst. Video Technol. **25**(5), 869–878 (2015)
15. Lowe, D.G.: Object recognition from local scale-invariant features. In: Proceedings of IEEE International Conference on Computer vision, vol. 2, pp. 1150–1157. IEEE (1999)
16. Milan, A., Leal-Taixé, L., Reid, I., Roth, S., Schindler, K.: MOT16: a benchmark for multi-object tracking. arXiv preprint arXiv:1603.00831 (2016)
17. Parikh, D., Grauman, K.: Relative attributes. In: Proceedings of International Conference on Computer Vision, pp. 503–510. IEEE (2011)

18. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: a unified embedding for face recognition and clustering. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 815–823 (2015)
19. Schultz, M., Joachims, T.: Learning a distance metric from relative comparisons. In: Advances in Neural Information Processing Systems, pp. 41–48 (2004)
20. Taylor, G.W., Spiro, I., Bregler, C., Fergus, R.: Learning invariance through imitation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 2729–2736. IEEE (2011)
21. Wang, J., et al.: Learning fine-grained image similarity with deep ranking. In: Proceedings of IEEE Computer Vision and Pattern Recognitionm, pp. 1386–1393 (2014)
22. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: Proceedings of IEEE International Conference on Image Processing, pp. 3645–3649. IEEE (2017)