Università degli Studi di Cagliari

PHD DEGREE

Electronic and Computer Engineering

Cycle XXXIII

TITLE OF THE PHD THESIS

Human Centered Computer Vision Techniques for Intelligent Video Surveillance Systems

Scientific Disciplinary Sector

ING-INF/05

PhD Student: Rita Delussu

Supervisor Prof. Giorgio Fumera

Final exam. Academic Year 2019 – 2020
Thesis defence: February 2021 Session

Human Centered Computer Vision Techniques for Intelligent Video Surveillance Systems

Rita Delussu

Department of Electrical and Electronic Engineering (DIEE)

University of Cagliari

A thesis submitted for the degree of

PhilosophiæDoctor (PhD), DPhil, Electronic and Computer Engineering

2021 February

Supervisor: Prof. Giorgio Fumera

Co-supervisor: Prof. Fabio Roli

Abstract

Nowadays, intelligent video surveillance systems are being developed to support human operators in different monitoring and investigation tasks. Although relevant results have been achieved by the research community in several computer vision tasks, some real applications still exhibit several open issues. In this context, this thesis focused on two challenging computer vision tasks: person re-identification and crowd counting. Person re-identification aims to retrieve images of a person of interest, selected by the user, in different locations over time, reducing the time required to the user to analyse all the available videos. Crowd counting consists of estimating the number of people in a given image or video. Both tasks present several complex issues. In this thesis, a challenging video surveillance application scenario is considered in which it is not possible to collect and manually annotate images of a target scene (e.g., when a new camera installation is made by Law Enforcement Agency) to train a supervised model. Two *human centered* solutions for the above mentioned tasks are then proposed, in which the role of the human operators is fundamental. For person re-identification, the *human-in-the-loop approach* is proposed, which exploits the operator feedback on retrieved pedestrian images during system operation, to improve system's effectiveness. The proposed solution is based on revisiting relevance feedback algorithms for content-based image retrieval, and on developing a specific feedback protocol, to find a trade-off between the human effort and re-identification performance. For crowd counting, the use of a *synthetic* training set is proposed to develop a scene-specific model, based on a minimal amount of information of the target scene required to the user. Both solutions are empirically investigated using state-of-the-art supervised models based on Convolutional Neural Network, on benchmark data sets.

Contents

List of Figures

LIST OF FIGURES

LIST OF FIGURES

LIST OF FIGURES

List of Tables

LIST OF TABLES

Chapter 1

Introduction

Intelligent video surveillance systems have had a great diffusion in order to guarantee security and to prevent acts of terrorism. For instance, a human operator, such as a Law Enforcement Agency (LEA) officer, often simultaneously monitors several videos acquired by a video surveillance system. This requires a great attention and a quick reaction in case of anomalous (e.g., panic escape, unexpected groupings, etc.). An intelligent video surveillance system can support human operator through different tools which, for instance,

- analyse human behaviour. In the case of anomalies such as panic escape, the system raises an alert and the human operator can act accordingly
- recognise a set of people. This can be used in monitoring the entrance in specific areas of employees in a company or looking for a criminal among all people who pass through the gait.
- track a suspicious person in real-time. It can be useful to monitor a criminal and to keep in track its movements
- estimate the number of people in a given area. It can be used, e.g., to guarantee security in case of overcrowding
- finding a suspicious person in a large amount of recorded videos. This can be useful in forensics investigation, e.g., to reconstruct the movements of that person

To mitigate the required human effort in real-time video monitoring and in analysing recorded videos, the research communities of computer vision, pattern recognition and

machine learning have developed several methods and tools such as object and event detection, object and pedestrian tracking and recognition. Despite the significant results achieved so far, many open issues remain in several real-world applications such as human behaviour analysis, pedestrian tracking in a crowded scene, and real-time crowd analysis. These tasks are challenging especially in unconstrained scenarios which are typical of video surveillance applications, making it difficult to satisfy user requirements.

This thesis focuses on two computer vision tasks that are of great relevance for real-world applications, namely person re-identification and crowd counting.

Person re-identification consists of recognising a person of interest, selected by the user, in a possibly very large amount of recorded videos from non-overlapping cameras. The main goal of a person re-identification system is to reduce the time required to human operator to analyse all the available videos.

To this aim, the system automatically compares the query against all pedestrian images detected in the available videos (template gallery). The output of this comparison is a ranked list of template gallery images based on their similarity to the query. The user can then hopefully find the person of interest (if present in the template gallery) near to the top ranks of that list.

Person re-identification is a challenging task in unconstrained scenarios due to several issues such as illumination variations, different poses, different cameras, low-resolution videos and occlusions. Due to these issues, it is not possible to use classical biometric techniques such as face recognition. "Soft biometric" cues have to be used instead, among which clothing appearance is the most widely used. State-of-the-art approaches are based on machine learning, and especially on convolutional neural networks, which however require a considerable amount of labelled data. To mitigate the lack of labelled images in many real-world applications, unsupervised approaches have been proposed, but they still require images (albeit unlabelled) of the target scene which is unfeasible or too demanding for some applications or end users. To overcome this issue, the human in the loop (HITL) approach is considered in this thesis. It consists of exploiting some feedback from a human user to improve system performance, leveraging the often complementary strengths by humans and machines. This approach has already been considered in some other computer vision tasks, such as image segmentation and fine-grained image recognition, as well as by a few authors for person

re-identification. In a person re-identification system, the inherent interaction with a human operator can be conveniently exploited, for instance by asking the operator a feedback on the presence of the query identity, or of similar or very different individuals, in the top positions of the ranked list; such a feedback can be used to automatically re-rank the gallery images, aiming at "pushing" images of the query identity (if any) to the top of the list, which in turn can result in further reducing the time required to the user to find them.

The second computer vision task considered in this thesis is named crowd counting. It consists of estimating the number of people in a given image or video frame. In particular, in this thesis the task of estimating in real-time the size of a possibly dense crowd from a video is addressed. This can lead to useful tools for supporting human operators (e.g., LEA officer) in monitoring a crowded area, e.g., during mass gathering events like demonstrations and concerts. Based on the estimated crowd size over time, further automatic functionalities related to the detection of anomalous crowd behaviours can be enabled, such as overcrowding and panic escape (which may result in a sudden decrease of crowd density). The system can then send an alert to the operator who can act accordingly. The main advantage is the reduction of user's monitoring effort, especially when several videos possibly coming from different areas have to be monitored simultaneously. Also this task presents several challenging issues due to perspective distortion, illumination changes, scale variations due to perspective, different and (often) complex background, and occlusions (especially in dense crowds). As in the case of person re-identification, state-of-the-art approaches are based on machine learning and convolutional neural networks, and require a considerable amount of manually labelled crowd images of the target scene, due to their sensitivity to background, perspective etc. However, in some challenging real-world application scenarios such as new installation of cameras by the LEAs that should be operational in a short time, collecting and manually labelling representative images of the target scene can be infeasible. This raises the issues of how to develop a *scene-specific* crowd counting system when a representative set of labelled images of the target scene is not available. To this aim, an approach based on building a *synthetic* training set is proposed in this thesis. The proposed approach only requires to the user basic information on the target scene: a background image, the region of interest (where people can appear) and the perspective map. This information is then used to automatically generate a synthetic

data set of crowd images of a suitable size, which are automatically labelled with the *exact* crowd size.

For both tasks it is essential that the computer vision systems effectively interacts with human operators with the aim of reducing their effort in routine and boring tasks; at the same time human capabilities related to tasks that are still challenging for machines should be leveraged. Therefore effective computer vision solutions should be *human-centred*.

1.1 Contributions

This work is the result of research activities that have benefited from the collaboration in the European Union Project LETSCROWD[1] under the HORIZON2020 program. The project goals included the development of prototype computer vision tools to support LEA operators in monitoring and investigation activities related to mass gathering events.

With regard to person re-identification, the HITL approach is used to address the cross-scene setting, such as the one considered in this thesis. The person re-identification problem can be considered as an image retrieval one and, as a consequence, content-based image retrieve with relevance feedback (CBIR-RF) techniques can be used to implement the HITL. This is disregarded by HITL approaches proposed at the state of the art, which are based on methods quite complex. The contribution of this thesis consists of investigating well-known CBIR-RF techniques for person re-identification, and in particular, of developing a feedback protocol which is different from existing HITL methods. Experimental comparison of UDA methods which are recently used at the state of the art in cross-scene settings (such as the one considered in this thesis), but too demanding in the examined application scenario. It was not possible to make a direct comparison with existing HITL methods since the code was made not available by authors, and it was not possible to re-implement them by using the information provided by authors.

Similarly to the person re-identification scenario mentioned above, also with regard to the crowd counting task, this thesis focuses on a challenging cross-scene application

[1] "Law Enforcement agencies human factor methods and Toolkit for the Security and protection of CROWDs in mass gatherings" https://letscrowd.eu/ (May 2017 - October 2019)

scenario in which collecting and annotating images of the target scene is difficult and too demanding to the end users. The first contribution of this thesis is an extensive experimental evaluation of the performance gap between cross- and same-scene performance of several state-of-the-art crowd counting methods, which is still missing in the literature. The main contribution of this thesis is the proposed solution to mitigate this gap, based on the construction of a synthetic training set of the target scene, inspired by the use of synthetic images in other computer vision tasks with similar motivations. The proposed solution is evaluated on methods mentioned above. The results of this thesis have been partly reported in the following publications:

- Delussu, Rita; Putzu, Lorenzo; Fumera, Giorgio. "Investigating Synthetic Data Sets for Crowd Counting in Cross-scene Scenarios". Proc. 15th International Joint Conference on Computer Vision Theory and Applications (VISAPP), Vol. 4, pp. 365-372, 2020.
- Delussu, Rita; Putzu, Lorenzo; Fumera, Giorgio. "An Empirical Evaluation of Cross-scene Crowd Counting Performance". Proc. 15th International Joint Conference on Computer Vision Theory and Applications (VISAPP), Vol. 4, pp. 373-380, 2020.
- Delussu, Rita; Putzu, Lorenzo; Fumera, Giorgio; Roli, Fabio. "Online Domain Adaptation for Person Re-Identification with a Human in the Loop". Proc. 25th International Conference on Pattern Recognition (ICPR), in press
- Delussu, Rita; Putzu, Lorenzo; Fumera, Giorgio. "Scene-specific Crowd Counting Using Synthetic Training Images". Pattern Recognition - (under review)

1.2 Organisation

This thesis is organised as follows:

- chapter 2 presents a complete overview of the approaches at the state of the art for person re-identification (section 2.1) and crowd counting (section 2.2)
- chapter 3 describes the proposed approach for person re-identification and its empirical evaluation, including experimental settings, data sets, performance metrics, and a discussion of the results

- chapter 4 explains in details the proposed approach for crowd counting and its empirical evaluation
- conclusions and future work are reported in chapter 5

Chapter 2

State of the art

As described in the previous chapter, this thesis focused on computer vision techniques that can support human operators, such as LEA officers in monitoring and investigation tasks.

The tasks of interest are known as person re-identification and crowd counting. Person re-identification consists of recognising images, acquired from different (non-overlapping) cameras, of a person of interest, which can be useful, e.g., to reconstruct the movements of a suspect individual [152, 163]. Crowd counting consists of counting people in a given image or video frame [85]. Both tasks present several and challenging issues caused by unconstrained acquisition scenarios that have been mentioned in the previous section and will be further described below. In particular for person re-identification, due to these issues, it is infeasible to use classic biometric techniques such as face recognition and, therefore, it is necessary to exploit different cues. A widely used cue is clothing appearance, which is however valid only for a limited amount of time. In other words, it is usually not possible to look for an individual of interest using a query acquired on a certain date in videos recorded in different days since the person might change its clothes. The input of a person re-identification system is an image of the person of interest chosen by the user, named query, and the output is a list of images of pedestrians automatically detected on the available videos, and sorted based on the similarity to the query. Early person re-identification techniques consisted of manually defined features (usually based on colour and texture) and similarity measure [163], whereas more recent ones, mostly based on Convolutional Neural Networks (CNNs), can fuse these aspects in a unique framework by automatically extracting discriminant

features and learning a similarity measure [152]. The main difference between early and recent approaches is that CNNs require a training phase, which in turn requires a large amount of manually annotated images.

Until recently, almost all proposed techniques were evaluated in an "ideal" scenario, i.e. when the training and the testing data belong to the same data set, i.e., when the corresponding images are acquired from the same set of cameras. However this scenario does not reflect many real application scenarios where the training and testing data are different, for instance because a pre-trained re-identification model is deployed by the end users to new target scenes acquired by different cameras. It is now clear that under such a cross-scene scenario, through cross-data set experiments, the performance of the state-of-the-art re-identification models can severely decrease [39, 119, 154]. To overcome this issue, approaches such as Domain Adaptation (DA) and later Unsupervised DA (UDA) have been proposed [16, 95]. DA is a technique in which a model trained on a source domain is applied on a different but related target domain. It is defined supervised DA or simply DA if the labels of the target domain are available; otherwise, it is defined UDA. Substantially, to use these techniques is necessary to re-train or fine-tune the source model by using images of the target domain. In a person re-identification system, the use of these approaches consists of training a model by using pedestrian images from one data set (source domain) and re-training or fine-tuning it by using images from another data set (target domain). Nevertheless, DA and UDA approaches require images of the target domain that in challenging real applications, may be not available. To overcome this issue, another approach, known as Domain Generalisation (DG), has been proposed. DG aims at improving generalisation capability by training the source model on several different source domains [51].

The input of a crowd counting system is an image and the output is an estimation of the number of people in the given image. Early approaches consist of extracting low-level features such as edges, textures, etc., and of learning a regression model to estimate the pedestrian count [85]. More recent approaches, as in the case of person re-identification, are based on CNNs [114]. Also in this task very high performances have been reported in the literature under the "ideal" same-scene scenario but in cross-scene settings the performances significantly degrade.

As in the case of person re-identification, DA and UDA approaches have been proposed to address the cross-scene settings [135, 154], but they still require representative

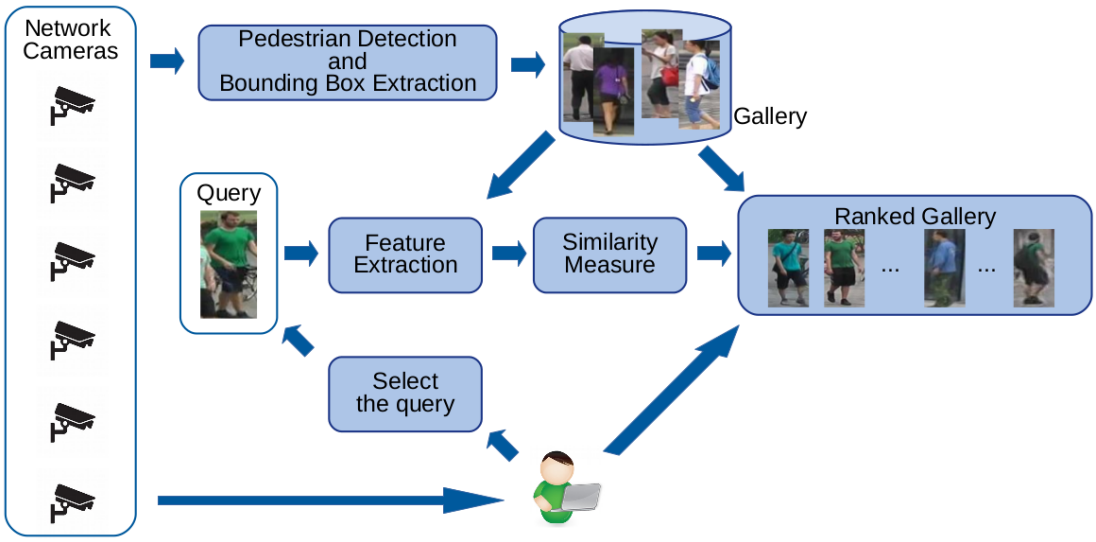


Figure 2.1: Scheme of a person re-identification system. All the videos acquired by network cameras are shown to the user. At the same time, these videos are processed to automatically detect and extract bounding boxes of pedestrians (pedestrian detector and bounding box module), then these images are stored in a database (gallery). The user can select a person of interest (query). The features of query and gallery images are extracted (feature extraction module) and, then, compared by using a similarity measure. Finally, the system shows the user a list of images (ranked gallery) ordered by the similarity to the query.

images of the target scene.

In the next sections, the state of the art of person re-identification and crowd counting tasks are described in section 2.1 and 2.2, respectively.

2.1 Person Re-Identification

Person re-identification consists of recognising images of a person of interest across several and non-overlapping cameras. It is a challenging task due to several issues such as illumination variations, different poses, occlusions, different viewpoints, different camera resolutions, etc. [152, 163] (fig. 2.2). In a video surveillance system, a person re-identification system can be used for monitoring and investigation tasks, e.g., to reconstruct movements of a suspect individual. The advantage for end user consists of limiting the time required to manually inspect all videos. In fig. 2.1 a scheme of a person re-identification system is shown. Videos recorded by a camera are shown to

the user. At the same time these videos are processed by the "pedestrian detector and bounding box extraction" module that detects and extract images of single pedestrians. All these images are stored in a database (gallery). The user can select a person of interest from a video and this image (query) is processed in order to retrieve all images of the same identity acquired from different cameras. The query is compared against all images contained in the gallery using the extracted features and using a given similarity measure. Finally, all images of the gallery are ordered based on their similarity to the query (ranked gallery).

Implementing the different modules of this system is difficult since each of them presents several issues. Pedestrian detection and accurate bounding box extraction is challenging since often there are occlusions either static (objects) and dynamic (overlapping among people). Pedestrian detectors based on tracking can mitigate this problem, but in case of temporary occlusions the identity labels can be switched, or a single track can be subdivided into two different ones. In the latter case, more than one bounding box per track can be extracted. In some cases, more than one person can appear in a single stored bounding box. In other cases, the bounding box can show a part of person only. For these reasons, the size of the gallery can be very large, and the bounding boxes may be inaccurate. All the problems mentioned above can significantly affect the performances of a person re-identification system. In the remainder of this thesis, figure 2.1 was simplified to focus on some parts of the scheme.

A large number of person re-identification methods have been proposed so far. These approaches can be categorised in hand-crafted, metric learning and, the more recent CNN-based ones. The term hand-crafted indicates approaches that do not require a learning phase; therefore features and the similarity measure are defined a priori [163]. In principle, hand-crafted methods should be effective in different kinds of scene. In practice, they achieve a lower accuracy than supervised approaches. Metric learning approaches, instead, require a learning phase since they focus on the definition of a similarity measure that keep images of similar individuals close in feature space, and images of dissimilar individuals farther. Therefore, these approaches are adapted to a specific scene, and their performances usually decreases in cross-scene settings.

Schemes of two systems that use hand-crafted and metric learning approaches are shown in figure 2.3. Each system consists of a module that extracts features of the image of the query and gallery images. The features of the query are compared against

Figure 2.2: Examples of person re-identification issues. Couple of images from the same person (acquired from two different cameras) show different poses (first column), different resolutions (second column), illumination variations (third column), occlusions (last column). Images taken from Market-1501 data set (more details can be found in section 3.2.3).

the images in the gallery by using a module (similarity measure or specific measure) that computes the similarity between image features. The output is, as described in the previous section, a list of gallery images (ranked gallery) based on the similarity to the query. In the ranked gallery, the top-rank images usually belong to individuals that exhibit a similar clothing appearance to the query, whereas the other images (different from the query) tend to be in the bottom of the list. Both methods present the same structure (fig. 2.3), but the main difference is that metric learning approaches learn the similarity measure using manually labelled data.

More recent approaches, based on CNN automatically learn discriminant features, and can also combine feature extraction and similarity measure learning in a single "framework". It is necessary to introduce some key points of the architecture of CNNs.

Figure 2.4 shows an example of CNN architecture. Manual feature extraction of early methods is replaced by a series of Convolutional (Conv.), Rectified Linear Unit (ReLU) and Pooling layers. A series of convolutional layers extract low- and high-level features such as edges, colour, gradient orientations, semantic features etc. ReLU layers replace all negative values by zeros. Pooling layers reduce the dimension of features retaining the most essential information. The second part of the architecture handles the classification that replaces the early similarity measure module. The flatten layer implements a reshaping of the features and transforms them into a vector. The Fully

Connected (FC) layer is a traditional multi-layer perceptron. In some architectures, this layer handles the classification of the extracted features into several classes (identities), whereas in other architectures the model is learnt to compute a similarity measure between two images (such as in a Siamese Network). Fig. 2.4 represents the first case, i.e., pedestrian identity recognition, in which classes correspond to identities (IDs). Therefore, the output of this layer is a score. The softmax layer takes this vector of scores and normalises it in order to obtain values between 0 and 1. Therefore for each image of a person, the score indicates the similarity to the i-th identity.

The learning phase of CNN is usually time consuming with respect to earlier approaches. During training, a large set of manually labelled examples (training set) is used to learn the model. After the training phase, the resulting model is evaluated on a testing set composed of images (queries and template gallery) of people different from those used for training. During this phase, a single query is compared against the images contained in the gallery. The output consists of a score vector that indicates (for each identity) the similarity to the query. By ordering (based on the similarity) this vector, it is possible to obtain the ranked gallery. In the remainder of this thesis, the architecture of a CNN is represented by a three-dimensional element (i.e., a cube) as shown in figure 2.5.

A CNN can be used only as a feature extractor or also for computing image similarity. A drawback of CNN-based approaches is that they require a significant amount of images for the training phase. Collecting and manually annotating a large set of images is difficult and laborious. Moreover, in some applications the target scenes where the system will be deployed are unknown at design (training) phase, which can significantly affect the performance of supervised approaches such as CNN-based and metric learning ones.

Supervised approaches reached a very high recognition accuracy on benchmark data sets under same-scene settings [152]. However, in real-world application scenarios, it may be difficult or infeasible to collect and label pedestrian images from the *target* camera views (i.e., the ones that will be used after system deployment) for training a person re-identification model. More recent approaches (mainly CNN-based ones) address more realistic cross-scene settings where the data used to train the model is different (e.g., comes from different cameras) from the target data. Some approaches are based on DA which however requires labelled images of the target scene to re-train

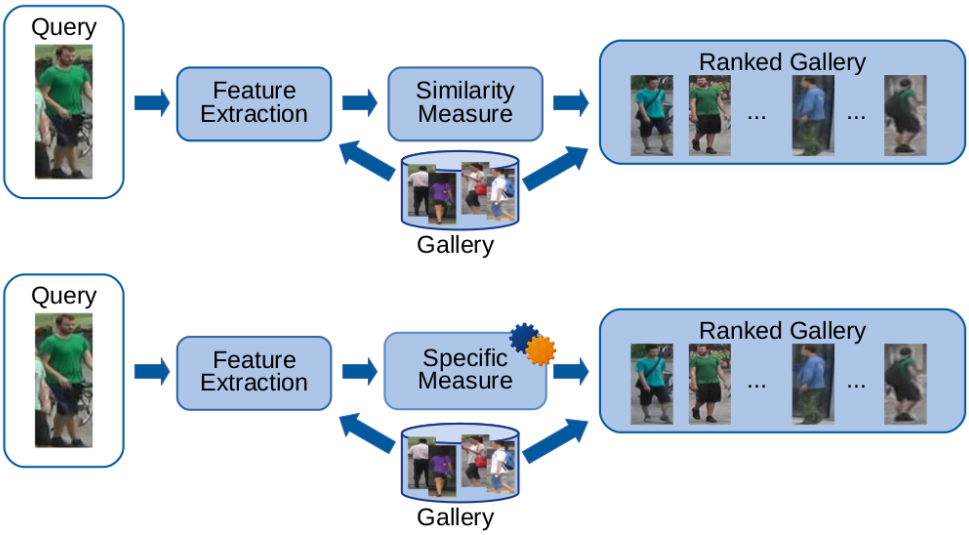


Figure 2.3: Two schemes of systems that show the difference between an hand-crafted approach (top) and a metric learning approach (bottom). The input (query) and the output (ranked gallery) of both systems are equal. The difference in these schemes is related to one of the two internal module, i.e. the module that computes the similarity measure. In a hand-crafted system, the similarity measure is defined a priori, whereas in the bottom system the measure is defined after a learning phase (gearwheels). Therefore, the second system aims to define a specific measure to compute the similarity between the query image and the images contained in the gallery.

the model. Other authors proposed UDA approaches, that require only unlabelled images from the target domain, and are typically based on learning shared feature space between the source and target domain. In other words, these approaches (DA and UDA) use an auxiliary (for instance a benchmark data set) labelled or unlabelled images of the target domain to re-train a model. One limitation of these approaches is that they *still* require a significant amount of images (even if unlabelled) of the target domain. However, in a scenario where a new camera installation by LEAs has to be operational in a short time, it may be unfeasible or too demanding to acquire (and manually label, if necessary) a sufficient large set of representative of the target scene In these cases, possible solution is to adopt a human-in-the-loop approach as explained in section 2.1.3.

Some recent approaches are based on the use of multiple source domains during the training phase to improve the generalisation capability of a person re-identification

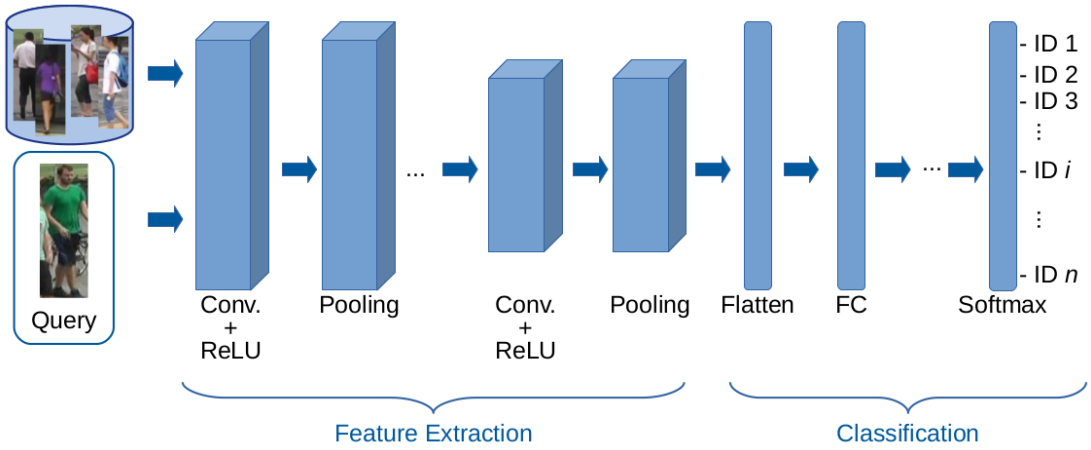


Figure 2.4: An example of CNN architecture used for person re-identification. The CNN fuses feature extraction and classification in a unique framework. Convolutional (Conv.), Rectified Linear Unit (ReLU) and Pooling layers handle feature extraction, whereas Flatten, Fully Connected and Softmax layers handle classification. During the training phase, a large set of images (top left corner) are used to learn the CNN, whereas a single query image (bottom left corner) is used to apply the trained model.

model [6, 51]. These approaches are known as Domain Generalisation (DG) [51]. In contrast, other authors have been proposed to use a fully unsupervised approach in which the model is trained or learnt by using data without identity labels [28, 76].

To sum up, it is possible to define three main categories of person re-identification techniques, regardless to the kind of learning (supervised or not), that can be defined as hand-crafted, metric learning and CNN-based approaches. In contrast to the other approaches, hand-crafted ones do not require any learning phase. The metric learning approaches aim to define a specific similarity measure. The more recent approaches based on CNNs can be categorised in turn into several dimensions (see sect. 2.1.2). In the next sections an overview of the traditional and CNN-based approaches is given.

2.1.1 Traditional approaches

As mentioned in the previous section, early approaches, named hand-crafted and metric learning ones, have a similar structure in the scheme of Fig. 2.3 but the second one requires a learning phase to define the similarity measure. In contrast, the first one does not require any learning phase since features and similarity measure are fixed.

The most common features used in literature are related to colours, using either RGB, or other colour spaces, such as YCbCR, YUV, Color LAB etc. [42, 89, 161]. In addition other features have been proposed related to texture, edges, and shape, either global or local [1, 42, 90, 98, 110, 136]. Such features are extracted either from the whole image, or from uniform horizontal stripes of the same size or by using a more complex partition of the image (e.g., into head, torso and legs) [5, 136]. Successively, several approaches have been proposed to improve the robustness of pedestrian descriptors. For instance, SIFT features have been proposed to achieve scale invariance [84]. In particular, SIFT extracts local information and is more robust to illumination, translation and rotation variations. In [151] Salient Colour Names Based Colour descriptor (SCNCD) is proposed to improve the robustness of RGB colour features. To address viewpoint changes a Local Maximal Occurrence (LOMO) and Symmetry-Driven Accumulation of Local (SDALF) feature have been proposed [29, 72]. With regard to the similarity measure ("similarity measure" module in fig. 2.3), traditional distance metrics have been used such as Euclidean, city block, cosine etc. To fit the distance measure to data distribution in feature space, some authors proposed to use distance metric learning techniques [163]. The most common formulation is based on the Mahalanobis distance defined as

$$D(x, x') = \sqrt{(x - x')^T M (x - x')} \tag{2.1}$$

where M is a positive semi-define matrix, whereas x and x' are feature vectors. Inspired by this formulation, some authors proposed different metric learning approaches [163]. In [96] local Fisher discriminant analysis (LFDA) was proposed. This approach combines the advantages of the Fisher discriminant analysis (FDA) [33] and of local preserving projection (LPP) [46]. By using LFDA it is possible to reduce data dimensionality preserving at the same time the local structure of data. In [140] large margin nearest neighbour learning (LMNN) strategy has been proposed to keeps close the neighbours from the same class (i.e., feature vectors corresponding to images of the same identity) and penalise the presence of impostors.

Another information-theoretic approach based on Mahalanobis distance was proposed by [17], named Information Theoretic Metric Learning (ITML); it allows finding a distance function that satisfies some constraints related to similar and dissimilar image pairs and to the relative ranking of different image pairs. Also in [61] an approach based on similarity and dissimilarity constraints, named KISSME (keep it simple and

straightforward method) was proposed. This approach is more scalable than the above-mentioned ones and replaces their optimization procedure with equivalent constraints based on likelihood ratio test.

The KISSME and ITML methods require a dimensionality reduction step. Instead in [89] a method capable of working on high-dimensional data was proposed. This method, named Pairwise Constrained Component Analysis (PCCA), is applicable on a sparse set of pairwise similarity constraints. It is an effective approach when the similarity information is available for a few pairs of points. PCCA learns a linear mapping function avoiding a regularization term and keeping a lower computational complexity than ITML. An extended version of KISSME, known as Cross-view Quadratic Discriminant Analysis (XQDA), was proposed [72]. This method was used to learn a cross-view discriminative subspace and a cross-view similarity measure at the same time.

To overcome the problem related to the high non-linearity of person appearance in feature space across different cameras, some authors proposed to use kernel-based distance metric learning methods [3, 91, 145]. The key idea of these methods is to perform a non-linear mapping from the input space to a higher-dimensional feature space. This allows a linear methods (e.g., KISSME, LFDA, XQDA, etc.) to be extended to the corresponding non-linear versions (e.g., such as k-KISSME, k-LFDA, k-XQDA).

2.1.2 CNN-based approaches

CNN-based approaches allow fusing the two main components of person re-identification model (pedestrian descriptor and similarity measure) into a single framework [163].

CNN-based methods can be categorised into several dimensions. A possible categorisation is the network architecture: i) Siamese [101, 112, 130, 133, 138, 141, 164] or triplet model [13], ii) a modification of a existing architecture (e.g., ResNet, GoogleNet, etc.) [10, 49, 59, 78, 120, 121, 123, 124, 132, 137, 150, 169] or iii) a specifically-devised architecture [66, 69, 128, 168].

The Siamese architecture, in formal terms, consists of two identical CNNs that share the same parameters [68] (fig. 2.5). For instance, each branch can have the architecture shown in figure 2.4. A Siamese CNN takes as input two images and output is a similarity score between them. The score is usually between 0 (no similarity) and 1 (full-similarity).

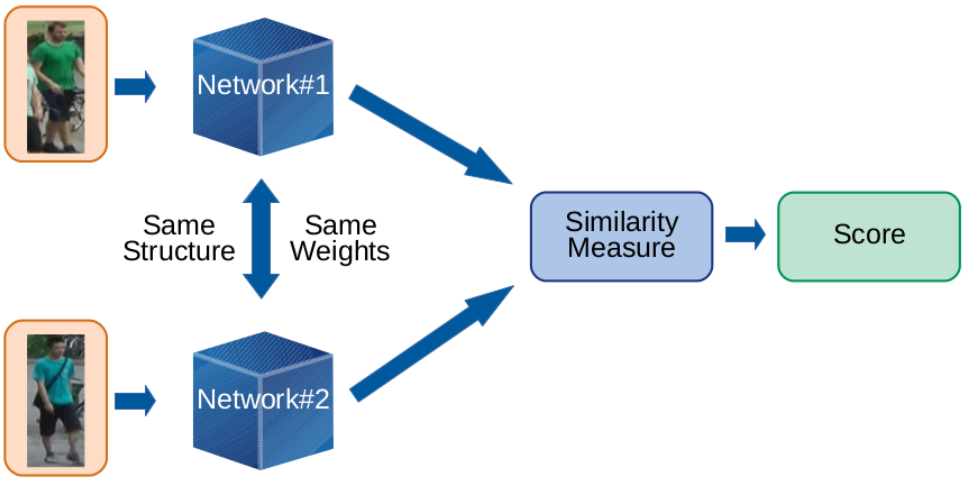


Figure 2.5: Siamese Network scheme. The two images are the input (orange boxes) of the two networks with the same structure and weights. The output features are compared by the similarity measure module that computes the similarity score (output in green).

A modified Siamese network has been proposed in [101] where the problem of matching pedestrian images at different scales was addressed. A multi-scale deep learning (MuDeep) model has been proposed to learn more discriminant features at different resolutions taking into account the cross-camera problem. The MuDeep structure presents two different outputs: one defines whether the two input images belong to the same identity or not, while the other one can predict the identity of the person. In [112] a module, named "Kronecker product matching", based on ResNet has been added to the CNN architecture to match two sets of multi-scale feature maps. Similarly, in [164] an attention module based on ResNet is added to the Siamese architecture, to focus on the foreground subject.

The triplet model consists of three identical networks with shared parameters. During training, it takes as input three images, where two of these images belong to the same person and the other one is from a different person. This framework has been used by [13] including a triplet loss based on intra- and inter-class constraints. That loss function aims at learning a feature space where the distance between similar images is lower than that between different images, with a predefined margin.

Most of the existing frameworks are based on well-known CNN architectures as ResNet [45], GoogleNet [125], CaffeNet [62], etc. A ResNet backbone was used in

[10, 49, 78, 124, 132, 137, 150, 169]; a GoogleNet backbone in [78, 120, 121]; and a CaffeNet backbone in [123]. All these works introduced some modifications to the corresponding architectures in order to extract more discriminant features or to learn an improved similarity metric, or both (fig. 2.6).

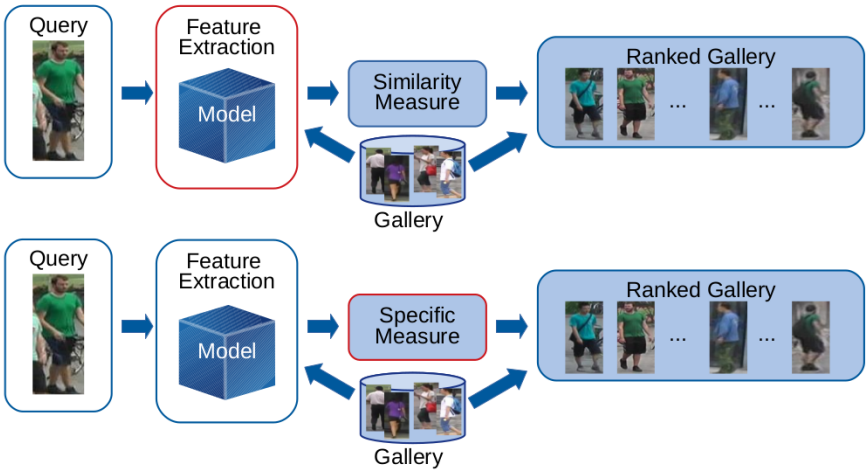


Figure 2.6: Different improvements (highlighted in red) to CNN-based approaches. Top: the extraction of more discriminant features is achieved by modifying one or more layers of the original model. Bottom: the improvement is based on the modification of an existing similarity metric, or on definition of a specific one.

Specifically-designed architectures aim at extracting robust features and at finding good similarity metrics.

As a consequence, another possible categorisation can be defined by focusing on which works use the CNN as a feature extractor and employ a standard (traditional) metric such as Euclidean distance, cosine distance etc. as shown in fig. 2.6 (top). This characteristic can be found in [133] where the CNN is used to extract global image features. In particular, single- and cross-image representations are considered, and a generalisation of the Euclidean distance is used as a metric. In [123] a singular vector decomposition (SVD) is used to extract global discriminant features, and a traditional Euclidean distance is utilised. The standard cosine distance has been used in [49] SIFT and other features that are designed to be robust to view changes.

An additional categorisation refers to frameworks that focus on improving either feature extraction or metric learning. Several strategies for improving feature extraction have been proposed, such as multi-scale analysis of pedestrian images [13, 101, 112, 150],

local and/or global features extraction [59, 69, 120, 121, 124, 128, 160], the use of attention modules [10, 118, 132, 164, 169] and other approaches [78, 130]. In particular in [124, 160] a decomposition of pedestrian body into several parts has been proposed. Also in [130] the pedestrian images are subdivided into several body parts to extract discriminant features, named LOMO (Local Maximal Occurrence) and ColorName. In [78] SIFT (Scale-invariant feature transform) has been used and combined with a view confusion learning strategy which aims to define features more robust to view changes.

Most of the works mentioned above use the softmax loss [59, 69, 101, 112, 120, 124, 128, 141, 164] or the triplet loss [10, 13, 78, 118, 121, 132, 137], or a combination of them [168, 169]. A modification of the softmax loss named contrastive loss was employed in [130].

Most of the above mentioned CNN approaches have been evaluated under an "ideal" setting where the training and the testing sets belong to the same data set, which correspond to a same-scene application scenario in which manually annotated training data are available from the same camera views that will be used during system deployment. This correspond to a supervised setting. More recently, several authors focused on a more realistic cross-scene scenario in which training and testing sets come from different data sets. This setting typically leads to a performance decrease [39, 119]. To overcome this issue, some authors have been proposed approaches based on DA that aim to learn different but related tasks. In other words, a model trained on a source domain is adapted to a different (but related) target domain. Typically, DA approaches require labelled samples of the target domain to train or fine-tune the underlying model [39]. However in many practical application scenarios, including different computer vision tasks such as person re-identification, it may be difficult to collect labelled samples of the target domain at the design phase. To address this issues, UDA methods have been proposed to exploit unlabelled target samples [16, 34, 52, 95, 119, 166]. A common approach in UDA methods is to learn a shared feature space between source and target domains by using a known architecture as a backbone [34, 37, 166] or a specifically design structure [40, 52, 70] to reduce the gap between source and target domain distributions. Another approach that tries to reduce the mentioned gap is based on adversarial learning, based on generative adversarial networks (GANs) [165]. In [80, 102] images with different poses are generated to improve robustness to pose variations.

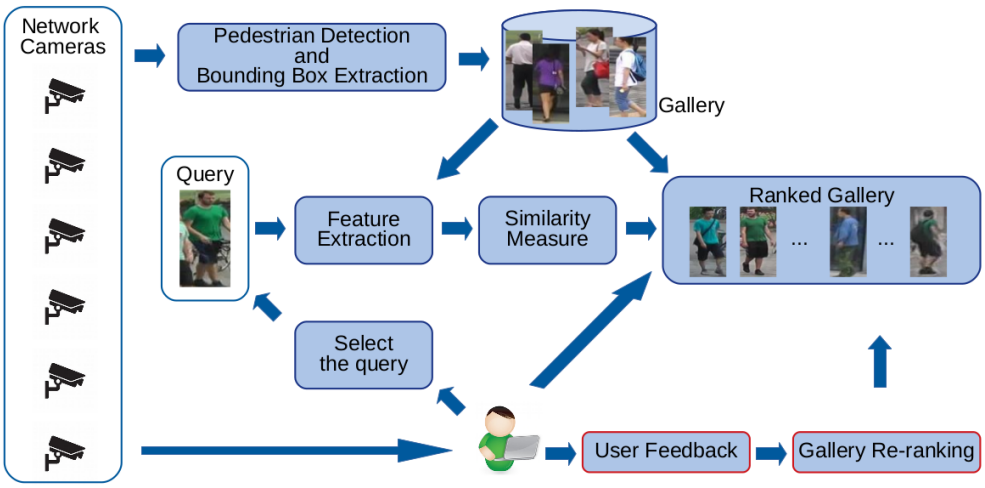


Figure 2.7: A person re-identification system with a human in the loop. The modules handle user's feedback are shown in red.

Other works focus on cross-camera variation to mitigate the camera discrepancy problem and different lighting conditions across cameras [4, 100, 126, 167].

Another recent approach, named Domain Generalisation (DG), uses training data from multiple source domains to improve the generalisation capability of the resulting model [51]. Since no training data from the target domain is used, its performance is worse with respect to DA and UDA methods [56].

2.1.3 The Human-in-the-Loop approach

The human-in-the-loop approach has been proposed in the pattern recognition and machine learning fields to exploit the complementary strengths by humans and machines in several application domains, including some computer vision tasks that are still very challenging for machine. This approach is based on some form of interaction between a human and a system, usually through a feedback which is used to improve performances of the system itself.

This kind of approach has been employed in different fields such as flight simulators [54], robotics [64], automation engineering [31], remote sensing [22], etc. In machine learning and image processing, the well known active learning (AL), online metric learning and content-based image retrieve with relevance feedback (CBIR-RF) can be considered as early HITL approaches. CBIR consists of retrieving images based

on visual content, instead of textual labels or tags, by using visual features such as colour, texture and shape [50, 75, 170]. The key idea of RF (which was previously used also in text retrieval systems), is to ask the user a feedback about relevant and irrelevant retrieved images This can allow to reduce the semantic gap in CBIR, between low-level image features and semantic image content [26, 57]. AL is used instead to limit the amount of training data that has to be manually labelled [22, 27]. It consists of automatically selecting the most informative samples among an unlabelled pool, that a user is asked to manually label, starting from a small amount of labelled data.

Recently, more refined HITL approaches have been proposed for some computer vision tasks, mainly fine-grained categorisation. The idea was to combine the strengths of human and machine in order to identify the correct class of an input image as soon as possible [7, 8, 131]. In these approaches, a human is asked to click on a specific object part, and/or to answer a limited number of questions based on visual attributes [8], or to multiple choices questions [7]. A different approach was proposed in [23] to allow humans to assist the machine in the selection of relevant features for fine-grained classification.

Despite person re-identification system inherently involve an interaction with a user, the HITL approach has been considered by few authors so far [2, 47, 63, 77, 134]. A general scheme of HITL person re-identification system is shown in figure 2.7. The user can give a feedback about the output of the system (top ranked gallery), which exploits it to update the ranked list. User feedback can be used to modify either the ranking module or the similarity measure, as shown in figure 2.8. In [47] each query is compared with the template gallery by using a given distance metric on a predefined feature space. The ranked list is shown to the user, who checks whether the query identity is present in the top-ranked images (fig. 2.9). If the query is not present, a discriminative model is learnt by using a negative example from the list. Then, an updated ranked list is shown to the user. More precisely, this system consists of two stages. In the first one, a set of features as intensity, colour, texture and other visual features are extracted, and a covariance-based distance is used to define a distance measure between features. The second stage is used only if the query is not present in the top-ranked images. In this case, local colour information and Haar features (edges and lines) are extracted to create a more discriminant descriptor. However this approach requires resources since a classifier has to be learnt in the second stage. In [2], the initial ranked list is obtained

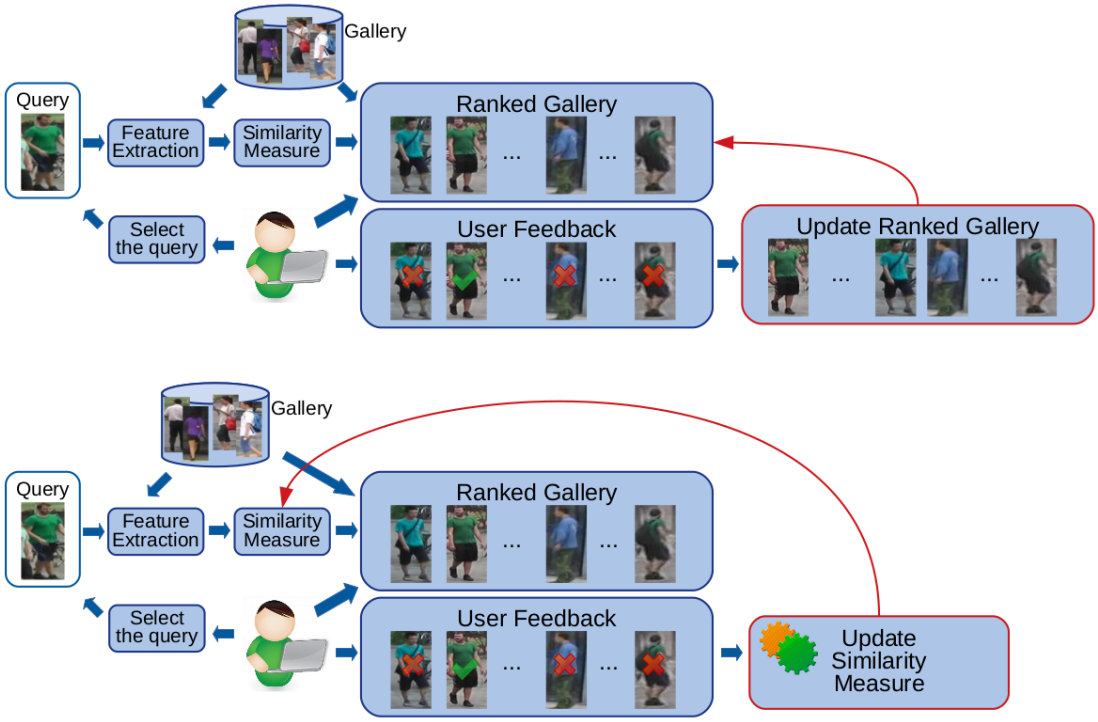


Figure 2.8: General scheme of HITL person re-identification system. Top: the feedback is used to update the ranked gallery. The images of the person of interest are selected as relevant (highlighted with the tick symbol) whereas the other ones are marked (either manually or automatically) as non-relevant ("x" mark). Relevant images are pulled to the top of the ranked list and non-relevant ones are pushed to the bottom. Bottom: the feedback is used to update the similarity measure.

by using Euclidean distance on descriptors consisting of colour space features and filter banks (Gabor and Schmid). During the first feedback iteration, the user labels images of the top-k ($k = 24$) as relevant and non-relevant. These images are used jointly with their nearest neighbours to learn a distance metric (Mahalanobis metric), which is then used to update the ranked list as shown in fig. 2.8 (bottom). Also this approach presents a disadvantage since it asks the user to label all images in the top ranked gallery as relevant and non-relevant. A different kind of feedback has been used in [77] that consists of selecting a "strong" negative and optionally a few "weak" negatives (fig. 2.10). The first feedback consists of an image that is dissimilar to the query one, whereas the second one concerns images of different identities that are similar to the query. The feedback is propagated through the graph and a ranking function is learnt to

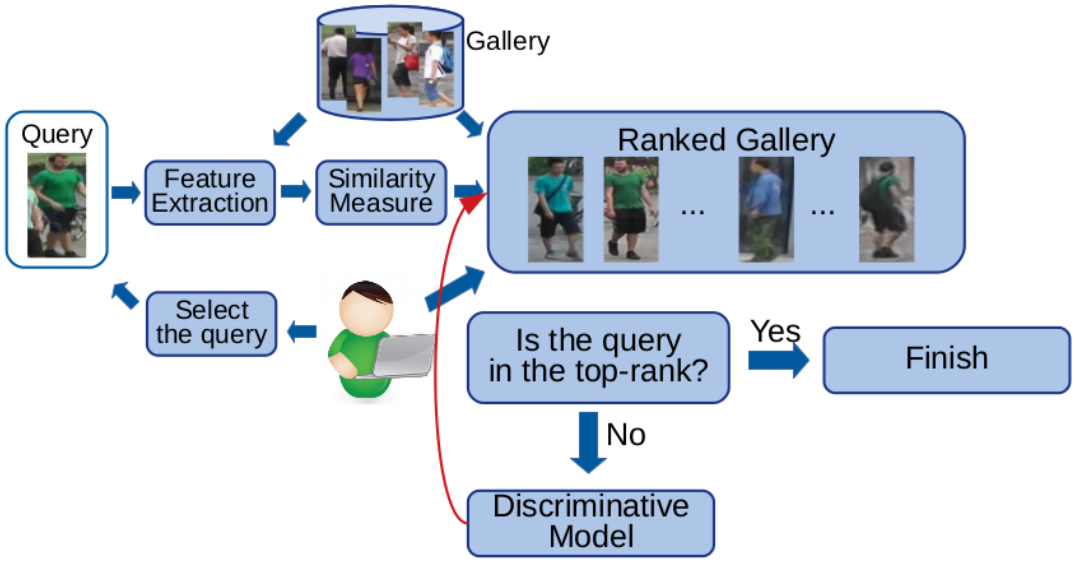


Figure 2.9: Scheme of the feedback approach proposed in [47]. The user checks whether the query identity is present among the top-ranked gallery images. If not, a discriminative model is learnt to extract more discriminant features. Finally a refined ranked gallery is shown to the user.

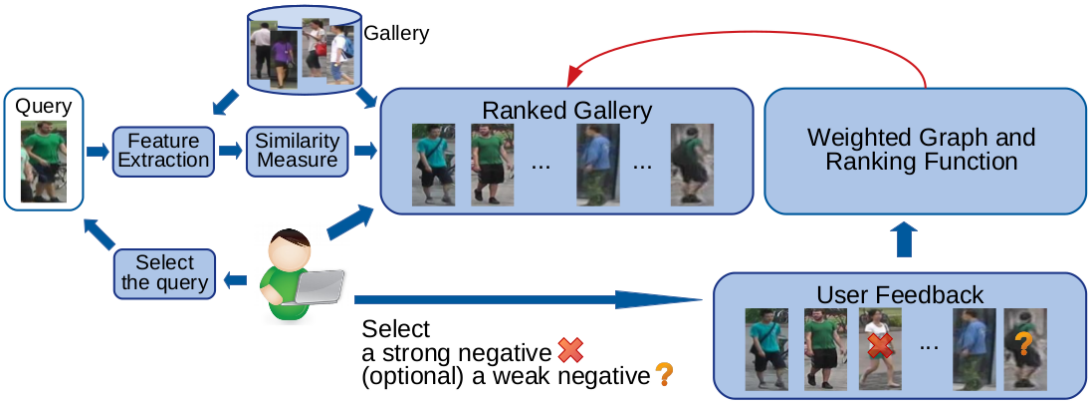


Figure 2.10: Scheme of feedback proposed in [77]. The user should select a dissimilar image with respect to the query (strong negative) and (optionally) a weak negative, i.e. a different identity looking similar to the query individual. User feedback is propagated in a weighted graph and a ranking function is learnt in order to update the ranked list.

improve the ranked list. However this approach can require significant resources when the gallery size increases since it built a graph. In other words, it is not scalable. An incremental approach has been proposed instead in [134] where the initial ranked list,

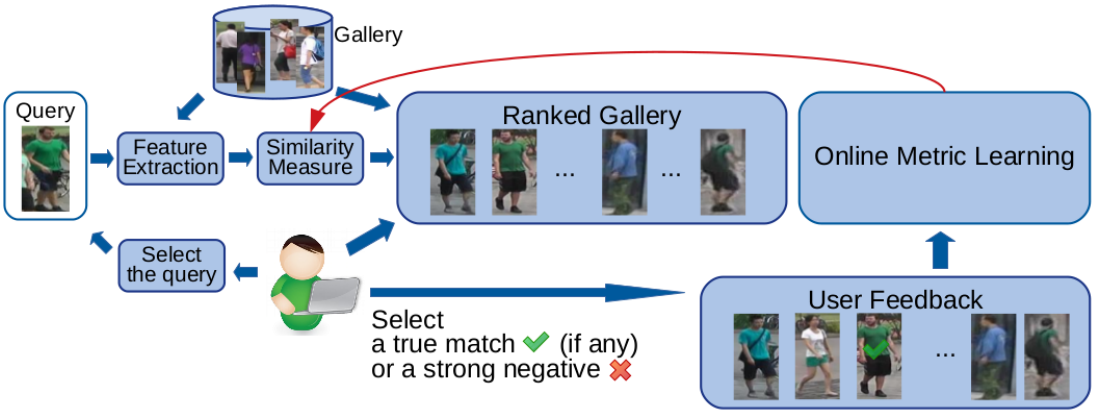


Figure 2.11: Scheme of the user feedback approach proposed in [134]. The feedback consists of selecting a true match (if any), or a strong negative. These information is used to update the ranked list by optimising incrementally the distance metric (Mahalanobis distance) in an online modality (online metric learning).

obtained by extracting predefined features and using Euclidean distance, is shown to the user who indicates a true match (if any) or a strong negative (fig. 2.11). Differently from previous methods, this one incrementally updates the distance metric, defined as Mahalanobis distance, after each user feedback, using an online metric learning algorithm. In principle, this allows the initial re-identification model to adapt to the target data. A sequential feedback approach has finally been proposed in [63], where the user is asked to indicate *only* true matches in the top positions of the initial ranked list (fig. 2.12). In other words, at each iteration the user checks whether the query is present in the ranked list of a single camera view. This feedback and the previous true matches (if any) are used to update the feature representation in the other camera views. A drawback of this approach is that if a true match is not present in the initial ranked list (it may happen in practice), the user should analyse the next k images of the ranked list. This process continues until a true match is found.

I finally points out that although the person re-identification problem can be considered as a CBIR problem, to my knowledge no work has investigated so far the possibility to use existing CBIR-RF techniques to implement the HITL approach.

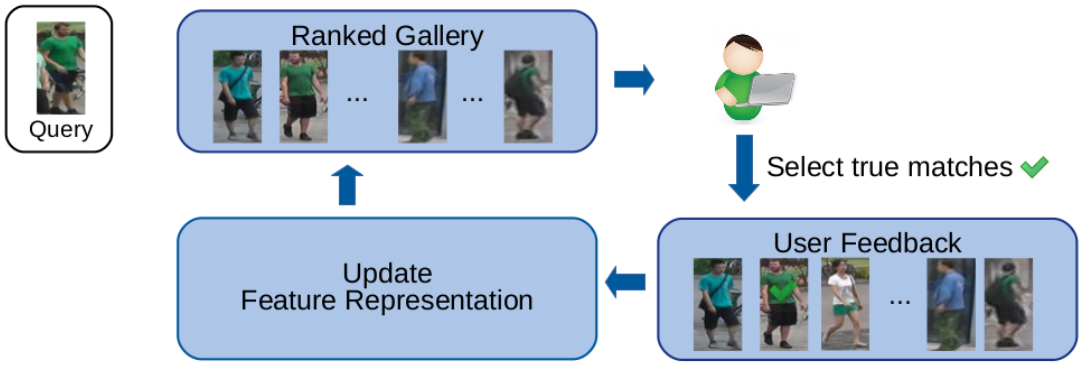


Figure 2.12: Scheme of the feedback approach proposed in [63]. The requested feedback is related only to true matches, and is used to update the feature representation.

2.2 Crowd counting

Crowd counting consists of estimating the number of people in a given image or video frame. In video surveillance systems, automatic crowd counting can support end users in monitoring activities and in guaranteeing the security of people in an area of interest, e.g., during demonstrations, concerts, etc. This computer vision task presents several issues such as different lighting conditions, occlusions, background and perspective distortions. In figure 2.13 some examples of these issues are shown. The lighting condition can change depending on the scene, which can be indoor or outdoor (first and second top images) or can depend on the weather in outdoor scenes (left column). The background of some scenes can be very complex (centre column). The perspective distortion is related to the distance of different cameras to the scene they are monitoring (right column). Some examples of occlusions are shown in figure 2.13 where they are caused by objects such as the palm tree (top centre image), the road sign (top and bottom left corner, bottom right corner), the mobile kiosk (top centre), and by overlapping among people. An additional issue is the crowd size, which can vary from a sparse crowd (a small number of pedestrians with few or limited occlusions) to a large and dense crowd, with sever overlapping between pedestrians.

In real application scenarios, crowd counting is still mainly made “manually”, i.e. human operators (e.g., LEA operators) estimate the number of people in an image or a video by using techniques such as Jacobs’ method [53]. It consists of dividing the image into several parts, determining the number of people in a single part and multiplying

Figure 2.13: Crowd counting issues. Several lighting conditions (left column), different background (centre column), perspective distortions (right column). Static and dynamic occlusions caused by objects (the palm tree, the road sign, the mobile kiosk) and overlapping among people. Images taken from PETS2009 and Mall data sets (more details can be found in sect. 4.3.2).

this number by the number of parts. Such approaches are laborious and require a significant human effort.

Therefore, several computer vision approaches have been proposed so far for automatic crowd counting. Early approaches were based either on pedestrian/head detection or tracking, and on regression models trained on low-level image features. More recent approaches are based on CNN, which usually estimate the crowd count by first estimating a "density map" of the input image.

Earliest methods focused on counting people by detecting their head or full-body, or by using tracking-based techniques. These techniques can be effective on small and sparse crowd, but are not suited to dense crowd [85]. In contrast, regression-based approaches are suited to dense crowd. Most recent methods are based on the use of CNN. They can be categorised into several dimensions, as is the case of person re-identification. Existing CNN architectures for crowd counting are either modifications of known ones (e.g., VGG) [113] or devised ad hoc for this task [116].

Also for this task most of the existing methods have been evaluated on a supervised or same-scene setting, i.e., using training and testing data belonging to the same scene or camera view. Despite the notable results achieved under this scenario, recent research

efforts are focusing on a more realistic cross-scene scenario where the target scene is different from those used for training. Accordingly, crowd counting methods are now often evaluated in a cross-data set setting, to simulate cross-scene application scenarios, which can lead to a significant performance decrease [154].

This issue is due to the fact that in real-world applications it is often unfeasible to collect and annotate a representative set of crowd images of the target scene. To address the cross-scene problem several authors proposed approaches based on DA [114]. However, this approach still requires manually annotated training images of the target scene. Approaches based on UDA have also been proposed [108], whose performance is however worse than that of DA and supervised methods, as one can expect.

As a consequence of the cross-scene issues, another relevant issue is the lack of benchmark data sets representative of different crowd scenes that can be of interest to crowd counting scenarios in video surveillance applications. In particular, only of a few dense crowd data sets are available.

To address similar issues in some computer vision tasks, including crowd counting by detection (which is a different approach than the one considered in this thesis), the use of *synthetic* data sets built by using computer graphics tools has been proposed [12, 79, 111, 122]. This solution can be potentially useful also for regression-based crowd counting, since it would allow to generate synthetic training images of the *target* scene and to automatically control every parameter of interest such as the number and location of pedestrians, the scene perspective, background and illumination. A similar solution has already been proposed, to my knowledge, only in [135] for CNN-based crowd counting, but in the context of a DA method which requires crowd images of the target scene for fine-tuning.

In the next sections, traditional crows counting methods and more recent CNN-based ones are described in details.

2.2.1 Traditional approaches

Early approaches can be divided into counting by detection, counting by clustering and counting by regression [85, 114]. Counting by detection is based on pedestrian detection from still images, either full-body [24, 65] or body part detection [73, 129]. The latter aims to overcome partial occlusions (e.g., overlapping among people) by detecting head and shoulders [36]. However, this approach is only effective for sparse

crowds with limited occlusions [85]. Counting by clustering is based on the idea that coherent feature trajectories can be grouped together to approximate the number of people; also this approach is effective only on sparse crowd scenes [85]. The counting by regression approach aims instead at mapping from low-level image features to the number of people or to the density map of a scene by supervised training of a regression model. Earliest methods estimated the number of people by using features such as foreground segment [86], edge [60] and, texture and gradient [92, 144]. Foreground segments can be obtained by using background subtraction, which is however sensitive to lighting variations. Edge, texture and gradient features provide information about local patterns. Noticeable examples of texture and gradient features are Gray-Level Co-occurrence Matrix (GLCM) [43] and Local Binary Pattern (LBP) [93].

Several regression models have been proposed so far, such as Linear Regression (LR) [67], Partial Least Squares Regression (PLSR) [38], Kernel Ridge Regression (KRR) [48], Support Vector Regression (SVR) [144], Gaussian Processes Regression (GPR) [103] and Random Forest regression (RF) [15]. The simplest model is LR which outputs a linear combination of the input variables; it is however effective only for sparse or small crowds with limited occlusions. The drawback of this method is that its computational complexity increases with data dimensionality, and it is sensitive ti highly co-linear features. To overcome these problems, either a PLSR model or KRR can be used. PLSR introduces a decomposition to maximise the covariance between score matrices, whereas KRR adds a regularisation term to find a trade-off between the correct measure and the penalty. A significant advantage of KRR, SVR and GPR is their higher flexibility. However GPR is not scalable to large data sets as RF, and is more sensitive to parameter values.

2.2.2 CNN-based approaches

Several CNN-based approaches have recently been proposed for crowd counting, based on a regression approach. In particular, most of them do not directly estimate the crowd count of input images, but estimate a density map instead, from which the crowd count is easily derived [114]. The density map consists of an image where each pixel value is proportional to the number of people per unit image area in the original image. To this aim, the ground-truth density map of training images is required, which is obtained by first manually annotating the head locations of each pedestrian; the corresponding

density map is then computed by summing 2D kernels whose volume is normalised to 1 (typically, Gaussian kernels) centred on the head of each pedestrian. Accordingly, at the prediction phase the number of pedestrians in a given input image is obtained by integrating the estimated density map, i.e., by summing up the value of each pixel in the map.

As in person re-identification, also the existing crowd counting approaches can be categorised into several dimensions. A first categorisation can be made according to the CNN architecture, that can be i) a modification of known, "generic" architectures such as VGG [35, 71, 81, 82, 83, 87, 107, 115, 135, 142, 143, 171], or ii) a specifically devised architecture [25, 94, 108, 116, 149, 153, 155, 157, 158]. Most authors proposed architectures (generic or specifically devised) based on branched structures; only in [71, 82, 87, 135, 155] a single-column architecture was proposed. For instance, in [115, 142, 143] a different number of parallel branches has been used, whereas in [107] a tree structure was used. Another possible categorisation is based on the type of information extracted from input images and how this information is combined to obtain the density map. Some authors focused on managing scale variations inside an input image [71, 81, 83, 87, 157]; other authors focused on extracting different features such as low- and high-level features [142, 155, 158], local and global information [153], and information from the region of interest (ROI) [82, 143]. For instance, in [158] the first part the proposed CNN architecture extracts low-level features, which are then used in a module that extracts multi-level information; a selection of the most informative channels is then made, followed by their concatenation. In [82] two concatenated modules have been used. The first one detects regions of interest by using an attention map and a value that indicates the congestion level of the scene. The second one processes the output information of the first module and extracts (in addition) low-level features before estimating the final density map. In [115] the information from different layers and different branches are fused to obtain the density map. The so-called main branch consists of a backbone, whereas the other branches extract information from high- and low-level context in order to fuse them. Also in [155] information extracted from different layers are merged to obtain the density map, but a single branch structure has been used with a single filter size. One additional categorisation can be based on the inference strategy, i.e. whether the CNN is trained by using image patches [82, 87, 94, 108, 142, 149, 153, 157, 158] or whole images [115, 143]. Moreover, data augmentation

techniques are usually adopted to increase the size of training data. The most used ones are random horizontal flipping [94, 153, 158] and random crop [82, 87, 94, 142, 157].

Some authors explicitly addressed the cross-scene issues mentioned above. A common solution to this problem is to use a multi-scene training sets that contain images from different scene [71, 81, 83, 87]. A different solution has been proposed in [117] where a weakly supervised approach has been used to train a CNN model by using six image-level labels (from zero- to very-high density). A DA-based approach has been used by some authors, such as in [135]. However, all the above approaches still require labelled images of the target scene. In [154] an unsupervised approach has been proposed, which however still requires representative images (albeit unlabelled) of the target scene. Such images are used to re-train or fine-tune the crowd counting model to improve its performances on the target scene. However, as mentioned above, in some challenging real-world application scenarios it can be difficult or infeasible to collect representative (even if unlabelled) images of the target scene. To mitigate this issue, the use of synthetic images has been proposed so far for other computer vision tasks including crowd behaviour analysis, pedestrian detection and person re-identification [12, 14, 44, 58, 79, 109, 111, 122]. The use of synthetic images has also been proposed for CNN-based crowd counting to my knowledge, only in [135], where a synthetic data set has been generated using the Grand Theft Auto V (GTA5) video game to pre-train a CNN model. However, this approach still requires images of the target scene, although unlabelled.

Chapter 3

Person re-identification with a human in the loop

In this chapter, the person re-identification with human-in-the-loop approach proposed in this work is described. First of all, it is necessary to focus on the goals of this thesis and on the considered application scenario. As described in the previous section, the aim of person re-identification system is to support human operators by limiting the time required to analyse all recorded videos. In this thesis, a challenging cross-scene scenario was considered when collecting a suitable amount of representative (even unlabelled) images of target camera view is unfeasible or too demanding for the end users, especially if the system should be operational in a short time. Under this scenario, it is not even possible to use UDA approaches, since they require images of the target scene, albeit unlabelled. In the considered scenario the human-in-the-loop (HITL) approach seems an interesting alternative to adapt a person re-identification system to the target scene, taking into account that such systems inherently include interaction with a user. The main idea of this approach is that of combining the complementary strength of machines and humans. In particular, since the person re-identification problem can be considered as an image retrieval problem, this thesis focuses on CBIR-RF techniques, which have been disregarded so far in literature. As a specific contribution, a feedback protocol different from those of existing HITL approaches to person re-identification is proposed, which takes into account the characteristics of both CBIR-RF techniques and person re-identification systems.

It is worth noting that the HITL approach in the considered application scenario

presents analogies with online domain adaptation (ODA) approach that has been investigated in several computer vision tasks such as face detection [55] and pedestrian detection [148]. For instance, in ODA methods target data are available only during system operation (online) and usually source data are not reused online for system update. For the above reasons, the HITL approach can also be viewed as a kind of ODA approach.

In next sections, the proposed HITL approach is described, followed by its empirical analysis.

3.1 Approach

As mentioned in previous sections, the HITL approach is based on the interaction with a user. It is an interesting approach to be investigated in person re-identification systems, since they include a inherent interaction with an operator. However, it has been proposed so far by only few authors. In a person re-identification system, the input query (selected by the user) is compared against the images contained in the gallery by using a similarity measure, as shown in figure 2.7. The feature extraction module can be a hand-crafted descriptor (section 2.1.1) or a CNN (section 2.1.2). The similarity measure module can be implemented in different ways, as described in section 2.1. The comparison between query and gallery images defines a ranked list based on their similarity.

In this context, a HITL method can be implemented by adding two other steps: user feedback and gallery re-ranking. Different feedback protocols have been proposed so far (more details in sect. 2.1.3) and all of them aim to update and improve the ranking of the template gallery (gallery re-ranking module). The updated ranked list has to be analysed again by the user (fig. 2.7).

In the following the main features and limits of existing HITL approaches to person re-identification, described in sect. 2.1.3, are pointed out. Limits can be categorised in two main groups: human effort and resource consuming. In [2] the user should label all images of the top-k ranked gallery as relevant and non-relevant. This is not practical in many real application scenario. In [47] a classifier has to be learnt for each query by using top-k ranked images as negative examples, requiring significant resources. A similar drawback can be found in [77] since a graph over all gallery images has to be

build. This approach is not scalable when a large gallery is used. All above mentioned approaches do not use an incremental approach, which instead is proposed in [134]. Although this approach can improve its performance, a complex optimisation problem has to be solved after each feedback to update the similarity metric. A similar approach has been proposed in [63], where a sequential methods is used. In contrast to above mentioned approaches, the ranked gallery shown to to the user is not limited. In other words, the user should analyse all images of the template gallery until a true match is found. This requires a huge human effort, especially when the template gallery is large.

Let's start with the consideration that the person re-identification task can be seen as a problem of image retrieval since the aim is to retrieve images of pedestrians similar to the query [163]. The kind of user interaction that can be used in person re-identification systems is similar or comparable to the one used by CBIR-RF algorithms, that have been used since a long time in the image processing field [99]. Additionally, many CBIR-RF algorithms present a relatively low computational complexity with respect to existing HITL methods for person re-identification. Despite this, they have not been considered so far for person re-identification, except for some experimental comparisons with the proposed HITL techniques [77, 134]. Toward the adoption of CBIR-RF algorithms for person re-identification, the most relevant aspect to investigate is the feedback protocol.

Feedback Protocol. Almost all existing HITL methods propose to use a feedback protocol that in this thesis is called "single-feedback", which consists of requiring the user to select a single image (and *optionally* another one) on the top-k of the ranked list. This protocol was analysed since it requires a small human effort. The single-feedback protocol considered in this thesis consists of asking the user to select a single image, that can be either the true match (if any) or a strong negative (i.e. a person image very different from the query) image in the top-k of the ranked list. However, this kind of approach appears sub-optimal for RF algorithms since they benefit from a large amount of feedback. In many real application, it is not possible to ask a user feedback on a large amount of images. To find a trade-off between a sufficient amount of images and the human effort, it can be useful to focus the human feedback on the top-k images; moreover, it is necessary to take into account two cases: i) the presence of true matches and ii) the absence of them. In the first case, the user can be interested to retrieve other true matches. In the second case, it is necessary take into account

that some RF algorithms require relevant and non-relevant images. For this reason, the feedback protocol proposed in this thesis consists of asking the user to select *all* true matches (if any) in the top-k ranked list, whereas the other images are *automatically* considered as non-relevant [21]. Accordingly, call this protocol *multi-feedback*. Examples of the two kind of feedback are reported in figure 3.1. In the first two rows, two examples of single-feedback are shown. In the first one, the user selects the first strong negative found, whereas in the other row a single true match (the top-ranked one) is selected. In the last row of fig. 3.1, the user selects *all* true matches in the top ranked list and the remaining images are *automatically* labelled as negative.

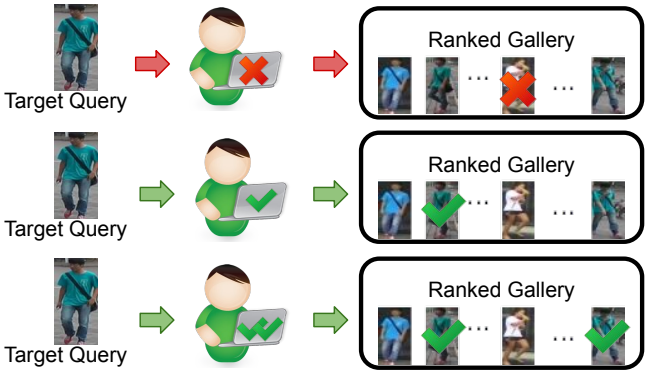


Figure 3.1: Single-feedback protocol: the user selects either a positive match (if any) among the top-k gallery images (middle row) or a strong negative (top row). Multi-feedback protocol (bottom row): the user selects all true matches (if any), and the remaining images are automatically labelled as negatives.

The proposed multi-feedback protocol may appear too demanding for the end user since it requires to select all true matches in the top-k ranks. Nevertheless, the value of k considered in existing HITL person re-identification methods is small. Moreover, for a very large template galleries that are typical of real-world applications involving hours of video footage acquired by many different surveillance cameras, the number of true matches in the top-k ranks is likely to be very limited, especially in the first feedback iterations. I also point out that selecting all true matches reflects application scenarios related to forensic investigations in which users want to retrieve *all* the images of query identity, e.g., to reconstruct the movements of the person of interest in the monitored area.

Now it is necessary to point out some aspects. In CBIR-RF algorithms, it is possible to use incremental techniques as in [134]. This allows to incrementally update the system module modified by the considered RF algorithm (e.g., similarity measure). In the one hand, the system performance improves. In the other hand, the processing time increases. For this reason, this thesis focuses on non-incremental RF algorithms. From a point of view, some CBIR-RF algorithms present a sort of limitation since they require relevant and non-relevant images. Nevertheless, they can be used under the proposed feedback protocol by adapting them (if necessary) to take into account if no true matches (relevant images) are found in the top-k ranked gallery. Therefore, it is possible to use CBIR-RF algorithms also in the case mentioned above, preserving a low processing time. With regard to possible low performances, it is possible to consider several feedback rounds. However, in this thesis the required human effort is take into account, and, for this reason, a fixed value of feedback rounds is considered.

3.2 Experiments

In this section, some well known CBIR-RF algorithms are considered to implement the HITL approach based on the proposed feedback protocol. A direct comparison should be done by using the existing HITL approaches at the state of the art. However, the corresponding codes were not made available by the authors and it is not easy re-implement these approaches by using information provided by authors. In addition, existing HITL approaches were applied on different and smaller data sets than the current benchmark ones. For these reasons, I had to renounce this comparison. Methods suited to a scenario that can be similar to that considered in this thesis are selected for comparison. These methods, i.e. UDA, however still require images of the target scene.

To simulate the considered cross-scene scenario, cross-data set experiments were carried out by using benchmark data sets (sect. 3.2.3). This kind of experiments consist of using one data set as the source domain and a different one as the target domain (target scene). For a fair comparison, it was necessary that the considered approaches present the same model trained on the source domain, and that the comparison is made on the test set of the target domain. For UDA approaches unlabelled images of training set of the target domain were used to fine-tune the underlying model. In contrast, RF

approaches were directly applied on the test of the target domain. In other words, their model was trained on the source domain only, and applied on the test of the target domain without any fine-tuning.

In the next sections, the person re-identification methods (UDA and HITL) considered in this work are described in details, as well as benchmark data sets and performance metrics. Then, a discussion about results and about the HITL approach is given.

3.2.1 Person re-identification methods

As mentioned in the previous section, for a fair comparison between HITL and UDA methods it is necessary to use for HITL methods the same UDA model trained *only* on the source domain. Among existing UDA methods only for Mutual Mean Teaching (MMT) [37] and Exemplar Camera Neighbour (ECN) invariance [166] the cose made available by the respective authors allows also to train the underlying model on the source data only. For this reason, only these two UDA methods have been used in these experiments. Both MMT and ECN use ResNet-50 pre-trained on ImageNet [106] as a backbone. **MMT** is a cluster-based approach that employs soft- and hard-pseudo labels in which pseudo-labels and better features are progressively learned. The architecture consists of two networks, initialised by using different weights, that are jointly learnt on the source domain. During this phase, hard pseudo-labels are generated for the target images. Simultaneous training is the key idea of the teacher-student approach proposed by [127]. Then, these networks are jointly trained on the target domain to predict soft pseudo-labels, starting from the hard ones. The final model of MMT is obtained by averaging models mentioned above and it is used as a feature extractor during the testing phase. **ECN** consists of a classification and exemplar memory modules developed after the ResNet backbone. The classification module is used for source data, therefore for labelled images. The other module is used for invariance learning and for target data (unlabelled images). The considered kinds of invariances are exemplar-invariance, camera-invariance and neighbour-invariance. The first one aims to distinguish the same identity from the others. The camera-invariance forces images of the same person but from different cameras to be close. The last one aims to relate similar images in feature space.

3.2.2 Relevance feedback algorithms

Concerning the HITL approaches, the Query Shift (QS) [74], also known as Rocchio, the Relevance Score (RS) [41], Efficient Manifold Ranking (EMR) [147], Passive-Aggressive (PA) online learning and Means-Relevance Score (M-RS) were considered. **QS** is a the classic RF algorithm [105]. Although it was not designed for CBIR systems, it was adopted in this kind of systems later and it is still widely used. QS is a cluster-based algorithm which works in feature space. It exploits user feedback to move the query near to positive samples and far from negative ones. Therefore, the query x_q is moved near to the positive (relevant) images and far from the non-relevant ones accordingly to the user feedback on the positive (N_p) and negative (N_n) images

$$x_q^{new} = \frac{1}{N_p} \sum_{i \in Np} x_i - \frac{1}{N_n} \sum_{i \in Nn} x_i \tag{3.1}$$

where N_p and N_n are the sets of positive and negative images, respectively, and x_i denotes the feature representation of the i-th image. **RS** is an effective RF algorithm and it is still widely used among algorithms that compute a score for each image [99]. RS uses the Nearest-Neighbour (NN) approach and defines, for each image, a relevance score by using distances in feature space [41]. This algorithm allows to increase the distance between positive and negative samples and produces a small distance between positive images. Formally, RS computes a relevance score $s_{NN}(x)$ as follows:

$$s_{NN}(x) = \frac{\|x - x_n^{NN}\|}{\|x - x_p^{NN}\| + \|x - x_n^{NN}\|} \tag{3.2}$$

where x_p^{NN} and x_n^{NN} denote the nearest positive and negative neighbouring images, respectively, and $\|\cdot\|$ is the metric used in the feature space, for example Euclidean distance. QS and RS do not use an online learning phase, and require a low processing time. **EMR** belongs to Manifold Ranking (MR) approaches which graph-based models are used [146]. The EMR algorithm was originally not designed for RF, but is considered in these experiments for completeness, since it was used for comparison in [134]. It is a graph-based Manifold Ranking (MR) algorithm but, rather than using the k-NN algorithm, it uses k-means which exhibits a lower complexity [147]. Moreover, to speed-up the calculation a lower-dimensional feature space is considered. The ranking function r used in EMR is defined as follows:

$$r = (I_n - \alpha H^T H)^{-1} y \tag{3.3}$$

where α is a smoothing parameter, $H = ZD^{-\frac{1}{2}}$ and y is the score vector. Z is the matrix representing the new feature space and D is a diagonal matrix where each element represents the sum of correlation between the i-th and j-th elements. **PA** algorithm is an interpretation of RF feedback in term of passive aggressive online learning approaches [97]. The PA algorithm is considered in this thesis to analyse its behaviour in a person re-identification system and to verify whether the images of query individual can be *linearly* separated to the other ones. In contrast to the other algorithms, it is based on online learning and its aim is to maximise the following objective function:

$$\sum_{\forall p \in N_p} \sum_{\forall n \in N_n} w(x_p - x_n) \tag{3.4}$$

where N_p and N_n are the sets of positive and negative images, and w is weight vector. The goal of PA is to adapt the weight vector in each iteration to obtain that behaviour $w \cdot x_p > w \cdot x_n$. In other words, it defines a linear function that separates positive from negative images. **M-RS** is a modification of RS approach evaluated in this work. It is considered to investigate if it is possible to create hyperspefic clusters. In contrast to RS, M-RS determines only two clusters (relevant and non-relevant). M-RS consists of using the mean of positive and negative images instead of the nearest positive and negative images to the query to compute the score (s_{M-RS}):

$$s_{M-RS} = \frac{\|x - \mu_n\|}{\|x - \mu_p\| + \|x - \mu_n\|} \tag{3.5}$$

where μ_p and μ_n denote the mean of positive and negative images, respectively.

3.2.3 Data set

Performances have been evaluated on three benchmark data sets, namely DukeMTMC-reID, Market-1501 and MSMT17.

Duke Multi-target multi-camera, known as **DukeMTMC-reID** consists of 1404 identities and (IDs) acquired from 8 cameras [159]. A Faster RCNN [104] was used to extract all pedestrian bounding boxes. **Market-1501** consists of 32,668 bounding boxes and 1501 IDs acquired from six cameras placed in front a supermarket [162]. A Deformable Part Model (DPM) was used to extract all bounding boxes from videos [30]. Market-1501 contains also "distractors" and "junk" images which present extensive variations in pose, resolution, etc. Distractors and junk images also denote false alarm

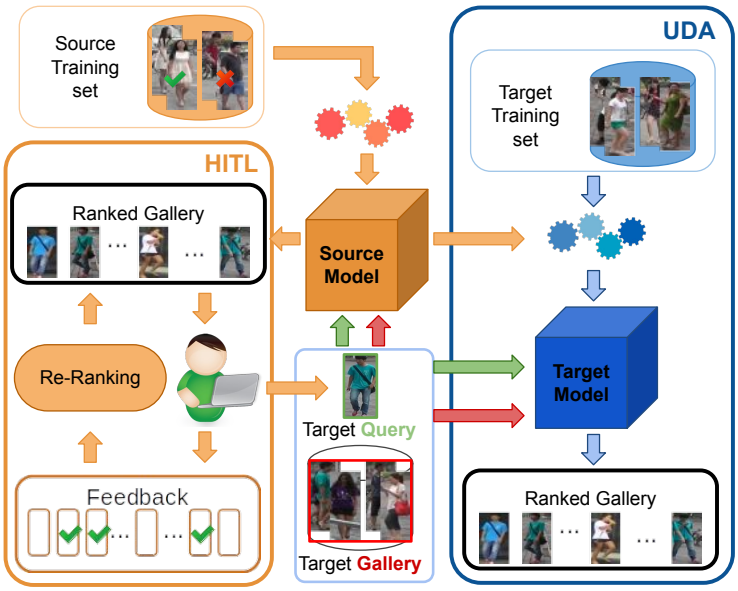


Figure 3.2: High-level view of UDA and HITL approaches to person re-identification. They both start from a model trained offline on source data. UDA refines it offline (before deployment) using unlabelled target data. HITL refines online (during operation) the ranked list of target gallery images provided by the source model, exploiting user's feedback (online updating of the distance metric, external to the source model, is also carried out in [134]).

or irrelevant images. The Multi-Scene Multi-Time (**MSMT17**) data set is larger and more recent than the previous ones. It consists of 126,441 bounding boxes and 4,101 identities acquired by using 15 cameras (12 outdoor and 3 indoor) [139]. All recorded videos were acquired with different weather conditions and at different times, and processed to extract all bounding boxes by using Faster RCNN [104].

Further details about the above data sets can be found in tab. 3.1.

	#IDs / #images			#Cameras
Data set	**Train**	**Query**	**Gallery**	-
Market-1501	751 / 12936	750 / 3368	751 / 15913	6
DukeMTMC-reID	702 / 16522	702 / 2228	1110 / 17661	8
MSMT17	1041 / 30248	3060 / 11659	3060 / 82161	15

Table 3.1: Data sets details: number of identities and of images (# IDs / # images) in the training, query and gallery sets, and number of cameras.

3.2.4 Metrics

Performances are evaluated by using two common metrics: Cumulative Matching Characteristic (CMC) curve at rank 1, 5, 10, 20 and mean average precision (mAP).

The CMC describes the likelihood of finding a correct identity within the top ranks. It is defined as the sum of $P(k)$:

$$CMC(k) = \sum_{r=1}^{k} P(k) \tag{3.6}$$

where $P(k)$ denotes the probability of associating the identity at the rank k.

The mean average precision is defined as follow:

$$mAP = \frac{1}{Q} \sum_{q=1}^{Q} AveP(q) \tag{3.7}$$

where Q is the number of queries and $AveP(q)$ is the average precision for a given query q.

3.2.5 Experimental Setup

Cross-data experiments were carried out to simulate the considered cross-scene scenario. Each of the considered data sets, in turn, was used as the source domain and the other ones as the target domain. As in [134], the user is asked a feedback on the top-k rank gallery, with $k = 50$. The proposed multi-feedback protocol is compared with the single one, used in the existing HITL approaches. In these experiments, some algorithms considered in this thesis, such as QS, RS and EMR, are evaluated under the single-feedback protocol since they were considered for experimental comparison in [77, 134]. However, it is necessary to point out that the single-feedback appear sub-optimal for RF algorithms that usually benefit from a large number of feedback. Simulating the user feedback, e.g., using volunteers, would have required a considerable human effort and a long time. Therefore, the feedback has been simulated by exploiting the ground truth of the considered data sets. In particular all the gallery images (multi-feedback) labelled with the same identity of the query image, among the top-k ranked ones, have been *automatically* selected as true matches, without the involvement of human users. On the one hand, this can be seen as a best-case feedback scenario, when an operator never makes errors in recognising the true matches. On the other hand,

this is not an unrealistic scenario; moreover, the proposed feedback protocol does not require the selection of *strong* or *weak* negatives as other HITL methods for person re-identification, which are much more difficult to simulate, instead (using the similarity scores to distinguish strong and weak false matches would not be effective, since the scores of top-k ranked images are usually very similar for values of k much smaller than the gallery size).

For performance evaluation, a subset of the query set from the test set of the considered data sets was used, due to the time required to collect user feedback. As in [134], the number of the considered queries was 300. Moreover, three feedback rounds were carried out, as in [134].

To further assess the performance of the considered RF algorithms, a value of k (lower than 50) was also considered, in particular, $k = 10$. In this case only the best RF algorithm under the multi-feedback protocol was evaluated. With regard to the UDA methods, they have been evaluated using the same number of queries mentioned above, for a fair comparison. For a complete overview of results, the baseline results are also reported, denoted as "source model", which corresponds to a model trained on a source data set and directly applied on a target data set (without using UDA nor HITL).

3.2.6 Results

In this section, results are reported and discussed. As mentioned above, experiments aimed to evaluate the effectiveness of the HITL approach to person re-identification implemented using CBIR-RF algorithms with the proposed feedback protocol, and to compare its performance to that of UDA methods which can also exploit unlabelled target images for model fine-tuning.

Comparison between HITL-RF and UDA approaches. Table 3.2 reports the overall results obtained by the baseline (source model), UDA methods (MMT and ECN) and the CBIR-RF algorithms used to implement HITL (QS, RS and EMR). In particular, results of HITL approaches after three feedback round under both single- and multi-feedback protocols are reported. In general, all methods achieved better performances when the target domain is Market, and DukeMTMC-reID is the source domain. UDA methods outperformed the source model (table 3.2). Moreover, MMT outperformed ECN in all ranks and in mAP in both target data sets. Also HITL

methods outperformed the source model in both target data sets, and achieved better performances under the proposed multi-feedback protocol than under single-feedback. By comparing each HITL approach under the two feedback protocols, it is possible to make the following observations. QS achieved a performance improvement from single to multi-feedback of around 9% in mAP and about 6.92% on average in all ranks when the target domain is DukeMTMC-reID; when Market-1501 is the target domain, the improvement was about 13.74% in mAP and 7.25% on average in all ranks. RS achieved the highest performance improvement when multi-feedback was used. Indeed, when DukeMTMC-reID is used as target domain, it achieved an improvement of around 18.07% in mAP and 9.67% on average in all ranks, and an improvement of 33.4% in mAP and 12.17% on average in all ranks when Market-1501 was the target domain. In contrast to QS and RS, EMR achieved limited performance improvements.

These results confirmed that CBIR-RF approaches benefit from a larger amount of feedback compared to the one used in HITL re-identification methods at the state of the art. Comparing HITL methods under the multi-feedback protocol with UDA methods, it can be seen that RS outperformed the best UDA method (MMT) in all rank and in mAP when DukeMTMC-reID was used as target data set. In contrast, when Market-1501 is the target data set, RS outperformed MMT in mAP and rank-1 only (tab. 3.2). It is worth reminding the reader that mAP measure gives a more complete overview than CMC curve of the ranks of *all* the images of the query identity in the ranked gallery: therefore, even if MMT outperformed RF for ranks 5, 10 and 20, the higher mAP achieved by RS means that the re-ranking obtained by using the proposed HITL feedback protocol is more effective than using of a large amount of unlabelled target images offline to fine-tune the source model.

A more detailed analysis of the results of CBIR-RF algorithms after each round under single- and multi-feedback protocols in Tables 3.3 and 3.4, respectively. By comparing such results to the one of the source model (tab. 3.2), it is possible to make the following observations. Under the single-feedback protocol, after the first feedback round QS and RS achieved the highest improvement in Market-1501 and DukeMTMC-reID, respectively (tab. 3.3). In particular, RS achieved an improvement of around 12.34% in mAP and about 10% on average in all ranks. In contrast, it achieved a lower improvement in each rank than QS when Market-1501 was the target domain. The improvement obtained by RS and QS when Market-1501 was used as a target, was

comparable in mAP and rank-1, but in the other ranks on average it is better for QS than RS. Under the multi-feedback protocol, after the first round, for both target data sets RS outperformed by a higher amount than the other RF algorithms (see tab. 3.4). RS achieved an improvement of 28.62% (31.15%) in mAP and around 21.25% (17.33%) on average in all ranks when the target was DukeMTMC-reID (Market-1501). The other algorithms achieved similar improvements with respect to the source model both in mAP and in all ranks when DukeMTMC-reID was the target domain, except for EMR when Market-1501 was the target domain: in this case, the improvement achieved by EMR in mAP was lower than that of QS. Although QS achieved a limited improvement over the feedback rounds compared to RS, it benefited from the multi-feedback protocol. In particular, for both target domains the improvement was 7% on average in all ranks, whereas mAP improved 9% on DukeMTMC-reID and 14% on Market-1501.

With regard to HITL and UDA, by comparing tables 3.3 and 3.2, it can be seen that RS achieved better results than ECN model in all ranks (except for rank-10 and rank-20) and in mAP since after second round, when DukeMTMC-reID was the target data set, whereas when Market-1501 was used as the target data set RS outperformed ECN only in mAP and rank-1. Among the other HITL algorithms, only QS (with the single-feedback protocol) achieved comparable results to ECN, overcoming it only in rank-1 for both target data set since after the second round. A similar behaviours can be observed for RS under the multi-feedback protocol (table 3.4). Indeed, it outperformed ECN in all ranks and mAP in both target data sets since after the first round. RS outperformed MMT in all ranks and mAP when DukeMTMC-reID was used as a target since after the second feedback round, whereas it exhibited results slightly lower than MMT when Market-1501 was the target data. Also, when Market-1501 was the target domain, the performances of QS improved under the multi-feedback protocol. In particular, it outperformed ECN in all ranks (except for rank-20) and mAP since after the first round. Instead, when DukeMTMC-reID is used as the target data set, QS outperformed ECN in mAP, rank-1 and rank-5 after the first round.

In general, improvements achieved by QS and EMR in each round were lower than those achieved by RS. In particular, QS exhibited a slight improvement among rounds under both feedback protocols, whereas the improvements of EMR were more limited than those of RS.

Table 3.2: Results of cross-data set experiments (source → target) for the source model (baseline), UDA methods (ECN, MMT), and HITL methods (QS, EMR, RS) after three rounds of single- and multi-feedback protocols. Best results in each column are highlighted in bold.

Method	Market-1501⟶ Duke					DukeMTMC-reID ⟶ Market				
	mAP	Rank-1	Rank-5	Rank-10	Rank-20	mAP	Rank-1	Rank-5	Rank-10	Rank-20
Source model	29.1	47.7	61.3	66.0	72.0	25.5	54.3	72.0	79.0	81.7
ECN	43.2	66.7	77.3	81.0	82.7	34.9	64.3	80.3	86.0	91.0
MMT	60.8	76.0	85.3	88.0	90.3	69.4	87.0	**95.3**	**97.0**	**97.7**
QS-single	42.71	68.67	74.33	76.33	78.0	33.9	71.0	78.0	82.0	84.0
EMR-single	36.79	72.33	72.67	73.33	73.67	30.41	70.33	72.0	73.33	75.33
RS-single	56.6	82.33	82.67	83.0	83.67	41.69	77.0	80.33	81.33	85.0
QS-multi	51.74	73.67	82.67	83.67	85.0	47.64	80.67	87.33	88.0	88.0
EMR-multi	47.23	74.0	74.33	74.33	74.33	36.13	70.67	71.67	72.0	72.0
RS-multi	**74.67**	**92.0**	**92.67**	**92.67**	**93.0**	**75.09**	**92.67**	92.67	93.33	93.67

These results showed that the HITL approach is able to reduce the performance gap in a cross-view scenarios, with respect to the UDA approach, by using the user feedback acquired *online* during system operation, even through relatively simple RF algorithms, and without requiring any (even unlabelled) target image *offline*, during system design. To this aim, it is necessary to clarify some details about the required user's feedback under the multi-feedback protocol. In the experiments, 18 positive matches on average were present among the top-50 ranks at the first feedback round. This means that the overall number of feedback the user had to provide in next feedback rounds was $k - 18$. After the first round, these images were present and then the user did not have to select them again. By comparing the number of images required by RS and UDA approaches, UDA approaches use a considerable amount of unlabelled training images of the target data. The number of unlabelled images is about two orders of magnitude larger than those used by RS.

Since HITL methods was more effective than UDA ones in term of the number of target images used to improve the results of the source model, the performances of HITL methods were further evaluated also for a lower value of k that the one ($k = 50$) considered in the previous experiments. In particular, $k = 10$ was considered which requires the user a significantly lower effort. In this experiment only RS approach was used since it achieved better results than the other ones under both feedback

Table 3.3: Results of cross-data set experiments for the HITL methods after each round of the **single**-feedback protocol. Best results in each column are highlighted in bold.

Method	Round	Market-1501 ⟶ DukeMTMC-reID					DukeMTMC-reID ⟶ Market-1501				
		mAP	Rank-1	Rank-5	Rank-10	Rank-20	mAP	Rank-1	Rank-5	Rank-10	Rank-20
QS	1	40.49	64.33	70.0	73.0	76.0	29.59	61.33	73.67	79.33	82.67
	2	41.79	69.0	74.33	76.0	78.33	32.4	67.33	75.0	79.67	83.0
	3	42.71	68.67	74.33	76.33	78.0	33.9	71.0	78.0	**82.0**	84.0
EMR	1	33.24	61.67	64.33	67.33	70.0	23.67	50.67	60.0	63.67	68.33
	2	36.62	68.67	71.0	72.33	73.33	27.93	65.33	68.33	70.33	74.0
	3	36.79	72.33	72.67	73.33	73.67	30.41	70.33	72.0	73.33	75.33
RS	1	41.44	65.67	70.33	74.0	77.33	29.09	61.0	70.0	76.33	80.67
	2	49.51	75.67	79.0	79.0	80.33	37.11	72.67	74.67	78.33	82.0
	3	**56.6**	**82.33**	**82.67**	**83.0**	**83.67**	**41.69**	**77.0**	**80.33**	81.33	**85.0**

Table 3.4: Results of cross-data set experiments for the HITL methods after each round of the **multi**-feedback protocol. Best results in each column are highlighted in bold.

Method	Round	Market-1501 ⟶ DukeMTMC-reID					DukeMTMC-reID ⟶ Market-1501				
		mAP	Rank-1	Rank-5	Rank-10	Rank-20	mAP	Rank-1	Rank-5	Rank-10	Rank-20
QS	1	47.68	71.0	79.0	79.67	81.33	43.23	79.67	85.67	86.67	88.0
	2	50.89	73.67	81.67	83.33	85.0	46.9	81.0	87.33	88.0	88.0
	3	51.74	73.67	82.67	83.67	85.0	47.64	80.67	87.33	88.0	88.0
EMR	1	46.27	76.0	76.33	76.33	76.33	34.86	71.0	71.33	71.33	71.33
	2	47.19	74.33	74.33	74.67	75.33	36.14	73.67	74.0	74.33	74.33
	3	47.23	74.0	74.33	74.33	74.33	36.13	70.67	71.67	72.0	72.0
RS	1	57.72	81.0	83.0	83.67	84.33	56.65	88.0	88.33	89.0	91.0
	2	68.06	87.67	88.0	89.0	90.33	68.44	91.33	92.33	92.33	92.67
	3	**74.67**	**92.0**	**92.67**	**92.67**	**93.0**	**75.09**	**92.67**	**92.67**	**93.33**	**93.67**

protocols. Table 3.5 reports cross-data set results for the source model, the two UDA methods, and RS under the multi-feedback protocol applied to the top-10 images in the ranked gallery. By using $k = 10$, RS was outperformed by MMT approach in both target domains (except for rank-1, when DukeMTMC-reID was the target domain). Nevertheless, the improvement of RS with respect to the source model was higher than 20% in mAP and 30% in rank-1, on both target data sets, after the third round; moreover, since the first round RS achieved an improvement of more than 13.5% in mAP and 20% in rank-1 for both target domains.

Moreover, further feedback rounds were carried out to investigate whether they could provide a further performance improvement, but the observed improvements (not

Table 3.5: Results of cross-data set experiments for the source model, UDA methods (ECN and MMT), and the HITL method RS after each round of the **multi**-feedback protocol on **top-10** gallery images. Best results in each column are highlighted in bold.

Method	Round	Market-1501 ⟶ DukeMTMC-reID					DukeMTMC-reID ⟶ Market-1501				
		mAP	Rank-1	Rank-5	Rank-10	Rank-20	mAP	Rank-1	Rank-5	Rank-10	Rank-20
Source model	–	29.1	47.7	61.3	66.0	72.0	25.5	54.3	72.0	79.0	81.7
ECN	–	43.2	66.7	77.3	81.0	82.7	34.9	64.3	80.3	86.0	91.0
MMT	–	**60.8**	76.0	**85.3**	**88.0**	**90.3**	**69.4**	**87.0**	**95.3**	**97.0**	**97.7**
RS	1	43.19	69.67	73.33	75.33	77.67	39.32	76.67	80.33	81.67	83.0
	2	48.43	75.67	78.0	79.0	81.67	44.32	81.67	83.33	86.0	86.33
	3	50.9	**79.0**	81.67	82.67	85.0	47.95	85.33	87.0	87.67	89.0

reported here) were very limited and therefore would not justify the corresponding user effort.

Further investigation on relevance feedback algorithms. The first set of experiments aimed to assess whether the HITL approach implemented using RF algorithms under the proposed feedback protocol can be an effective alternative to UDA in the considered scenario. Since they exhibited good performances, to extend this analysis other RF algorithms were considered, PA and M-RS. They were evaluated *only* under the multi-feedback protocol, which clearly outperformed the single-feedback one in previous experiments. In particular, PA is an algorithm that uses an online metric learning approach, whereas M-RS is a simple modification of RS that, in contrast to RS, defines only two clusters (relevant and non-relevant). These experiments were carried out on the same data sets used in the previous experiments, as well as on to another data set (i.e., MSMT17) that is the largest one among all data set considered in this thesis. Tables 3.6, 3.7 and 3.8 report results of these experiments. For an easier comparison, the results of the previous set of experiments are also reported. In general, PA and M-RS confirmed that RF under the proposed feedback protocol is more effective than UDA. By comparing results obtained by RF algorithms, it is possible to note that, as expected, M-RS exhibited a similar performance to RS on both target domains, since it is a simple modification of RS. In contrast, PA achieved better performances than those obtained by other RF algorithms. In particular, RS and M-RS achieved after the third round performances comparable to those achieved by PA after the second round.

Before further analysing the results on MSMT17 I focus on the two previous data sets, by comparing the performances of PA and M-RS against UDA and RF algorithms.

These results are reported in table 3.6 (left) and table 3.7 (left). Both PA and M-RS achieved good performances compared to UDA methods (MMT and ECN). PA outperformed MMT in mAP and in all ranks when the target domain was DukeMTMC-reID, after the third round. In particular, it achieved an improvement of 20% in mAP and on average around 9.2% in all ranks. M-RS achieved similar results; indeed, it outperformed MMT in mAP and in all ranks. Moreover, both PA and M-RS outperformed ECN since the first round. It can also be observed that PA achieved the best results when Market-1501 and DukeMTMC-reID were used as source and target domains, respectively. In contrast, these when DukeMTMC-reID and Market-1501 were the source and target domains (tab. 3.7). PA and M-RS outperformed MMT only in mAP and in rank-1, whereas they outperformed ECN in all ranks (except for rank-20 of M-RS) and mAP.

Regarding the performances of RF algorithms obtained by using MSMT17 either as the source or the target domain, in most cases they were better than those of both UDA methods. In particular, ECN was outperformed in mAP and in all ranks since after the first round when MSMT17 was either the source domain or the target domain as shown in table 3.6, table 3.7 and table 3.8. Only QS did not outperform ECN, in rank-20, when MSMT17 was the target domain and DukeMTMC-reID was the source domain, as shown in table 3.8 (right). Moreover, performances achieved by RF algorithms are comparable with each other, except for QS that exhibited in most cases slightly lower results than the other ones.

Considering the results achieved when MSMT17 was used as source domain (tables 3.6, 3.7 - right), it is possible to note that the best performances was achieved by one or more RF algorithms. In particular, when the target domain is DukeMTMC-reID the best performances was achieved by PA in terms of mAP and rank-1, and by M-RS in the other ranks (after the third feedback round). Moreover, these algorithms outperformed MMT of around 25% (PA) and 21% (M-RS) in mAP, after the third round. In some cases, RF algorithms outperformed UDA methods since after the first round. For instance, this is the case in which DukeMTMC-reID was used as the target domain (see table 3.6 - right). When Market-1501 was used as target domain, three RF algorithms outperformed MMT (table 3.7 - right). In particular, the best result in terms of mAP was achieved by PA, which showed an improvement (with respect to MMT) of around 23%, whereas the best results in all ranks were achieved by RS

and M-RS, showing an improvement of around 5%. Also, in this case, RF algorithms outperformed UDA methods before the third feedback round. In particular, M-RS and PA outperformed MMT in mAP and in all ranks since the second round, whereas RS did not outperform MMT only in rank-20. Only QS exhibited lower performances than the other algorithms and did not outperform MMT. In other words, after the third round, QS achieved similar performances that the other RF algorithms exhibited after the first round. It is worth noting that, when MSMT17 was used as the target domain, the performances of all the considered HITL and UDA methods decreased. In particular, UDA methods exhibited a significant performance decrease (see table 3.8), whereas the one of RF algorithms was more limited. More precisely, UDA methods showed a reduction in mAP of one order of magnitude with respect to RF algorithms. The best performances were achieved by M-RS and PA when Market-1501 and DukeMTMC-reID were the source domains. In the first case (i.e., when Market-1501 was the source domain), PA achieved the best results in terms of mAP, rank-1, rank-5 and rank-10, whereas for rank-20 the best performances were achieved by M-RS (table 3.8 - left). In particular, after the third round PA and M-RS exhibited an improvement with respect to MMT of around 17% and 13%, respectively, in mAP, and and of about 23% on average in all ranks. RF algorithms outperformed UDA since the first or the second feedback round in mAP and in all ranks, in most of the cases. Also in the other case (DukeMTMC-reID as the source domain) PA and M-RS achieved the best results (table 3.8 - right). In particular, PA achieved the best result in terms of mAP and M-RS in all ranks, after the third feedback round. Although M-RS is a simple modification of RS, it achieved an improvement with respect to RS of around 5% in mAP and in all ranks when DukeMTMC-reID was the source domain.

Influence of pedestrian detection and bounding box extraction. During the above experiments, some issues related to the considered data sets were noticed, that might penalise the HITL approach more than UDA. Although different images of the same individual present typical issues for person re-identification (e.g., different lighting conditions and different colour calibration), also differences in the image aspect ratio turned out to be present. They can be caused by the Deformable Part Model pedestrian detection algorithm that was used in Market-1501 to extract bounding boxes with a *fixed* size. Some examples of this issue can be found in figure 3.3. This kind of image may make more difficult for the user to recognise different identities exhibiting a similar

Table 3.6: Results of cross-data set experiments (source ⟶ target) for source model, UDA methods (ECN, MMT), and RF algorithms (QS, RS, M-RS, PA) after each round of **multi-feedback** protocol. In particular, ResNet models were trained on Market-1501 (left) and MSMT17 (right), and tested on the target domain DukeMTMC-reID. Best results for each column are highlighted in bold.

Methods	Round	Market-1501 ⟶ DukeMTMC-reID					MSMT17 ⟶ DukeMTMC-reID				
		mAP	Rank-1	Rank-5	Rank-10	Rank20	mAP	Rank-1	Rank-5	Rank-10	Rank20
Source model	-	29.1	47.7	61.3	66.0	72.0	42.6	60.3	72.3	79.3	83.7
ECN	-	43.2	66.7	77.3	81.0	82.7	42.8	67.0	78.3	81.0	83.7
MMT	-	60.8	76.0	85.3	88.0	90.3	64.3	78.7	86.7	90.0	91.7
QS	1	47.68	71.0	79.0	79.67	81.33	60.72	81.67	87.33	88.67	90.33
	2	50.89	73.67	81.67	83.33	85.0	64.92	85.67	92.0	92.67	94.0
	3	51.74	73.67	82.67	83.67	85.0	66.35	86.67	92.33	93.33	94.33
RS	1	57.72	81.0	83.0	83.67	84.33	69.71	89.33	91.0	91.33	92.0
	2	68.06	87.67	88.0	89.0	90.33	79.93	94.0	94.33	94.67	94.67
	3	74.67	92.0	92.67	92.67	93.0	84.81	95.33	95.33	95.33	96.0
M-RS	1	56.44	79.0	82.67	84.67	87.0	69.22	88.67	90.33	91.67	93.0
	2	69.71	89.67	91.33	91.33	92.33	80.8	94.0	96.67	97.33	98.0
	3	75.46	91.33	92.67	93.33	93.33	86.21	97.33	**98.33**	**98.33**	**98.33**
PA	1	59.1	81.0	84.0	88.0	90.0	71.17	89.0	92.33	92.33	93.33
	2	74.83	91.67	93.0	93.67	93.67	84.83	96.33	97.33	97.33	98.0
	3	**80.8**	**94.0**	**94.0**	**94.0**	**94.33**	**89.75**	**98.0**	98.0	98.0	98.0

appearance. This issue is not likely to be present in a real person re-identification system, where images extracted by a pedestrian detector would be not resized.

Besides the above mentioned aspect ratio issue, some annotation errors were also observed. In particular, in the Market-1501 data set different images of the same individual were found to be annotated with different IDs, whereas some images belonging to different individuals were annotated with the same ID. Some examples of annotation errors are reported in figure 3.4. Moreover, also in DukeMTMC-reID annotation errors due to pedestrian tracking were observed due to static and dynamic occlusions. These annotation errors can influence the performances of HITL methods to a higher extent than UDA methods.

3.3 Discussion

In this chapter the use of CBIR-RF algorithms with a specifically devised feedback protocol was proposed to implement HITL person re-identification system, focusing on cross-view application scenarios, in which representative images of the target domain

Table 3.7: Results of cross-data set experiments (source ⟶ target) for source model, UDA methods (ECN, MMT), and RF algorithms (QS, RS, M-RS, PA) after each round of **multi-feedback** protocol. In particular, ResNet models were trained on DukeMTMC-reID (left) and MSMT17 (right) and tested on the target domain Market-1501. Best results for each column are highlighted in bold.

Methods	Round	DukeMTMC-reID ⟶ Market-1501					MSMT17 ⟶ Market-1501				
		mAP	Rank-1	Rank-5	Rank-10	Rank20	mAP	Rank-1	Rank-5	Rank-10	Rank20
Source model	-	25.5	54.3	72.0	79.0	81.7	31.7	57.3	74.0	80.3	85.7
ECN	-	34.9	64.3	80.3	86.0	91.0	38.3	68.7	82.7	87.0	92.0
MMT	-	69.4	87.0	**95.3**	**97.0**	**97.7**	67.2	84.7	93.3	95.3	96.3
QS	1	43.23	79.67	85.67	86.67	88.0	53.07	86.0	91.0	91.67	93.0
	2	46.9	81.0	87.33	88.0	88.0	57.29	87.67	93.0	93.33	93.67
	3	47.64	80.67	87.33	88.0	88.0	58.32	88.33	93.33	93.67	94.0
RS	1	56.65	88.0	88.33	89.0	91.0	65.41	92.0	93.33	93.33	94.0
	2	68.44	91.33	92.33	92.33	92.67	76.87	95.0	95.33	95.67	96.0
	3	75.09	92.67	92.67	93.33	93.67	84.17	**97.33**	**97.33**	**97.33**	**97.33**
M-RS	1	54.72	86.0	88.0	89.0	90.67	65.94	92.0	93.33	93.67	95.0
	2	69.31	90.67	92.67	93.0	93.33	79.71	95.67	96.0	96.67	97.0
	3	76.89	92.67	94.0	94.0	94.0	86.39	**97.33**	**97.33**	**97.33**	**97.33**
PA	1	57.2	86.0	88.33	90.33	91.67	70.09	93.0	94.33	94.33	94.67
	2	76.85	93.67	93.67	93.67	94.0	85.43	96.0	96.0	96.0	96.33
	3	**82.75**	**94.67**	94.67	94.67	94.67	**89.97**	96.33	96.33	96.33	96.33

Figure 3.3: Examples of considerable appearance changes in Market-1501: colour (first three columns) and shape (last two columns).

are not available during system design. The performance of the proposed approach was evaluated on three benchmark data sets in cross-data set setting, and compared with representative methods of UDA approach, that require unlabelled images of the target domain during design. A direct comparison with existing HITL methods for

Table 3.8: Results of cross-data set experiments (source ⟶ target) for source model, UDA mehtods (ECN, MMT), and RF algorithms (QS, RS, MM, PA) after each round of **multi-feedback** protocol. In particular, ResNet models were trained on Market-1501 (left) and DukeMTMC-reID (right) and tested on target domain MSMT17. Best results for each column are highlighted in bold.

Methods	Round	Market-1501 ⟶ MSMT17					DukeMTMC-reID ⟶ MSMT17				
		mAP	Rank-1	Rank-5	Rank-10	Rank20	mAP	Rank-1	Rank-5	Rank-10	Rank20
Source model	-	1.9	6.3	10.7	14.0	18.7	2.8	10.0	16.0	21.7	28.0
ECN	-	3.8	8.7	19.3	22.7	31.3	5.6	19.0	28.3	35.0	40.3
MMT	-	5.6	14.3	25.0	29.3	33.7	7.4	19.7	31.0	35.0	42.0
QS	1	4.69	20.0	24.0	26.33	27.33	7.39	27.33	33.33	36.33	38.33
	2	5.6	22.0	26.67	29.67	30.0	9.27	32.33	37.67	40.0	40.33
	3	6.02	22.67	27.0	30.0	30.67	10.1	34.33	39.0	40.67	41.0
RS	1	7.44	26.67	29.0	30.0	33.0	11.87	37.67	42.0	46.0	49.33
	2	13.21	37.33	39.0	39.67	41.0	21.38	55.0	57.0	58.33	61.67
	3	18.11	44.0	45.0	45.0	45.33	28.06	64.67	65.67	66.0	66.67
M-RS	1	7.0	25.33	29.67	33.67	35.0	11.93	34.33	42.67	49.33	56.0
	2	13.72	39.0	40.33	41.33	43.67	24.13	61.33	62.67	65.33	67.67
	3	18.16	46.33	47.67	**48.67**	**50.0**	33.1	**69.67**	**70.33**	**70.67**	**72.67**
PA	1	7.97	28.33	32.33	35.0	37.0	12.96	38.33	47.0	51.0	57.0
	2	16.68	44.33	44.33	44.67	46.0	27.74	64.67	65.0	65.33	66.33
	3	**22.21**	**47.67**	**48.33**	48.33	48.33	**37.16**	68.67	68.67	68.67	69.0

Figure 3.4: Examples of images of the same individual labelled with different IDs (first two columns: Market-1501; third and fourth column: DukeMTMC-reID), and of images of different individuals labelled with the same IDs (last two columns, DukeMTMC-reID).

person re-identification was not possible, due to the unavailability of their source code and to their much higher complexity than RF algorithms, that did not allow a re-implementation. First, the considered implementation of HITL approach clearly out-

performed the source model in cross-data set experiments, since the first round, in all target data sets. Moreover, the proposed multi-feedback protocol allowed the considered RF algorithms to achieve a significant performance improvement with respect to the single-feedback protocol used in existing HITL person re-identification methods. Finally, even when implemented through simple CBIR-RF algorithms, in most cases the HITL approach achieved comparable or better performances than UDA methods, despite the latter exploited a much higher amount of target data (although unlabelled) at the design phase. This shows that, CBIR-RF algorithms can attain an effective trade-off between re-identification performances and processing cost in challenging cross-view scenarios, without requiring images of the target view during system design, as the UDA approaches, but leveraging instead the feedback that the operator can give online, during system operation, on a much smaller number of target images.

Some issues of the proposed HITL approach are now discussed. One limitation might be represented by the size of the gallery, which in application scenarios characterised by hours of video footage acquired by many cameras can be very large. In this case, the algorithm should process all images of the ranked gallery to retrieve other relevant images taking into account the feedback provided on the top-k ranks. This process can require more time. In this case, HITL approach to speed up the process could examine only a subset of the ranked gallery assuming that images of people from a certain position in the ranked list are too different or dissimilar from the query.

Another issue concerns the human effort, i.e., the number of feedback. In the proposed implementation of the HITL approach, the user should select all true matches in the top-k ranked gallery images, contrary to existing HITL methods that require a single feedback. Selecting all true matches may appear too demanding for the end user. On the other hand, as the size of the gallery increases, it becomes less likely to find several true matches in the top-k ranks (I point out that in real application scenarios the gallery can be order of magnitude larger than in benchmark data sets). In particular, in my experiments 18 true matches on average were present among the top-50 ranks in the first feedback round; moreover, in the next rounds such images are still present in the top-k ranks and therefore the user does not have to select them again. This means that the human effort after each round would be limited also under the multi-feedback protocol. Another issue, related to the considered RF algorithms to implement the HITL approach, is that with respect to state-of-the-art methods

[134], it does not modify or update the feature representation or the similarity measure incrementally, over different queries. On the one hand, this may limit the capability of the proposed implementation to adapt the source model to the target domain. On the other hand, this makes the considered HITL implementation much faster in re-ranking the gallery images, which is useful for very large galleries.

Thanks to the participation of LETSCROWD project, I had the opportunity to test a prototype of the person re-identification system developed in this thesis during practical demonstrations on a real and challenging application scenario. This allowed collecting feedback of potential end users (LEAs officers) on different aspects of the prototype, beside its re-identification performance. The practical demonstrations were fundamental to assess performances of the proposed re-identification system on real data, and in particular the gap with the performance on data from benchmark data sets, as well as its usability by the end users and their general expectations about this kind of supporting tool.

As expected, other issues emerged beside the ones mentioned above. One is related to the presence of "false positives" in the ranked gallery obtained for a given query image, i.e., images of individuals different from the query individual, and sometimes exhibiting a significantly different clothing appearance. This is unavoidable in the person re-identification task, especially when the size of the template gallery is very large (which is typical of real applications). It is worth reminding that the images (in term of bounding boxes) contained in the template gallery are automatically extracted by a pedestrian detector or tracker. These bounding boxes can be not precise, i.e., the bounding box can present missing parts of a person, or can be not tight enough. As a consequence, the background can cover a considerable portion of the bounding box, affecting the feature extraction significantly, which in turn negatively influences the matching phase. As a consequence, the ranked gallery can present several "false positives" even in the top ranks. However, it turned out that the end users involved in practical demonstrations had relatively demanding expectations about person re-identification tools: they did not expect false positive at the top of the ranked gallery, not even images of people with similar appearance as the query, but only images of the *same* person of interest acquired by the available cameras. Therefore false positives were interpreted simply as errors made by the system: from their viewpoint, false positives indicated a low accuracy of the system. This highlights that for an effective and useful

deployment of computer vision supporting tools, even if semi-automatic (e.g., with a human in the loop), a suitable training of the end users is necessary, to allow them understand the operating principles and characteristics such as the expected accuracy in different operating conditions. This would enable end users to understand the actual potentiality and limitations of a computer vision tool, and to correctly interpret its outputs.

Another issue is related to regulatory constraints, and in particular privacy-related ones, due to , e.g., General Data Protection Regulation (GDPR)[1] or to country-specific regulations. It is worth knowing that, in some countries like Germany, person re-identification systems cannot be used in real-time, i.e., during an event. More precisely, a person re-identification system can be used only offline, during an official investigation, authorised by the judicial authority, and therefore in a post-event phase. In this kind of application constraints on processing time are likely to be less severe, and it may be possible to trade a higher accuracy (e.g., through more complex and more discriminative features) for a higher processing time.

[1] https://gdpr.eu/

Chapter 4

Scene-specific crowd counting with synthetic training images

Crowd counting and density estimation are useful functionalities useful in several security-related applications involving monitoring and analysis of crowds through video surveillance systems. In the computer vision research community, this is still a challenging task for unconstrained scenes, due to issues like perspective distortions, illumination changes, occlusions, scale variations, complex backgrounds etc. An additional issue is related to the need of labelled data by state-of-the-art supervised techniques, which might be very difficult or too demanding to collect, especially in real, cross-scene application scenarios where the target scene where a crowd counting system will be used during operation are unknown during system design. For instance, this is the case of a new camera installation by LEAs which should be operational in a short time. In such a case it is unfeasible to require to the end users (e.g., LEA operators) to collect and annotate a sufficient amount of representative crowd images of the target scene. To evaluate the cross-scene performance of crowd counting methods, cross-data experiments are reported in literature using source and target benchmark data sets that are different in terms of background, illumination, perspective, scale, size of crowd, etc. Reported results clearly show a significant performance drop by even state-of-the-art supervised techniques with respect to the same-scene scenario, when training (source) and testing (target) data come from the same data set [18, 154]. Accordingly, recent research effort is being devoted to developing solutions robust to cross-scene issues, including the use of DA and UDA.

This thesis focuses on a possible solution to address challenging cross-scene application scenarios where it is not possible to obtain labelled nor unlabelled images of the target scene during system design for fine-tuning a crowd counting model. The first contribution of this work is an extensive assessment of the performance gap between same- and cross-scene scenarios both in traditional (i.e., regression-based) and in recent (i.e., CNN-based) crowd counting approaches [18]. The second contribution of this work consists of an alternative technique to reduce this gap, with respect to the ones proposed so far in the literature. Inspired by recent work in other computer vision tasks [12, 14, 44, 58, 79, 109, 111, 122], this thesis propose the use of *synthetic* images was proposed to mitigate the lack of representative image to build a scene-specific training set for a specific target camera view (target scene), by requiring minimal effort to end users. This approach has some potential advantages: i) synthetic images can be *automatically* annotated, and therefore do not require any annotation effort to human users; ii) for the same reason synthetic images can be free from annotation errors; iii) since no collection and annotation effort is required, it is possible to create a large amount of representative (synthetic) images of the target scene, capable to fulfil the needs of demanding supervised methods such as the ones based on CNNs.

In particular, the approach proposed in this work requires from the end user a *single* background image of the target scene, the ROI in terms of a binary map, and the perspective map. This information can be provided with low effort through an appropriate user interface, e.g., embedded in software suites currently used to manage video surveillance camera networks.

In the next sections, motivations of the proposed approach are presented 4.1, followed by its description 4.2, and by its extensive experimental evaluation and comparison on several state-of-the-art methods 4.3. Finally, the discussion is reported in sect. 4.5.

4.1 Motivations

This thesis focuses on real-world, cross-scene scenarios in which it is infeasible for the end users to collect and possibly annotate a sufficient amount of representative images of the target scene, for instance, when a new camera installation is operated by LEAs, which should be operational in a short time. Under this scenario it is not possible to use

crowd counting approaches that require either annotated or non-annotated images of the target scene to address the cross-scene scenario, such as DA and DA. In particular, manual image annotation requires a huge human effort by state-of-the-art supervised CNN-based approaches, since their ground truth consists not of the number of people in an image, but of the head position of each pedestrian. To overcome this issue, inspired by previous work in other computer vision tasks [12, 14, 44, 58, 79, 109, 111, 122], the use of synthetic training images was proposed in this thesis.

The main advantages of such a solution, mentioned above, are particularly appealing for CNN-based methods, as synthetic images allow automatic generation and annotation of a large and representative training set of target scene, e.g., containing different crowd sizes and configurations of pedestrians. This is beneficial for the generalisation capability of any supervised crowd counting model, and especially for CNN-based ones, which require a significant amount of training data. In particular, in this challenging computer vision task it is useful that training images have the same background and perspective of the target scene. The former is important since features used by crowd counting models are typically affected by the background. The latter is useful to limit performance degradation in a cross-scene scenario [18]. Clearly, these two requirements might be not fulfilled in cross-scene scenarios in which the target scene is not known in advance. Despite its potential advantages, to my knowledge this kind of approach (i.e. synthetic training images) has not been used for regression- and CNN-based crowd counting, the only exception is [135], which however still requires images of the target scene, although not annotated, to fine-tune the source model *during design*: this is not feasible under the application scenario considered in this thesis.

4.2 Approach

Starting from a background image of the target scene, its ROI and perspective map, the idea of the approach proposed in this thesis is to build a training set of synthetic images obtained by superimposing to the background image pedestrian images placed in random locations of the ROI, properly scaled through the perspective map. The generated synthetic images should present different crowd sizes and different configurations of people to favour generalisation capability, and their number should be sufficiently large to fulfil the requirements of the underlying supervised crowd counting model.

The three elements mentioned above can be provided by the user with a low effort: a background of the target scene, its ROI and its perspective map [20]. A *background* image of the target scene should be provided, possibly without pedestrians or other non-fixed objects (for instance, cars etc.). Such an image can be easily collected during camera set-up. If the acquired image contains pedestrians or non-fixed objects, it is possible to remove them by using image subtraction algorithms on a short video (a few frames can be sufficient). The ROI of the background image can be provided in terms of *binary map* which defines the image region where the pedestrians can appear. Static objects (if any) should be manually removed from the image. The ROI can be easily drawn by the human operator through an appropriate user interface, e.g., as a polygon area. The *perspective* map is an image in which the value of each pixel corresponds to the height, in pixels, of a standard adult at the corresponding location. This allows to properly scale synthetic pedestrian images at each image location. The perspective map can be automatically computed by asking the user to manually select the bounding boxes of a few pedestrians in images of the target scene acquired during camera set-up. In particular, for a flat target scene, three bounding boxes are sufficient [85]. An *additional* information that can be defined by the user is the expected range of crowd size during system operation (e.g., the maximum crowd size which is expected by LEAs during a public demonstration, in the venue region corresponding to the camera view). It allows to precisely limit the range of the number of pedestrians to be generated in synthetic images.

Once the above information has been acquired by the user, it is possible to proceed with the generation of synthetic data set of the target scene.

To this aim, a collection of pedestrian images is necessary. These images can be collected from the web or can be generated by computer graphics tools, by system designers. In particular, these images should not include background (i.e., they should present a transparency layer), and should be set to a standard height, to simplify their re-scaling according to the perspective map. The proposed approach consists of superimposing on the background image of the target scene synthetic images of pedestrians, positioned in random locations on the ROI and scaled according to the perspective map. To reproduce a realistic overlapping among people, pedestrian images can be added one at time from the farthest to the nearest location to the camera. By

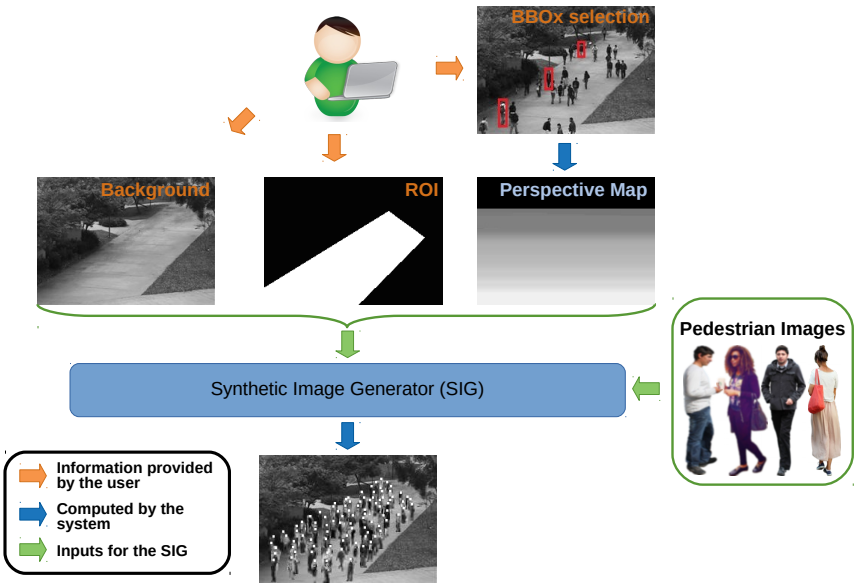


Figure 4.1: Scheme of the proposed procedure for generating synthetic training images of the target scene. The orange arrows indicate the information required to the user, i.e. an image of the background, the region of interest (ROI) and the selection of three bounding boxes in the target scene. The ROI defines the area in which pedestrian can appear. The selection of three bounding boxes is necessary to automatically compute the perspective map that in turn is used to compute the correct size of synthetic pedestrian images. The blue arrows mean that the corresponding output images are automatically computed. In particular, the perspective map is computed by using the selected bounding boxes; and the final synthetic image is generated through the Synthetic Image Generator (SIG) by using the input information highlighted in green. The white dots in the output (synthetic) image represent the automatically annotated head positions of pedestrians, which are required by CNN-based methods.

positioning pedestrians in random locations, it is possible to build images of the target scene with different configurations of people.

To generate images with different crowd sizes the following procedure can be followed, starting from the expected range of crowd size provided by the end user. The number N of synthetic images to be generated can be set by system designers, depending on the underlying crowd counting model. The number of pedestrians n in each synthetic image can be defined based on the maximum number of pedestrians $n_{\max}$ provided by the user: if $n_{\max} = qN$, where $q \in \mathbb{R}^+$, each synthetic image will contain n pedestrians, with $n = 1, \lceil 1+q, \lceil 1+2q, \cdots, n_{\max}$. Each generated image can be *auto-*

matically annotated in terms of either the number of pedestrians, or the head locations of each pedestrian and the corresponding density map, as required by the underlying crowd counting model. To automatically annotate the head location of each synthetic pedestrian basic human anatomy notions can be exploited, according to which the head is 1/8 of the total body [88] and the head point is located at 1/16 height and 1/2 width of the image. The above procedure is shown in Fig. 4.1. All the synthetic images generated by using the proposed approach are used as training data to learn the considered crowd counting model. It is worth noting that the proposed approach is computational much simpler than the ones of [135], which is based on graphical engine and on GANs.

4.3 Experiments

The experiments reported in this section are aimed first at assessing the performance gap of state-of-the-art crowd counting methods in cross-scene settings with respect to the same-scene setting, and then at evaluating the effectiveness of the proposed approach in mitigating such a gap, for the same crowd counting methods. Among state-of-the-art methods, four early regression-based methods and nine CNN-based ones were selected (sect. 4.3.1). They are applied to six benchmark data sets (more details in sect. 4.3.2). In particular, five of the selected data sets are single-scene, whereas the other one is a multi-scene data set. The term "multi-scene" refers to data sets that contain images of many different scenes, with one or few images of each scene. To simulate a cross-scene setting, each single-scene data set (one by one) is used as the target scene, and a different one is used as the source (training) data. Moreover, the multi-scene data set is also used as the training set, which is a possible proposed in literature solution for improving cross-scene performances. Experiments under same-scene setting (i.e., training and testing on the same data set) have then been carried out to evaluate the performance gap.

In subsequent experiments, for each target scene (data set), a synthetic training set is built by using a background image of the same target scene (see sect. 4.2). Performances obtained by using a synthetic data set as training set have been evaluated for all the crowd counting methods considered before. In the next sections the experimental set-up and the obtained results are described in details.

4.3.1 Crowd counting methods

Regression-based crowd counting methods require the selection of a set of low-level features and a regression model (sect. 2.2.1). Among all features used in early regression-based approaches, two common foreground features as segment [86] and edges [60], and Grey-Level Co-occurrence Matrix (GLCM) [43] and Local Binary Pattern (LBP) texture [92] features are considered in the following experiments. Segment and edge features focus on extracting complementary information, i.e., global and local information, respectively. In particular, global features extract properties of the image as area and perimeter; instead, edge features extract information as the number and the orientation of edge pixels. GLCM aims to extract information about the spatial relationship of pixels, whereas LBP identifies texture patterns in an image. In the following experiments the above features are concatenated into a single feature vector. It is worth pointing out that the above features are seriously affected by the background.

Concerning the **regression models**, global ones are considered in these experiments. In particular, Linear Regression (LR), Partial Least Square (PLS) regression, Random Forest (RF) and Support Vector Regression (SVR) with Radial Basis Function (RBF) kernel have been considered. The first two are linear models, whereas the others are non-linear. LR is the most simple approach that combines the input variables linearly. As described in sect. 2.2.1, the drawback of this approach is related to its computational complexity that grows excessively with high-dimensional data. In contrast, PLSR does not present this disadvantage; in particular, a decomposition of input and target variable is used to maximise the covariance between score matrices. GPR model is an approach more flexible than the previous ones. However it is not scalable to large data sets and is more sensitive to parameter values than the other methods. One of these issues is overcome by RF, that is scalable to large data sets.

Regarding the **CNN-based** models, nine state-of-art methods, whose source code was made available by the authors have been considered. The Multi-Column CNN (MCNN) architecture can be defined as multi-branch or multi-column; indeed it consists of three columns that share the same configuration except for the size of filters (large, medium and small, respectively) [157]. All the information extracted by these branches are merged by a final module to obtain the density map. Also the Cascaded Multi-Task Learning (CMTL) architecture is a multi-column. In particular, CMTL is composed

of two columns that work on two related sub-tasks which share the first layers [116]. The first sub-task focuses on the crowd count categorisation into ten crowd levels, whereas the second one defines the density map. The Deformation Aggregation Network (DAN) consists of two parts [171]. The first one, which is made up of the first eight layers of VGG, is used to obtain feature maps, whereas the other part consisting of deform blocks aims to preserve the correlation between the image and the density map. Finally, an adaptive fusion module is used to combine the feature map and the output of deform blocks by using different weights. The Spatial Fully Connected Network (SFCN) uses ResNet as a backbone and an encoder, placed in the top of ResNet, is used to improve the density map estimation [135]. The Congested Scene Recognition Network (CSRN) consists of VGG backbone and dilated modules that aim to extract discriminant information without increasing the required resources [142]. The Context-Aware Network (CAN) uses VGG as a backbone and combines feature maps extracted from the backbone with weighted feature maps to obtain the density map [83]. The Spatial-/ Channel-wise Attention Regression (SCAR) network is based on two modules that can be defined as two variant of self-attention modules [35]. The first one, named Spatial-wise Attention Module (SAM), aims at improving the accuracy of head detection. The other, called Channel-wise Attention Module (CAM), defines the relationship between channel-maps mitigating errors. The Deep Structured Scale Integration Network (DSSI-Net, for short DSSI) consists of three columns that share the same parameters [81]. Each column considers a different scale of the input image to address scale variation issues. Then, all outputs are fused to obtain the density map. The Bayesian Loss (BL+) approach uses a loss function to alleviate the large scale variation [87]. Table 4.1 summarises the main features of CNN-based methods including other information as augmentation techniques, loss function, type of kernel used to compute the density map, and details about the inference time.

4.3.2 Real data sets

The scene-specific setting considered in this thesis implies that crowd counting models should be evaluated on a testing set made up of video frames coming from a *same* scene (different from the ones used for training); moreover, dense crowds are more interest than sparse crowds for the considered application scenario (e.g., monitoring of mass gathering events by LEAs). However, current benchmark data sets do not satisfy these

Table 4.1: Main features of the CNN-based methods used in these experiments. Network architecture: pre-trained backbone network (– denotes a network trained from scratch), number of columns, loss function (Mean Squared Error, MSE; Binary Cross Entropy, BCE; Bayesian loss). Input: type of input images (whole image or crop, and augmentation technique – flip, noisy, scale), and kernel used for computing the density map. Speed: inference time (in ms) on a reference input image of size 640×480

Method	Network architecture			Input		Speed
	backbone	columns	loss	images	kernel	
MCNN [157]	–	3	MSE	Crop	Fixed	130
CMTL [116]	–	2	MSE&BCE	Crop&Flip&Noisy	Fixed	350
DAN [171]	VGG16	5	MSE	Crop	Fixed	210
SFCN [135]	ResNet	–	MSE	Whole	Fixed	900
CSRN [71]	VGG16	–	MSE	Crop&Flip	Fixed	480
CAN [83]	VGG16	4	MSE	Crop&Flip	Fixed	450
SCAR [35]	VGG16	2	MSE	Whole	Fixed	412
DSSI [81]	VGG16	3	MSE	3 scales	Adaptive	510
BL+ [87]	VGG19	–	Bayesian	Crop&Flip	Adaptive	260

requirements. Indeed, data sets of dense crowds mainly contain distinct images coming from different scenes; in the case when several video frames coming from a same scenes are available, their number is too limited for the purpose of these experiments. For these reasons, only three benchmark data sets containing small crowds are considered, namely Mall, UCSD and PETS2009, since they are made up of a sufficient number of video frames from a same scene. Although other data sets such as ShanghaiTech, UCF-QNRF and World Expo Shanghai 2010 contain dense crowd images, they do not contain a significant number of images of a same scene. Although the selected data sets (Mall, UCSD and PETS2009) do not contain dense crowd scenes, they present challenging scenarios such as lighting variations, perspective distortion and occlusions.

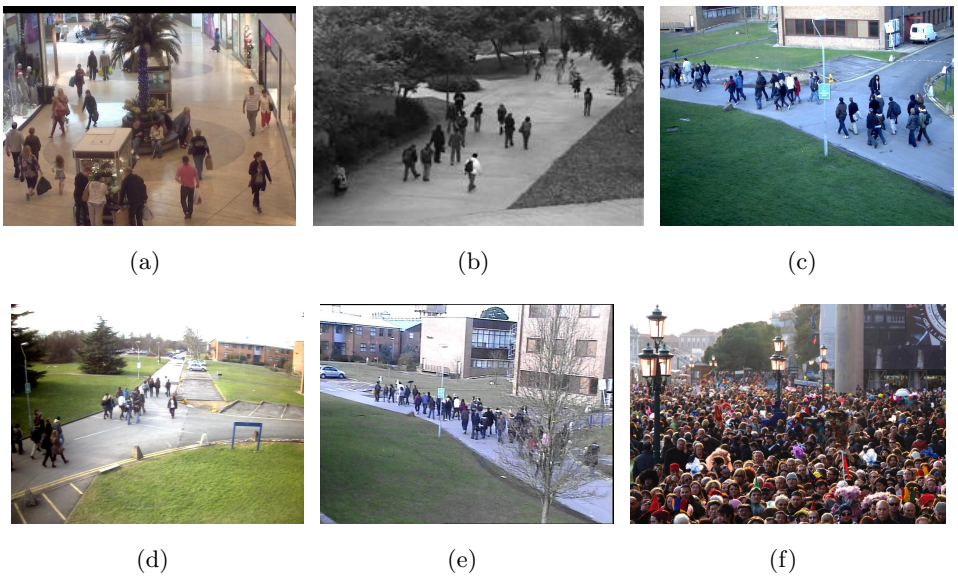
Mall consists of 2000 frames from a single scene acquired by using a camera placed in a shopping mall [11]. The resolution of images is fixed at 320×240 pixels. This data set contains 62,325 pedestrians, with 13 to 53 people per frame. The first 800 frames are used as the training set, and the remaining ones for testing.

UCSD consists of 49,885 pedestrians acquired by a low-resolution camera in a University campus with a frame size of 238×158 [9]. The training set contains 800 frames, and the testing set consists of the remaining 1,200 frames.

PETS2009 is composed of different challenging tasks, such as pedestrian count and

density estimation, people tracking and event recognition [32]. The first part, named S1, is related to crowd counting and is divided into three different parts with different challenging levels. This data set does not contain a single scene, and the available frames are taken from different camera views. In this case, frames from the same camera view are grouped together to obtain a single scene data set for the purpose of these experiments. Then, three single-scene data sets, named PETSview1, PETSview2 and PETSview3, have been created. These new datasets contain 1,229 total frames, that I split into training, validation and test set of size 361, 128 and 740 respectively. Since the original PETS2009 data set does not contain the head locations for each frame, I used the ground truth provided in [156].

ShanghaiTech is instead a multi-scene data set made up of distinct images acquired from different scenes. It contains 1,198 frames and 330,165 pedestrians in total acquired with different cameras, different resolutions, and different crowd size [157]. The data set is divided into two parts, named Part_A and Part_B, that contain 482 and 716 images, respectively. Both partitions are further divided into a training and a testing set made up of 300/182 images for Part_A and 400/316 images for Part_B. Fig. 4.2 shows examples of frames from the above data sets, whereas table 4.2 summarises their main features.



(a) (b) (c)

(d) (e) (f)

Figure 4.2: Example of frames from the data sets used in the experiments: (a) Mall, (b) UCSD, (c) PETSview1, (d) PETSview2, (e) PETSview3, (f) ShanghaiTech.

Table 4.2: Statistics of real and synthetic data sets used in the experiments.

Type	Data set	Image size	Number of images				Pedestrian count			
			total	training	validation	test	total	min	avg	max
Real	Mall	480 × 640	2,000	600	200	1,200	62,235	13	31	53
	UCSD	158 × 238	2,000	600	200	1,200	49,885	11	25	46
	PETSview1	576 × 768	1,229	361	128	740	32,719	1	27	40
	PETSview2	576 × 768	1,229	361	128	740	36,458	2	30	40
	PETSview3	576 × 768	1,229	361	128	740	41,873	11	34	40
Synthetic	Mall	480 × 640	1,000	800	200	–	50,500	1	50	100
	UCSD	158 × 238	1,000	800	200	–	50,500	1	50	100
	PETSview1	576 × 768	1,000	800	200	–	50,500	1	50	100
	PETSview2	576 × 768	1,000	800	200	–	50,500	1	50	100
	PETSview3	576 × 768	1,000	800	200	–	50,500	1	50	100

4.3.3 Synthetic data set

In a real application scenario, once the information (see sect. 4.2) has been collected, it is possible to proceed with the generation of the synthetic data set. The first requirement is the background. To this aim, in these experiments for each data set (sect. 4.3.2) a few images of the target scene are used to obtain the background by using the image subtraction technique. Then, it is necessary to obtain the ROI in terms of a binary map (second requirement). It is defined as a polygon drawn on the background image by the user. The third requirement is the perspective map that is computed by selecting three bounding boxes at different locations by the user on one or more images of the target scene during camera set-up. Finally, the number of synthetic images to be generated had to be defined. Considering the size of original data sets, and to guarantee an equal number of images for each $n_{\max}$ value, a value of $N = 1000$ (the total number of images of the synthetic data set) was selected. The maximum number of pedestrian $n_{\max} = 100$ was set while taking into account the characteristics of the target scene and the size of the ROI. For each value in range $[1, n_{\max}]$, $N/n_{\max} = 10$ synthetic images were generated obtaining 50500 pedestrian in total. The resulting synthetic data set was divided into a training and a validation set. In particular, 800 images were used for training, whereas the remaining images were used for validation. In table 4.2 details about the generated synthetic data set are reported.

4.3.4 Metrics

Crowd counting accuracy is evaluated by using two common metrics: mean absolute error (MAE) and root mean squared error (RMSE)

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\eta_i - \hat{\eta}_i| \ , \tag{4.1}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\eta_i - \hat{\eta}_i)^2} \ , \tag{4.2}$$

where η_i and $\hat{\eta}_i$ are the exact count (ground truth) and the estimated count for the i-th image respectively, and N is the total number of images. The MAE evaluates the mean of absolute error between the actual count and the estimated one. The RMSE evaluates the root squared error between the exact count and the estimated one, and therefore penalises larger errors more than smaller ones.

4.3.5 Results

The experiments aimed at evaluating i) the performance gap between cross- and same-scene setting of early and state-of-the-art crowd counting methods and ii) whether the use of a synthetic training set of the target scene is an effective approach to reduce such gap when representative images of the target scene are not available. In particular, cross-scene experiments were carried out by using single- and multi-scene training sets as described in section 4.3. As mentioned before, the cross-scene setting was simulated by training a model on a single-scene or multi-scene data set, and testing it on a *different* (target) data set. The same-scene performance was evaluated by training and testing on the *same* data set. This results reported in the following extend the ones published in [18, 19]. Finally, cross-scene performances were compared with results obtained by using a synthetic training set of the corresponding target scene.

Tables 4.3 and 4.4 report performances of early regression-based and CNN-based approaches by using single-scene data sets, respectively. Table 4.5 show results obtained by using a multi-scene training set. For an easy comparison, the best and the worst cross-scene performances are reported in table 4.4. Table 4.6 reports performances obtained by using synthetic data sets. Also in this table, for an easy comparison, the best same-scene and the best cross-scene performances are reported.

Table 4.3: Cross-scene MAE and RMSE of early regression-based methods (LR, RF, SVR and PLS) using single-scene training sets. Same-scene results (when training and testing images belong to the same data set) are also reported for comparison, highlighted in grey. The best cross-scene result for each target data set is reported in bold.

	Training set	Testing set (target scene) Mall MAE	Mall RMSE	UCSD MAE	UCSD RMSE	PETSview1 MAE	PETSview1 RMSE	PETSview2 MAE	PETSview2 RMSE	PETSview3 MAE	PETSview3 RMSE
LR	Mall	2.74	3.49	9.59	11.63	289.2	294.4	348.7	349.0	268.1	270.9
	UCSD	67.3	78.75	2.9	3.54	334.6	347.9	369.2	374.0	128.2	146.6
	PETSview1	276.9	277.0	577.1	577.2	6.25	7.91	33.43	38.04	9.35	11.17
	PETSview2	210.2	210.3	308.4	308.4	97.86	127.0	4.85	5.98	159.4	160.2
	PETSview3	12.15	14.01	29.09	29.93	110.3	110.7	125.1	126.6	6.84	8.42
RF	Mall	3.82	4.85	5.12	7.42	9.27	12.43	12.15	13.96	**4.44**	**6.59**
	UCSD	5.83	6.98	3.82	4.66	9.12	11.45	**8.06**	**10.46**	5.22	5.94
	PETSview1	3.89	5.07	6.92	8.12	9.47	11.03	13.59	14.98	8.36	9.31
	PETSview2	6.88	8.57	5.38	7.31	**8.01**	**8.94**	9.56	11.05	6.27	8.14
	PETSview3	5.52	7.07	6.34	7.73	10.11	11.54	11.59	12.54	11.41	12.49
SVR	Mall	4.8	6.29	8.15	9.18	9.56	10.45	9.8	10.68	8.74	9.55
	UCSD	7.68	9.32	5.38	7.31	10.74	12.08	12.09	13.15	12.86	13.88
	PETSview1	12.26	13.57	6.21	8.52	12.82	15.25	14.85	16.79	17.67	18.56
	PETSview2	8.54	10.12	5.13	7.3	11.06	12.62	12.6	13.81	13.78	14.8
	PETSview3	**5.11**	**6.71**	7.52	8.61	9.76	10.61	10.2	11.04	9.5	10.37
PLS	Mall	3.16	4.1	110.7	110.9	51.97	65.77	16.97	20.94	53.4	61.05
	UCSD	266.3	268.0	2.6	3.23	99.38	109.1	428.7	429.9	460.9	467.7
	PETSview1	49.0	49.37	13.0	14.21	8.46	10.13	20.39	24.53	21.07	26.56
	PETSview2	23.01	23.42	103.9	104.1	57.72	68.15	7.65	9.06	103.1	103.8
	PETSview3	18.05	18.67	**5.1**	**7.27**	14.55	16.86	25.12	26.75	9.03	10.06

Regression-based approaches achieved a remarkable same-scene performance (table 4.3) on Mall and UCSD data sets. The best same-scene performances are reported by LR and PLR, whereas they exhibited worse performance than other methods in the cross-scene setting. In particular, these methods achieved an MAE lower than 3.2 in the same-scene setting, but the lower cross-scene MAE is 49. In contrast, RF and SVR exhibited a limited performance degradation in most cross-scene experiments. For instance, the RF model trained on Mall and tested on UCSD achieved an MAE lower than 5.5, whereas the same-scene (Mall-Mall) MAE is 3.82. In some cases, the best cross-scene performances (highlighted in bold) of RF and SVR are comparable, and sometimes they are even better than those obtained in the same-scene setting. This last behaviour was exhibited by RF when the training-testing sets pairs were PETSview2-PETSview1, UCSD-PETSview2 and Mall-PETSview3 whose scenes are similar in terms of perspective and scale, especially in the case of Mall and of the three views of PETS2009.

Table 4.4: Cross-scene MAE and RMSE of CNN-based methods using single-scene training sets. Same-scene results are also reported for comparison, highlighted in grey. The best cross-scene result for each target data set is reported in bold.

	Training set	Testing set (target scene) Mall MAE	Mall RMSE	UCSD MAE	UCSD RMSE	PETSview1 MAE	PETSview1 RMSE	PETSview2 MAE	PETSview2 RMSE	PETSview3 MAE	PETSview3 RMSE
MCNN	Mall	5.33	6.17	24.64	25.75	5.94	7.83	9.67	10.95	9.9	11.22
	UCSD	86.39	88.04	2.3	2.84	144.9	149.6	49.4	56.85	180.6	181.2
	PETSview1	19.54	20.16	24.18	25.28	6.2	7.86	22.05	23.59	9.77	11.75
	PETSview2	3.39	4.27	19.62	20.92	20.93	22.19	4.23	5.08	24.29	27.72
	PETSview3	4.31	5.35	21.28	22.47	19.54	21.63	10.37	11.66	4.18	5.13
CMTL	Mall	5.53	6.39	23.42	24.58	**5.77**	**7.42**	17.65	19.28	11.41	12.79
	UCSD	189.1	191.1	2.04	2.50	213.7	217.9	111.9	113.7	298.5	300.8
	PETSview1	9.93	10.73	24.18	25.13	5.11	6.29	15.56	17.20	4.46	5.95
	PETSview2	4.68	5.95	24.63	25.76	36.85	38.49	4.80	6.06	47.34	50.96
	PETSview3	4.61	5.79	21.94	23.12	21.90	24.54	11.50	13.97	4.23	5.06
DAN	Mall	5.43	6.42	25.42	26.54	7.51	9.43	11.7	13.14	8.84	10.27
	UCSD	164.1	166.1	5.18	6.39	185.9	192.1	61.76	66.53	227.3	228.5
	PETSview1	7.97	9.06	26.1	27.09	4.92	6.15	16.41	19.12	6.34	7.74
	PETSview2	28.95	29.54	27.86	29.0	26.43	28.38	28.68	30.37	32.89	33.38
	PETSview3	7.9	9.48	**18.8**	**20.12**	18.02	20.45	13.2	15.15	4.63	5.92
SFCN	Mall	4.05	5.02	28.15	29.27	19.37	20.85	27.66	28.72	71.38	71.87
	UCSD	880.2	882.1	2.91	3.64	853.5	859.6	634.3	635.5	988.4	990.6
	PETSview1	8.33	9.64	27.13	28.1	6.32	7.57	12.83	14.5	10.74	12.05
	PETSview2	36.55	38.35	25.93	26.85	85.29	87.81	8.1	9.81	106.9	108.6
	PETSview3	14.78	15.98	28.23	29.36	11.49	13.64	10.03	12.74	4.35	5.68
CSRN	Mall	6.57	7.73	24.51	25.8	21.55	23.89	19.08	21.61	15.37	16.38
	UCSD	70.78	71.46	6.2	7.01	57.52	61.86	28.29	31.21	69.06	69.36
	PETSview1	14.51	14.96	27.33	28.43	5.54	6.83	15.62	17.46	20.57	21.11
	PETSview2	12.15	12.66	27.06	28.16	10.14	11.82	7.09	7.9	8.42	9.53
	PETSview3	9.21	9.89	27.49	28.62	5.84	6.8	9.66	10.56	2.9	3.76
CAN	Mall	2.59	3.21	28.09	29.23	8.28	10.36	17.49	20.02	29.54	30.11
	UCSD	281.6	283.1	4.73	6.16	173.5	176.9	133.4	135.2	252.0	252.4
	PETSview1	10.5	11.17	27.5	28.56	6.33	7.5	8.43	9.25	**3.94**	**4.84**
	PETSview2	27.59	28.51	27.1	28.15	24.62	26.03	6.07	7.67	5.09	6.77
	PETSview3	6.73	7.7	27.55	28.7	7.5	9.07	11.54	12.78	6.82	7.84
SCAR	Mall	3.99	4.75	372.28	372.8	42.3	45.41	55.78	56.46	93.3	93.51
	UCSD	19.43	20.98	4.19	5.24	19.45	21.11	**6.67**	**8.19**	15.3	17.83
	PETSview1	265.37	265.53	503.0	504.1	3.38	4.07	122.04	128.9	134.72	135.23
	PETSview2	314.63	315.81	574.18	577.12	13.47	17.53	5.09	6.32	123.88	124.16
	PETSview3	36.1	37.14	575.91	578.83	11.88	13.53	38.03	44.03	8.39	10.32
DSSI	Mall	5.44	7.09	37.35	37.81	22.81	23.56	18.1	19.03	13.78	14.98
	UCSD	25.6	26.84	21.75	23.2	27.36	28.53	26.92	28.11	26.52	27.72
	PETSview1	9.87	14.1	69.02	69.8	18.0	20.44	12.63	15.0	10.31	11.36
	PETSview2	8.02	12.5	66.81	67.57	20.25	22.26	14.64	16.51	11.31	12.21
	PETSview3	**4.14**	**6.47**	62.54	62.8	24.09	24.75	17.32	18.22	11.46	12.46
BL+	Mall	2.18	2.74	152.76	153.63	6.9	7.86	15.12	16.08	8.22	9.98
	UCSD	23.96	25.05	2.5	3.57	22.65	23.8	21.17	22.0	23.66	24.77
	PETSview1	10.09	11.81	127.26	129.71	3.75	5.12	12.41	14.34	10.49	12.86
	PETSview2	15.73	17.91	77.63	80.9	15.35	17.78	5.8	6.57	10.22	11.68
	PETSview3	26.01	26.69	132.99	133.57	18.69	19.53	7.44	9.0	4.72	5.61

Also the **CNN-based methods** achieved high performances in same-scene settings, with some exceptions such as DAN in PETSview2, and DSSI in UCSD and in all views of

PETS. In cross-scene settings, also these approaches exhibited a performances decrease. In particular, the gap between same- and cross-scene performance is clear when UCSD is used either as training or as testing set (target scene). This can be explained by the fact that this data set exhibits a significant difference in terms of perspective and scale from the other data sets. The best cross-scene performances (highlighted in bold in table 4.4) in some cases are comparable to the ones obtained in same-scene settings. In some other cases, cross-scene performances are even better than those obtained in same-scene setting. For instance, this result was observed when CAN was tested on PETSview3, and when DSSI was tested on Mall.

By comparing tables 4.3 and 4.4 related to **regression- and CNN-based approaches**, respectively, it can be seen that the performances of the latter are better in the same-scene setting. In contrast, regression-based methods (in particular RF and SVR) achieved better cross-scene performances than the CNN-based ones. Therefore in cross-scene setting, regression-based approaches turned out to be as more robust than CNN-based ones.

As mentioned before, a multi-scene data set (ShanghaiTech) was also considered in these experiments as the training set, since this is the one of the approaches employed in literature to improve cross-scene performances of CNN models [71, 81, 83, 87]. No experiments using a multi-scene training set have been carried out for early regression-based approaches, since a regression model needs different background images of each scene to compute edges and foreground features, whereas a single image for each scene was available in ShanghaiTech. To speed up these experiments, models already trained on ShanghaiTech and made available by the authors were used. In table 4.5 the corresponding cross-scene results of CNN-based methods are reported. The highlighted results show in which cases using a multi-scene training set provided a better performance than the best cross-scene one achieved using a single-scene training set, on the same target data. In particular, the performances achieved by CAN, DSSI and BL+ using a multi-scene training set were comparable to the best cross-scene achieved on a single-scene training set. In some cases, models achieved significantly different results when were trained on Part_A or Part_B: this is the case of such as MCNN and CMTL on PETSview3, and of SFCN on UCSD, Mall and PETSview3. In contrast, performances of the SCAR model were poor in all target data sets. The reason of this behaviour is not clear, and are probably due to overfitting.

Table 4.5: Cross-scene MAE and RMSE of CNN-based methods when the multi-scene ShanghaiTech data set was used for training, either part_A (ShTechA) or part_B (ShTechB). For comparison, best and worst cross-scene results achieved on single-scene training data (S-best and S-worst) are reported from Table 4.4. For each method and target data set, multi-scene results that are better than the *best* single-scene ones are highlighted in boldface.

	Training set	Testing set (target scene) Mall MAE	Mall RMSE	UCSD MAE	UCSD RMSE	PETSview1 MAE	PETSview1 RMSE	PETSview2 MAE	PETSview2 RMSE	PETSview3 MAE	PETSview3 RMSE
MCNN	ShTechA	16.16	16.77	**18.88**	**19.64**	9.3	10.04	10.26	11.98	33.9	38.67
	ShTechB	21.03	21.58	22.01	22.86	7.51	8.58	23.2	24.86	**6.55**	**8.12**
	S-best	3.39	4.27	19.62	20.92	5.94	7.83	9.67	10.95	9.77	11.75
	S-worst	86.39	88.04	24.64	25.75	144.9	149.6	49.4	56.85	180.6	181.2
CMTL	ShTechA	17.71	18.33	**21.0**	**21.84**	8.51	9.39	**10.36**	**11.92**	33.46	40.68
	ShTechB	13.92	14.6	22.26	23.02	10.32	11.38	17.95	19.89	9.61	12.39
	S-best	4.61	5.79	21.94	23.12	5.77	7.42	11.5	13.97	4.46	5.95
	S-worst	189.1	191.1	24.63	25.76	213.7	217.9	111.9	113.7	298.5	300.8
DAN	ShTechA	16.76	17.32	23.96	24.67	8.88	10.21	14.49	16.56	15.68	16.68
	ShTechB	18.02	18.64	22.82	24.01	8.93	10.71	19.19	22.03	20.13	21.11
	S-best	7.9	9.48	18.8	20.12	7.52	9.43	11.7	13.14	6.34	7.74
	S-worst	163.1	166.1	27.86	29.0	185.9	192.1	61.76	66.53	227.3	228.5
SFCN	ShTechA	773.2	777.4	**5.42**	**7.55**	30.59	31.5	802.1	802.3	683.6	687.4
	ShTechB	31.21	32.4	322.7	323.7	**10.88**	**12.46**	238.5	238.5	33.8	34.3
	S-best	8.33	9.64	25.93	26.85	11.49	13.64	10.03	12.74	10.74	12.05
	S-worst	880.2	882.1	28.23	29.36	853.5	859.6	634.3	635.5	988.4	990.6
CSRN	ShTechA	14.64	15.1	26.58	27.63	8.58	10.08	**8.92**	**10.17**	15.45	16.55
	ShTechB	10.61	11.1	28.06	29.2	10.97	12.11	12.28	13.83	15.44	16.62
	S-best	9.21	9.89	24.51	25.8	5.84	6.8	9.66	10.56	8.42	9.53
	S-worst	70.78	71.46	27.49	28.62	57.52	61.86	28.29	31.21	69.06	69.36
CAN	ShTechA	9.72	10.28	**27.04**	**28.16**	**5.04**	**5.87**	**6.2**	**7.46**	10.3	11.67
	ShTechB	**3.6**	**4.56**	28.05	29.18	**6.53**	**8.25**	10.31	11.49	15.57	16.55
	S-best	6.73	7.7	28.09	29.23	7.5	9.07	8.43	9.25	3.94	4.84
	S-worst	281.6	283.1	28.09	29.23	173.5	176.9	133.4	135.2	252.0	252.4
SCAR	ShTechA	738.4	739.2	520.4	521.1	997.9	999.5	918.9	919.9	911.7	913.5
	ShTechB	512.9	513.5	**326.2**	**327.2**	813.5	815.5	829.9	811.7	825.6	826.1
	S-best	19.43	20.98	372.3	372.8	11.88	13.53	6.67	8.19	15.3	17.83
	S-worst	314.6	315.8	575.9	578.8	42.3	45.4	122.5	128.9	134.7	135.2
DSSI	ShTechA	8.44	9.16	**20.41**	**21.06**	**7.91**	**9.46**	**8.91**	**9.9**	11.73	13.55
	ShTechB	12.93	13.47	**26.24**	**27.2**	**13.47**	**15.52**	**9.88**	**11.68**	25.65	26.1
	S-best	4.14	6.47	37.35	37.81	20.25	22.46	12.63	15.0	10.31	11.36
	S-worst	25.6	26.84	69.02	69.8	27.83	28.53	26.92	28.11	26.52	27.72
BL+	ShTechA	**6.07**	**7.05**	**16.63**	**17.08**	**5.28**	**6.28**	7.77	9.48	16.51	17.36
	ShTechB	**6.78**	**7.57**	**18.52**	**19.2**	**4.21**	**5.34**	**7.05**	**8.9**	10.07	11.85
	S-best	10.09	11.81	77.63	80.9	6.9	7.86	7.44	9.0	8.22	9.98
	S-worst	26.01	26.69	152	153.63	22.65	23.8	21.17	22.0	23.66	24.77

Table 4.6: Comparison between the performance (MAE and RMSE) attained by all the considered crowd counting methods using as a training set: scene-specific synthetic images of the target data set ("Synthetic"), real images from the same scene ("Real-same"), and real images from different scenes ("Real-cross": best results over all single-scene training sets for early regression-based methods, and over the two ShanghaiTech training sets for CNN-based methods). For each data set and method the cases in which using synthetic training sets outperformed the best cross-data set results are highlighted in bold.

Method	Training set	Testing set (target scene) Mall MAE	Mall RMSE	UCSD MAE	UCSD RMSE	PETSview1 MAE	PETSview1 RMSE	PETSview2 MAE	PETSview2 RMSE	PETSview3 MAE	PETSview3 RMSE
LR	Real-same	2.74	3.49	2.9	3.54	6.25	7.91	4.85	5.98	6.84	8.42
	Real-cross	12.15	14.01	9.59	11.63	97.86	127.0	33.43	38.0	9.35	11.17
	Synthetic	14.94	16.34	**4.74**	**7.09**	**23.25**	**27.08**	**19.14**	**30.16**	26.6	33.19
RF	Real-same	3.82	4.85	3.82	4.66	9.47	11.03	9.56	11.05	11.41	12.49
	Real-cross	3.89	5.07	5.12	7.42	8.01	8.94	8.06	10.46	4.44	6.59
	Synthetic	6.76	8.1	**3.12**	**3.59**	**7.51**	9.13	18.35	23.41	7.82	9.61
SVR	Real-same	4.8	6.29	5.38	7.31	12.82	15.25	12.6	13.81	9.5	10.37
	Real-cross	5.11	6.71	5.13	7.3	9.56	10.45	9.8	10.68	8.74	9.55
	Synthetic	7.98	9.57	**2.85**	**4.13**	**6.96**	**8.66**	**8.83**	**10.63**	**4.6**	**6.54**
PLS	Real-same	3.16	4.1	2.6	3.23	8.46	10.13	7.65	9.06	9.03	10.06
	Real-cross	18.05	18.67	5.1	7.27	14.55	16.86	16.97	20.94	21.07	26.56
	Synthetic	**13.39**	**16.29**	5.16	**6.46**	17.06	21.32	29.1	30.59	**11.22**	**14.05**
MCNN	Real-same	5.33	6.17	2.3	2.84	6.2	7.86	4.23	5.08	4.18	5.13
	Real-cross	16.16	16.77	18.88	19.64	7.51	8.58	10.26	11.98	6.55	8.12
	Synthetic	20.73	21.68	**2.94**	**3.65**	12.22	13.43	17.86	18.67	11.39	13.69
CMTL	Real-same	5.53	6.39	2.04	2.50	5.11	6.29	4.80	6.06	4.23	5.06
	Real-cross	13.92	14.6	21.0	21.84	8.51	9.39	10.36	11.92	9.61	12.39
	Synthetic	22.96	23.47	**8.4**	**9.65**	9.43	11.09	**9.39**	**10.57**	**8.74**	**11.19**
DAN	Real-same	5.43	6.42	5.18	6.39	4.92	6.15	28.68	30.37	4.63	5.92
	Real-cross	16.76	17.32	22.82	24.01	8.88	10.21	14.49	16.56	15.68	16.68
	Synthetic	17.51	18.49	**10.31**	**12.21**	**4.05**	**5.37**	19.37	22.32	**10.55**	**12.56**
SFCN	Real-same	4.05	5.02	2.91	3.64	6.32	7.57	8.1	9.81	4.35	5.68
	Real-cross	31.21	32.4	5.42	7.55	10.88	12.4	238.5	238.5	33.8	34.3
	Synthetic	**17.76**	**18.57**	6.34	**7.34**	15.56	16.85	**23.22**	**24.82**	**10.19**	**12.46**
CSRN	Real-same	6.57	7.73	6.2	7.01	5.54	6.83	7.09	7.9	2.9	3.76
	Real-cross	10.61	11.1	26.58	27.63	8.58	10.08	8.92	10.17	15.45	16.55
	Synthetic	19.9	20.18	**3.45**	**4.8**	13.35	15.42	21.33	23.78	20.01	20.55
CAN	Real-same	2.59	3.21	4.73	6.16	6.33	7.5	6.07	7.67	6.82	7.84
	Real-cross	3.6	4.56	27.04	28.16	5.04	5.87	6.2	7.46	10.3	11.67
	Synthetic	16.77	17.26	**7.35**	**8.0**	12.78	14.4	16.99	19.19	30.95	31.36
SCAR	Real-same	3.99	4.75	4.19	5.24	3.38	4.07	5.09	6.32	8.39	10.32
	Real-cross	512.93	513.47	326.24	327.2	813.47	815.52	829.88	811.68	825.65	826.1
	Synthetic	**23.54**	**24.0**	**7.83**	**8.88**	**8.35**	**9.59**	**7.77**	**10.53**	**15.18**	**16.61**
DSSI	Real-same	5.44	7.09	21.75	23.2	18.0	20.44	14.64	16.51	11.46	12.46
	Real-cross	8.44	9.16	20.41	21.06	7.91	9.46	8.91	9.9	11.73	13.55
	Synthetic	28.91	29.5	**14.86**	**16.91**	19.18	21.81	21.29	23.58	29.48	30.02
BL+	Real-same	2.18	2.74	2.5	3.57	3.75	5.12	5.8	6.57	4.72	5.61
	Real-cross	6.07	7.05	16.63	17.08	4.21	5.34	7.05	8.9	10.07	11.85
	Synthetic	15.5	15.87	**7.85**	**8.59**	8.01	10.1	12.23	13.71	18.74	19.42

Table 4.6 reported the results obtained by using the proposed approach. For an easy comparison, the best cross-scene and the "ideal" (same-scene) results on training sets made up of real images are reported in table 4.6. For most of crowd counting methods, the use of synthetic images improved performances, especially when the UCSD data set was used as the target scene. For early regression-based approaches, the use of synthetic images as a training set lead in many cases to better or comparable performances with respect to the corresponding best cross-scene results. These results are highlighted in bold in table 4.6. Moreover, for RF and SVR these performances were even better than in the same-scene setting. Only in few cases the performance of synthetic images provided worse results than the best cross-scene ones: this is the case of LR, RF and SVR models when Mall was used as the target scene.

Also for CNN-based approaches synthetic images generally provided better or comparable performances with the cross-scene ones. In few cases these results were even better than the "ideal" performance, such as for DAN on PETSview1, and CSRN and DSSI on UCSD. The use of synthetic images significantly improved the performance of SCAR model, that however exhibited very poor performances using real training data, as mentioned above (see tables 4.1 and 4.5). The largest gap between synthetic images and best cross-scene performance can be observed for MCNN, DAN, CSRN, CAN, DSSI and BL+ for some data sets. For instance, MCNN reduced the MAE of the best cross-scene case by six times, whereas CSRN achieved a reduction of seven times.

The influence of the synthetic images on CNN-based methods was further investigated, since these methods present an intermediate phase in which the density map is estimated. Interestingly, this analysis showed that the use of synthetic images improves the quality of the estimated density map. In particular, pedestrians are located in target images with higher precision with respect to the case when real images of a different scene are used for training. Figure 4.3 shows an example of density maps produced by the MCNN model on two frames of PETSview1 and PETSview2, when the training set was made up of synthetic images (top), real images from PETSview3 (middle) and real images from the multi-scene ShanghaiTech data set (bottom). The improvement of the density map is evident also in data sets that did not exhibit improvements in crowd counting accuracy; for instance, this is the case of MCNN on

PETSview1 and PETSview2: despite synthetic training images were not beneficial for its crowd counting accuracy, they lead to a more precise density map.

Figure 4.3: Density maps produced on two frames of PETSview1 (left) and PETSview2 (right) by the MCNN model trained on: synthetic images (top), the single-scene PETSview3 data set (middle), and the multi-scene ShanghaiTech PartB data set (bottom). The ground truth (red) and estimated (green) density maps are superimposed on the original frames. The yellow regions are the ones where the two maps coincide, corresponding to prefect localisation of pedestrians. The highest localisation accuracy is achieved when synthetic training images are used (top).

4.4 Ablation study

In this section, the influence of parameters N (number of synthetic training images) and $n_{\max}$ (maximum number of pedestrians in synthetic images) on the accuracy of crowd counting models are evaluated. To this aim, different subsets of synthetic images with different values of N and $n_{\max}$ were randomly selected. Since re-training all the considered models in each image subset was too time-consuming, a representative set of methods has been selected with the following criteria: one early regression-based method, one CNN-based method trained from scratch, one CNN-based method trained using image patches, one CNN-based method trained using whole images, one CNN-based method which uses a fixed kernel and one CNN-based method which uses an adaptive kernel. The methods that fulfilled the above criteria are RF, MCNN, DAN and BL+.

In fig. 4.4 the MAE values on five target scene (Mall, UCSD, PETSview1, PETSview2 and PETSview3) are reported for the selected methods as a function of the size N of the training set. In most of cases, the increase of N did not provide a decreasing MAE. On the contrary, for RF the MAE slightly increased on the target scene PETSview2. Interestingly, this means that it might not be necessary to use a large synthetic data set of the target scene; as a consequence, the training phase can require a limited amount of time.

In fig. 4.5, the MAE values are reported as a function of the maximum number of pedestrians in training images, $n_{\max}$. In this case the considered range of crowd size was from 20 to 100 with a step of 20, and the total number of training images was set to 1,000 (the same value considered in previous experiments). The generated synthetic data set was divided into a training and a validation set: 200 images were used for validation, whereas the remaining ones were used for training. In these experiments, the behaviour of RF was significantly different from the other selected CNN-based models. In particular, it attained a minimum MAE when $n_{\max}$ was closest to the maximum number of pedestrians present in the corresponding target scene. The MAE of CNN-based models (in particular, MCNN and BL+) exhibited a slightly decreasing behaviour as the number of pedestrians increased, except for DAN. These results show that the early regression-based methods are more sensitive than CNN-based ones to the maximum number of pedestrians $n_{\max}$.

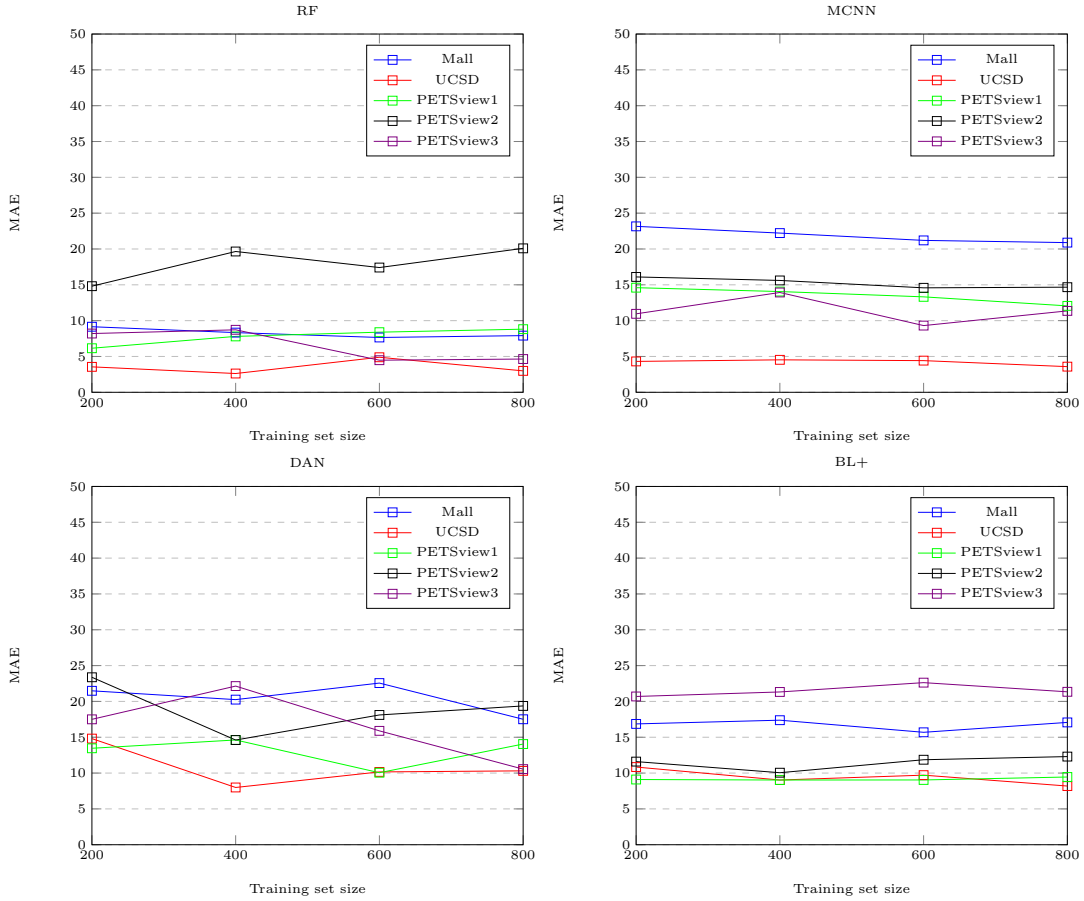


Figure 4.4: MAE values achieved by RF, MCNN, DAN and BL+ on the five target scenes using synthetic training data, as a function of training set size.

4.5 Discussion

In this chapter the use of synthetic images to train scene-specific crowd counting models was proposed for challenging cross-scene application scenarios where it is not possible to collect and annotate real images of the target scene. The proposed approach requires minimal effort to the end users.

Although the proposed approach is easier than other approaches based on synthetic images proposed for other computer vision tasks [79], as well as for crowd counting [135], it exhibits some limitations. Although the generation of synthetic data sets required about 15 minutes (on a single machine with the following configuration: Intel(R) Core(TM) i9-8950HK @ 2.90GHz CPU with 32 GB RAM and NVIDIA GTX1050

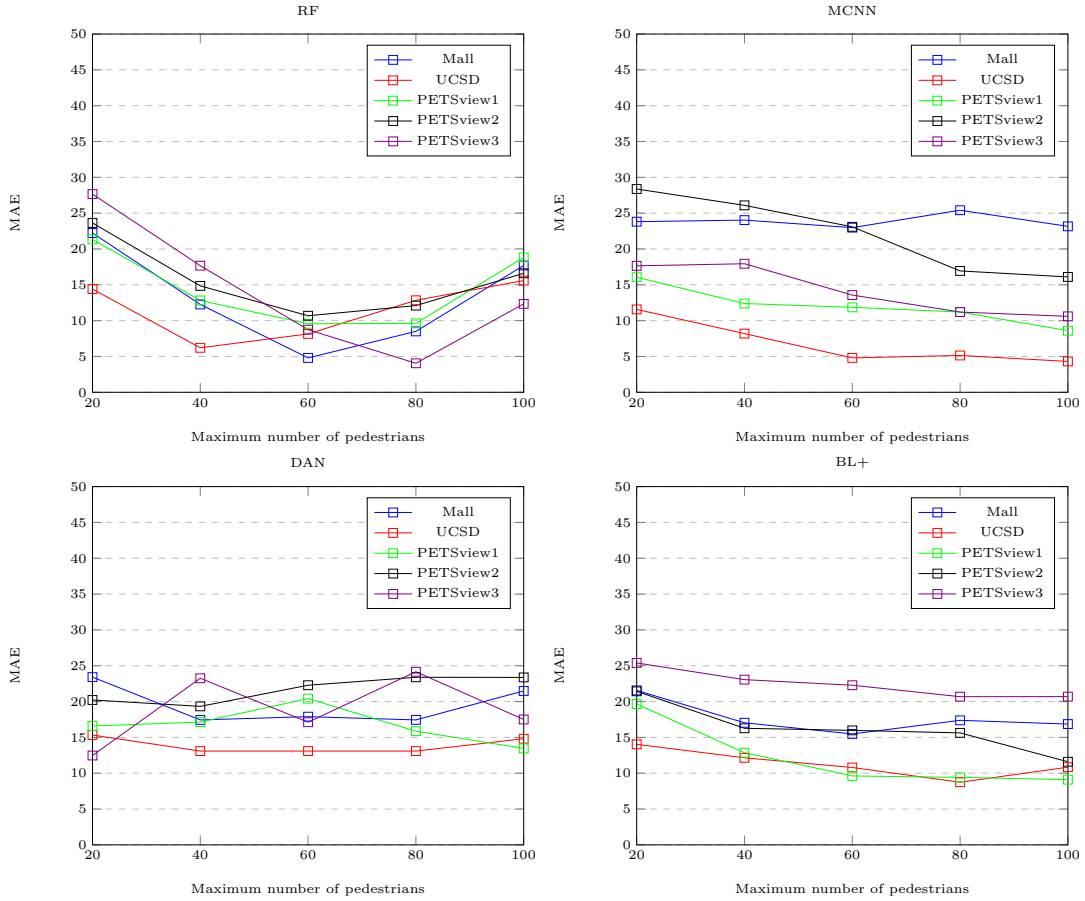


Figure 4.5: MAE values achieved by RF, MCNN, DAN and BL+ on the five target scenes using synthetic training data, as a function of the maximum number of pedestrians in training images.

Ti 4GB GPU), the processing time required for training some of the considered CNN models using 1,000 training images took a few days, which can be excessive for some application scenarios. In this case, CNN models relatively fast to train should be used.

Another issue is the computation of the perspective map on a non-flat scenes (e.g., a concert in a stadium where some people are on the ground and other in the terraces). In this case, more than three bounding boxes should be selected by the end users on an image of the target scene.

Another issue is that the proposed approach focuses on *fixed* camera views, and may be ineffective for pan-tilt-zoom camera, known as PTZ cameras that allow the operator to change the camera view. To address this issue, different crowd counting

models could be trained on different target scenes selected by the operator as the most useful ones, for a given PTZ camera.

As for person re-identification task (see sect. 3.3) a prototype of crowd counting and density estimation system was also developed and tested on a real and challenging scenario during practical demonstrations of the LETSCROWD project. As for person re-identification prototype, the practical demonstrations allowed collecting feedback of potential end users about the system's performance and usability, as well their expectations about this kind of tool.

One of their expectations is related to the accuracy under different environmental conditions, e.g., illumination, crowd size, etc. It is worth reminding that this kind of tool provides an estimation of the number of people and the accuracy depends on the application scenario. For instance, consider a venue of maximum predefined capacity. When the number of people approaches the maximum capacity, even slight variations are relevant, and therefore it is important to detect them. In contrast, when the number of people is significantly lower than the maximum capacity, it is possible to tolerate a lower accuracy. It is well known that crowd counting and density estimation accuracy can be affected by environmental conditions, i.e., during the day the accuracy might be higher than during the night with artificial illumination, or under different weather conditions (rain, fog, etc.). Therefore, it would be useful that the end users are aware of the expected accuracy under different environmental conditions.

Another issue is related to the kind of density map produced by existing methods. Usually it is represented as a heat map where colours from blue to red indicate low to high density. This is easy to interpret in principle. However, density maps produced by existing methods indicate the number of people *per pixel*, whereas end users are interested in the number of people *per unit area* on the ground plane. Due to the typical image perspective of the surveillance cameras, the number of people per pixel is not proportional to the number of people per unit area with a constant proportionality factor over the whole image. For instance, by considering an image where the crowd density is constant in terms of the number of people per unit area in the whole region of interest, it is clear that the number of people per pixel will be lower in image regions nearest to the camera than in farther regions. By supposing that the perspective map is known, as in the proposed method of this thesis, a possible solution to fulfil the expectations of the end users, can be to include a post-processing phase of the estimated

density map to correct it; as an alternative solution, the perspective information could be embedded into a specific CNN architecture that capable of producing a density map with the above mentioned requirement.

In general, as mentioned in sec. 3.3 about the person re-identification task, end users' training is necessary in order to enable them to correctly interpret the results produced by tool.

Chapter 5

Conclusions

This thesis focused on two computer vision tasks that are relevant to intelligent video surveillance system in real-world applications, namely person re-identification and crowd counting. They can provide useful functionalities to support human operators, who often simultaneously monitor or analyse several videos acquired by a video surveillance system. In particular, in this thesis, a challenging cross-scene scenario was considered when collecting a suitable amount of representative images (even unlabelled) of the target scene is infeasible or too demanding for the end users. Supervised state-of-the-art methods exhibit a significant performance decrease in a cross-scene scenario. For both the above mentioned computer vision tasks, a human centred approach was proposed.

With regard to person re-identification, the HITL approach is considered since it can be an interesting alternative to approaches like UDA to adapt a person re-identification system to the target scene. In particular, in this thesis person re-identification was considered as an image retrieval problem, and for this reason I focused on CBIR-RF techniques which have been disregarded so far in literature. A specific contribution of my thesis is a feedback protocol different from the ones used in other HITL person re-identification methods, which takes into account the characteristics of CBIR-RF techniques and of the considered computer vision tasks. The proposed feedback protocol, named multi-feedback, consists of asking the end user to select *all* true matches in the top ranked list, if any (instead of a *single* true match or strong negative match), whereas the other top ranked images are automatically considered as non-relevant. Experimental results showed that CBIR-RF algorithms with the proposed feedback protocol are able to improve re-identification performance in cross-scene scenarios (sect. 3.2.6). The

main limit of the experimental analysis is the lack of comparison with existing HITL methods for person re-identification, since their source code was not made available by the authors and their complexity did not allow to re-implement them. Some issues of the proposed approach were discussed in sect. 3.3, i.e., the number of feedback to be provided by the end user, and the fact that the considered RF algorithms are not incremental. The number of feedback required by the proposed approach may appear too demanding for the end user. Nevertheless, in a real application scenario the gallery can be mach larger than in benchmark data sets, and therefore finding several true matches in the top ranked images is unlikely. Moreover, the true matches provided in a feedback round should not be selected again in the next round. This means that the required user effort is not significantly larger than in the existing HITL methods under the proposed multi-feedback protocol. Although the use of non-incremental RF algorithms might limit the adaptation of the source model to the target scene, the considered HITL solution is fast in re-ranking the gallery.

With regard to crowd counting, the first contribution of this thesis was an extensive evaluation of the performance gap between same- and cross-scene performance of several state-of-the-art methods (early regression-based and CNN-based), which is still missing in the literature. The second contribution is a possible solution to mitigate this gap, by using synthetic images of the target scene to build a scene-specific training set for supervised models, requiring minimal human effort compatible with the considered application scenario (sect. 4.2). In particular, the information required to the end user is a single background image of the target scene, the ROI in terms of binary map, and the selection of a few pedestrian bounding boxes in images of the target scene that allows the perspective map to be automatically computed. This information can be easily provided during camera set-up. An additional information that can be provided by the user is the expected range of crowd size. The construction of the synthetic training set consists of superimposing to the background image pedestrian images placed in random locations of the ROI, scaled accordingly to the perspective map. Experimental results showed that in most cases the proposed solution, when implemented on the considered crowd counting methods, allow them to achieve better or comparable cross-scene (sect. 4.3), and in few cases performances were even better than under the "ideal" same-scene setting. An interesting result is related to the estimation of density map produced by CNN-based methods as an intermediate step:

the use of synthetic training images turned out to improve the quality of the density map, i.e., pedestrians are located in the target images with a higher precision with respect to the use of real training images of a different scene. Such improvement was also observed in cases where synthetic images did not provide improvements in crowd counting accuracy. The proposed approach has a lower complexity than approaches based on synthetic images proposed for other computer vision tasks; at the same time it has some limitations (sect. 4.5). One limitation is related to the computation of the perspective map in particular scenarios, e.g. a non-flat scene. Such scenes require the selection of several bounding boxed to compute the perspective map. Another limitation is related to the fact that a fixed camera view is considered in this thesis, which does not take into account PTZ camera. A possible solution could be to train different crowd counting models on different target scenes selected by the end user as the most interesting ones on a given PTZ camera.

References

[1] Timo Ahonen, Abdenour Hadid, and Matti Pietikäinen. **Face recognition with local binary patterns**. In *European conference on computer vision (ECCV)*, pages 469–481. Springer, 2004. Available from: http://dx.doi.org/10.1007/978-3-540-24670-1_36. 15

[2] Saad Ali, Niels Haering, and Takeo Kanade. **Interactive retrieval of targets for wide area surveillance**. In *ACM Multimedia*, pages 895–898, 2010. Available from: http://dx.doi.org/10.1145/1873951.1874106. 21, 32

[3] T. M. Feroz Ali and Subhasis Chaudhuri. **Cross-View Kernel Similarity Metric Learning Using Pairwise Constraints for Person Re-identification**. *CoRR*, **abs/1909.11316**, 2019. 16

[4] Slawomir Bak, Peter Carr, and Jean-François Lalonde. **Domain Adaptation Through Synthesis for Unsupervised Person Re-identification**. In *European Conference on Computer Vision (ECCV)*, pages 193–209, 2018. Available from: http://dx.doi.org/10.1007/978-3-030-01261-8_12. 20

[5] Nathaniel D. Bird, Osama Masoud, Nikolaos P. Papanikolopoulos, and Aaron Isaacs. **Detection of loitering individuals in public transportation areas**. *IEEE Transactions on Intelligent Transportation Systems*, **6**(2):167–177, 2005. Available from: http://dx.doi.org/10.1109/TITS.2005.848370. 15

[6] Gilles Blanchard, Gyemin Lee, and Clayton Scott. **Generalizing from Several Related Classification Tasks to a New Unlabeled Sam-**

ple. In *Neural Information Processing Systems (NIPS)*, pages 2178–2186, 2011. 14

[7] Steve Branson, Grant Van Horn, Catherine Wah, Pietro Perona, and Serge J. Belongie. **The Ignorant Led by the Blind: A Hybrid Human-Machine Vision System for Fine-Grained Categorization**. *International Journal of Computer Vision*, **108**(1-2):3–29, 2014. Available from: https://doi.org/10.1007/s11263-014-0698-4. 21

[8] Steve Branson, Catherine Wah, Florian Schroff, Boris Babenko, Peter Welinder, Pietro Perona, and Serge J. Belongie. **Visual Recognition with Humans in the Loop**. In *European conference on computer vision (ECCV)*, pages 438–451, 2010. Available from: https://doi.org/10.1007/978-3-642-15561-1_32. 21

[9] Antoni B Chan, Zhang-Sheng John Liang, and Nuno Vasconcelos. **Privacy preserving crowd monitoring: Counting people without people models or tracking**. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–7, 2008. Available from: http://dx.doi.org/10.1109/CVPR.2008.4587569. 63

[10] Guangyi Chen, Chunze Lin, Liangliang Ren, Jiwen Lu, and Jie Zhou. **Self-Critical Attention Learning for Person Re-Identification**. In *IEEE International Conference on Computer Vision (ICCV)*, pages 9636–9645, 2019. Available from: http://dx.doi.org/10.1109/ICCV.2019.00973. 16, 18, 19

[11] Ke Chen, Chen Change Loy, Shaogang Gong, and Tony Xiang. **Feature mining for localised crowd counting.** In *British Machine Vision Conference (BMVC)*, pages 1–11, 2012. Available from: http://dx.doi.org/10.5244/C.26.21. 63

[12] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. **Instance-Guided Context Rendering for Cross-Domain Person Re-Identification**. In *IEEE International Conference on Computer Vision (ICCV)*, pages 232–242, 2019. Available from: http://dx.doi.org/10.1109/ICCV.2019.00032. 27, 30, 56, 57

[13] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. **Person Re-identification by Multi-Channel Parts-Based CNN with Improved Triplet Loss Function**. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1335–1344, 2016. Available from: http://dx.doi.org/10.1109/CVPR.2016.149. 16, 17, 18, 19

[14] Nicolas Courty, Pierre Allain, Clement Creusot, and Thomas Corpetti. **Using the AGORASET dataset: Assessing for the quality of crowd video analysis methods**. *Pattern Recognition Letters*, **44**:161–170, 2014. Available from: http://dx.doi.org/10.1016/j.patrec.2014.01.004. 30, 56, 57

[15] Antonio Criminisi, Jamie Shotton, and Ender Konukoglu. **Decision forests for classification, regression, density estimation, manifold learning and semi-supervised learning**. *Microsoft Research Cambridge, Tech. Rep. MSRTR-2011-114*, **5**(6):12, 2011. Available from: http://dx.doi.org/10.1609/aaai.v33i01.33018868. 28

[16] Gabriela Csurka. **A Comprehensive Survey on Domain Adaptation for Visual Applications**. In *Domain Adaptation in Computer Vision Applications*, pages 1–35. 2017. Available from: http://dx.doi.org/10.1007/978-3-319-58347-1_1. 8, 19

[17] Jason V. Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S. Dhillon. **Information-theoretic metric learning**. In *Machine Learning, Proceedings of the Twenty-Fourth International Conference*, pages 209–216, 2007. Available from: http://dx.doi.org/10.1145/1273496.1273523. 15

[18] Rita Delussu, Lorenzo Putzu, and Giorgio Fumera. **An Empirical Evaluation of Cross-scene Crowd Counting Performance**. In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISIGRAPP*, pages 373–380, 2020. Available from: http://dx.doi.org/10.5220/0008983003730380. 55, 56, 57, 66

[19] Rita Delussu, Lorenzo Putzu, and Giorgio Fumera. **Investigating Synthetic Data Sets for Crowd Density Estimation**. 2020. 66

[20] Rita Delussu, Lorenzo Putzu, and Giorgio Fumera. **Scene-specific Crowd Counting Using Synthetic Training Images**. *Pattern Recognition*, (under review). 58

[21] Rita Delussu, Lorenzo Putzu, Giorgio Fumera, and Fabio Roli. **Online Domain Adaptation for Person Re-Identification with a Human in the Loop**. *IEEE International Conference on Pattern Recognition (ICPR)*, (accepted). 34

[22] Begüm Demir and Lorenzo Bruzzone. **A Novel Active Learning Method in Relevance Feedback for Content-Based Remote Sensing Image Retrieval**. *IEEE Transactions on Geoscience and Remote Sensing*, **53**(5):2323–2334, 2015. Available from: http://dx.doi.org/10.1109/TGRS.2014.2358804. 20, 21

[23] Jia Deng, Jonathan Krause, Michael Stark, and Fei-Fei Li. **Leveraging the Wisdom of the Crowd for Fine-Grained Recognition**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **38**(4):666–676, 2016. Available from: https://doi.org/10.1109/TPAMI.2015.2439285. 21

[24] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. **Pedestrian detection: An evaluation of the state of the art**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **34**(4):743–761, 2011. Available from: http://dx.doi.org/10.1109/TPAMI.2011.155. 27

[25] Zihao Dong, Ruixun Zhang, Xiuli Shao, and Yumeng Li. **Scale-Recursive Network with point supervision for crowd scene analysis**. *Neurocomputing*, **384**:314–324, 2020. Available from: http://dx.doi.org/10.1016/j.neucom.2019.12.070. 29

[26] Carlos Arango Duque, Ani Khachatryan, and Sule Yildirim Yayilgan. **A relevance feedback based image retrieval approach for improved performance**. In *Colour and Visual Computing Symposium (CVCS)*, pages 1–6, 2015. Available from: http://dx.doi.org/10.1109/CVCS.2015.7274883. 21

[27] Ehsan Elhamifar, Guillermo Sapiro, Allen Y. Yang, and S. Shankar Sastry. **A Convex Optimization Framework for Active Learning**. In *IEEE International Conference on Computer Vision (ICCV)*, pages 209–216, 2013. Available from: http://dx.doi.org/10.1109/ICCV.2013.33. 21

[28] Hehe Fan, Liang Zheng, Chenggang Yan, and Yi Yang. **Unsupervised Person Re-identification: Clustering and Fine-tuning**. *ACM Trans. Multim. Comput. Commun. Appl.*, **14**(4):83:1–83:18, 2018. Available from: https://doi.org/10.1145/3243316. 14

[29] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. **Person re-identification by symmetry-driven accumulation of local features**. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2360–2367, 2010. Available from: http://dx.doi.org/10.1109/CVPR.2010.5539926. 15

[30] Pedro F. Felzenszwalb, Ross B. Girshick, David A. McAllester, and Deva Ramanan. **Object Detection with Discriminatively Trained Part-Based Models**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **32**(9):1627–1645, 2010. Available from: http://dx.doi.org/10.1109/TPAMI.2009.167. 38

[31] Lu Feng, Clemens Wiltsche, Laura R. Humphrey, and Ufuk Topcu. **Synthesis of Human-in-the-Loop Control Protocols for Autonomous Systems**. *IEEE Transactions on Automation Science and Engineering*, **13**(2):450–462, 2016. Available from: http://dx.doi.org/10.1109/TASE.2016.2530623. 20

[32] James Ferryman and Ali Shahrokni. **Pets2009: Dataset and challenge**. In *2009 Twelfth IEEE international workshop on performance evaluation of tracking and surveillance (PETS)*, pages 1–6, 2009. Available from: http://dx.doi.org/10.1109/PETS-WINTER.2009.5399556. 64

[33] Ronald A Fisher. **The use of multiple measurements in taxonomic problems**. *Annals of eugenics*, **7**(2):179–188, 1936. 15

REFERENCES

[34] Yang Fu, Yunchao Wei, Guanshuo Wang, Yuqian Zhou, Honghui Shi, Uiuc Uiuc, and Thomas Huang. **Self-Similarity Grouping: A Simple Unsupervised Cross Domain Adaptation Approach for Person Re-Identification**. In *IEEE International Conference on Computer Vision (ICCV)*, pages 6111–6120, 2019. Available from: http://dx.doi.org/10.1109/ICCV.2019.00621. 19

[35] Junyu Gao, Qi Wang, and Yuan Yuan. **SCAR: Spatial-/channel-wise attention regression networks for crowd counting**. *Neurocomputing*, **363**:1–8, 2019. Available from: http://dx.doi.org/10.1016/j.neucom.2019.08.018. 29, 62, 63

[36] Weina Ge and Robert T Collins. **Marked point processes for crowd counting**. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2913–2920, 2009. Available from: http://dx.doi.org/10.1109/CVPR.2009.5206621. 27

[37] Yixiao Ge, Dapeng Chen, and Hongsheng Li. **Mutual Mean-Teaching: Pseudo Label Refinery for Unsupervised Domain Adaptation on Person Re-identification**. *CoRR*, **abs/2001.01526**, 2020. 19, 36

[38] Paul Geladi and Bruce R Kowalski. **Partial least-squares regression: a tutorial**. *Analytica chimica acta*, **185**:1–17, 1986. Available from: http://dx.doi.org/10.1016/0003-2670(86)80028-9. 28

[39] Anil Genç and Hazim Kemal Ekenel. **Cross-dataset person re-identification using deep convolutional neural networks: effects of context and domain adaptation**. *Multimedia Tools and Applications*, **78**(5):5843–5861, 2019. Available from: http://dx.doi.org/10.1007/s11042-018-6409-3. 8, 19

[40] Muhammad Ghifary, W. Bastiaan Kleijn, Mengjie Zhang, David Balduzzi, and Wen Li. **Deep Reconstruction-Classification Networks for Unsupervised Domain Adaptation**. In *European Conference on Computer Vision (ECCV)*, pages 597–613, 2016. Available from: http://dx.doi.org/10.1007/978-3-319-46493-0_36. 19

[41] Giorgio Giacinto. **A nearest-neighbor approach to relevance feedback in content based image retrieval**. In *ACM International Conference on Image and Video Retrieval (CIVR)*, pages 456–463, 2007. 37

[42] Douglas Gray and Hai Tao. **Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features**. In *European Conference on Computer Vision (ECCV)*, pages 262–275, 2008. Available from: http://dx.doi.org/10.1007/978-3-540-88682-2_21. 15

[43] Robert M. Haralick, K. Sam Shanmugam, and Its'hak Dinstein. **Textural Features for Image Classification**. *IEEE Transactions Systems, Man, and Cybernetics*, **3**(6):610–621, 1973. Available from: http://dx.doi.org/10.1109/TSMC.1973.4309314. 28, 61

[44] Hironori Hattori, Vishnu Naresh Boddeti, Kris M. Kitani, and Takeo Kanade. **Learning scene-specific pedestrian detectors without real data**. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3819–3827, 2015. Available from: http://dx.doi.org/10.1109/CVPR.2015.7299006. 30, 56, 57

[45] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. **Deep Residual Learning for Image Recognition**. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. Available from: http://dx.doi.org/10.1109/CVPR.2016.90. 17

[46] Xiaofei He and Partha Niyogi. **Locality Preserving Projections**. In *Neural Information Processing Systems (NIPS)*, pages 153–160, 2003. 15

[47] Martin Hirzer, Csaba Beleznai, Peter M. Roth, and Horst Bischof. **Person Re-identification by Descriptive and Discriminative Classification**. In *Scandinavian Conference on Image Analysis (SCIA)*, pages 91–102, 2011. Available from: http://dx.doi.org/10.1007/978-3-642-21227-7_9. iv, 21, 23, 32

[48] Arthur E Hoerl and Robert W Kennard. **Ridge regression: Biased estimation for nonorthogonal problems**. *Technometrics*, **12**(1):55–67, 1970. Available from: http://dx.doi.org/10.1080/00401706.2000.10485983. 28

REFERENCES

[49] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. **Interaction-And-Aggregation Network for Person Re-Identification**. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9317–9326, 2019. Available from: http://dx.doi.org/10.1109/CVPR.2019.00954. 16, 18

[50] Rongyao Hu, Xiaofeng Zhu, Debo Cheng, Wei He, Yan Yan, Jingkuan Song, and Shichao Zhang. **Graph self-representation method for unsupervised feature selection**. *Neurocomputing*, **220**:130–137, 2017. Available from: http://dx.doi.org/10.1016/j.neucom.2016.05.081. 21

[51] Shoubo Hu, Kun Zhang, Zhitang Chen, and Laiwan Chan. **Domain Generalization via Multidomain Discriminant Analysis**. In *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI*, page 101, 2019. 8, 14, 20

[52] Yangru Huang, Peixi Peng, Yi Jin, Junliang Xing, Congyan Lang, and Songhe Feng. **Domain Adaptive Attention Model for Unsupervised Cross-Domain Person Re-Identification**. *CoRR*, **abs/1905.10529**, 2019. 19

[53] Haroon Idrees, Muhmmad Tayyab, Kishan Athrey, Dong Zhang, Somaya Al-Máadeed, Nasir M. Rajpoot, and Mubarak Shah. **Composition Loss for Counting, Density Map Estimation and Localization in Dense Crowds**. In *European Conference on Computer Vision (ECCV)*, pages 544–559, 2018. Available from: https://doi.org/10.1007/978-3-030-01216-8_33. 25

[54] Mario Innocenti, Andrea Balluchi, and Aldo Balestrino. **Modeling of nonlinear human operator in the control loop: Preliminary results**. *Journal of Guidance, Control, and Dynamics*, **23**(4):736–739, 2000. 20

[55] Vidit Jain and Erik G. Learned-Miller. **Online domain adaptation of a pre-trained cascade of classifiers**. In *Computer Vision and Pattern Recognition (CVPR)*, pages 577–584, 2011. Available from: http://dx.doi.org/10.1109/CVPR.2011.5995317. 32

[56] Jieru Jia, Qiuqi Ruan, and Timothy M. Hospedales. **Frustratingly Easy Person Re-Identification: Generalizing Person Re-ID in Practice**. In *British Machine Vision Conference (BMVC)*, page 117, 2019. 20

[57] Simon Jones, Ling Shao, Jianguo Zhang, and Yan Liu. **Relevance feedback for real-world human action retrieval**. *Pattern Recognition Letters*, **33**(4):446–452, 2012. Available from: http://dx.doi.org/10.1016/j.patrec.2011.05.001. 21

[58] Julio Cezar Silveira Jacques Junior, Soraia Raupp Musse, and Claudio Rosito Jung. **Crowd Analysis Using Computer Vision Techniques**. *IEEE Signal Process. Mag.*, **27**(5):66–77, 2010. Available from: http://dx.doi.org/10.1109/MSP.2010.937394. 30, 56, 57

[59] Mahdi M. Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E. Kamasak, and Mubarak Shah. **Human Semantic Parsing for Person Re-Identification**. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1062–1071, 2018. Available from: http://dx.doi.org/10.1109/CVPR.2018.00117. 16, 19

[60] Dan Kong, Douglas Gray, and Hai Tao. **Counting Pedestrians in Crowds Using Viewpoint Invariant Training**. In *British Machine Vision Conference (BMVC)*, **1**, page 2. Citeseer, 2005. Available from: http://dx.doi.org/10.5244/C.19.63. 28, 61

[61] Martin Köstinger, Martin Hirzer, Paul Wohlhart, Peter M. Roth, and Horst Bischof. **Large scale metric learning from equivalence constraints**. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2288–2295, 2012. Available from: http://dx.doi.org/10.1109/CVPR.2012.6247939. 15

[62] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. **ImageNet Classification with Deep Convolutional Neural Networks**. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1106–1114, 2012. Available from: http://dx.doi.org/10.1145/3065386. 17

[63] Navaneet K. L., Ravi Kiran Sarvadevabhatla, Shashank Shekhar, R. Venkatesh Babu, and Anirban Chakraborty. **Operator-in-the-Loop Deep Sequential Multi-Camera Feature Fusion for Person Re-Identification**. *IEEE Transactions Information Forensics and Security*, **15**:2375–2385, 2020. Available from: http://dx.doi.org/10.1109/TIFS.2019.2957701. v, 21, 24, 25, 33

[64] Adam Eric Leeper, Kaijen Hsiao, Matei Ciocarlie, Leila Takayama, and David Gossow. **Strategies for human-in-the-loop robotic grasping**. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, pages 1–8, 2012. Available from: http://dx.doi.org/10.1145/2157689.2157691. 20

[65] Bastian Leibe, Edgar Seemann, and Bernt Schiele. **Pedestrian detection in crowded scenes**. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, **1**, pages 878–885, 2005. Available from: http://dx.doi.org/10.1109/CVPR.2005.272. 27

[66] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang. **Learning Deep Context-Aware Features over Body and Latent Parts for Person Re-identification**. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7398–7407, 2017. Available from: http://dx.doi.org/10.1109/CVPR.2017.782. 16

[67] Jingwen Li, Lei Huang, and Changping Liu. **Robust people counting in video surveillance: Dataset and system**. In *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 54–59, 2011. Available from: http://dx.doi.org/10.1109/AVSS.2011.6027294. 28

[68] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. **DeepReID: Deep Filter Pairing Neural Network for Person Re-identification**. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 152–159, 2014. Available from: http://dx.doi.org/10.1109/CVPR.2014.27. 16

[69] Wei Li, Xiatian Zhu, and Shaogang Gong. **Harmonious Attention Network for Person Re-Identification**. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2285–2294, 2018. Available from: http://dx.doi.org/10.1109/CVPR.2018.00243. 16, 19

[70] Yu-Jhe Li, Fu-En Yang, Yen-Cheng Liu, Yu-Ying Yeh, Xiaofei Du, and Yu-Chiang Frank Wang. **Adaptation and Re-Identification Network: An Unsupervised Deep Transfer Learning Approach to Person Re-Identification**. In *Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 172–178, 2018. Available from: http://dx.doi.org/10.1109/CVPRW.2018.00054. 19

[71] Yuhong Li, Xiaofan Zhang, and Deming Chen. **Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes**. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1091–1100, 2018. Available from: http://dx.doi.org/10.1109/CVPR.2018.00120. 29, 30, 63, 69

[72] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z. Li. **Person re-identification by Local Maximal Occurrence representation and metric learning**. In *Conference on Computer Vision and Pattern Recognition, CVPR*, pages 2197–2206, 2015. Available from: http://dx.doi.org/10.1109/CVPR.2015.7298832. 15, 16

[73] Sheng-Fuu Lin, Jaw-Yeh Chen, and Hung-Xin Chao. **Estimation of number of people in crowded scenes using perspective transformation**. *IEEE Transaction on Systems, Man, and Cybernetics*, **31**(6):645–654, 2001. Available from: http://dx.doi.org/10.1109/3468.983420. 27

[74] Wei-Chao Lin, Zong-Yao Chen, Shih-Wen Ke, Chih-Fong Tsai, and Wei-Yang Lin. **The effect of low-level image features on pseudo relevance feedback**. *Neurocomputing*, **166**:26–37, 2015. Available from: http://dx.doi.org/10.1016/j.neucom.2015.04.037. 37

[75] Wei-Chao Lin, Zong-Yao Chen, Shih-Wen Ke, Chih-Fong Tsai, and Wei-Yang Lin. **The effect of low-level image features on pseudo relevance feedback**. *Neurocomputing*, **166**:26–37, oct 2015. Available from: http://dx.doi.org/10.1016/J.NEUCOM.2015.04.037. 21

[76] Yutian Lin, Xuanyi Dong, Liang Zheng, Yan Yan, and Yi Yang. **A Bottom-Up Clustering Approach to Unsupervised Person Re-Identification**. In *AAAI Conference on Artificial Intelligence*, pages 8738–8745, 2019. Available from: https://doi.org/10.1609/aaai.v33i01.33018738. 14

[77] Chunxiao Liu, Chen Change Loy, Shaogang Gong, and Guijin Wang. **POP: Person Re-identification Post-rank Optimisation**. In *IEEE International Conference on Computer Vision (ICCV)*, pages 441–448, 2013. Available from: http://dx.doi.org/10.1109/ICCV.2013.62. v, 21, 22, 23, 32, 33, 40

[78] Fangyi Liu and Lei Zhang. **View Confusion Feature Learning for Person Re-Identification**. In *IEEE International Conference on Computer Vision (ICCV)*, pages 6638–6647, 2019. Available from: http://dx.doi.org/10.1109/ICCV.2019.00674. 16, 18, 19

[79] Jianlei Liu, Yun Zhou, Lingchuan Sun, and Zhuqing Jiang. **Similarity Preserved Camera-to-Camera Gan for Person Re-Identification**. In *IEEE International Conference on Multimedia & Expo Workshops (ICME)*, pages 531–536, 2019. Available from: http://dx.doi.org/10.1109/ICMEW.2019.00097. 27, 30, 56, 57, 75

[80] Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu. **Pose Transferrable Person Re-Identification**. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4099–4108, 2018. Available from: http://dx.doi.org/10.1109/CVPR.2018.00431. 19

[81] Lingbo Liu, Zhilin Qiu, Guanbin Li, Shufan Liu, Wanli Ouyang, and Liang Lin. **Crowd Counting With Deep Structured Scale Integration Network**. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1774–1783, 2019. Available from: http://dx.doi.org/10.1109/ICCV.2019.00186. 29, 30, 62, 63, 69

[82] Ning Liu, Yongchao Long, Changqing Zou, Qun Niu, Li Pan, and Hefeng Wu. **ADCrowdNet: An Attention-Injective Deformable Convolutional Network for Crowd Understanding**. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3225–3234, 2019. Available from: http://dx.doi.org/10.1109/CVPR.2019.00334. 29, 30

[83] Weizhe Liu, Mathieu Salzmann, and Pascal Fua. **Context-Aware Crowd Counting**. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5099–5108, 2019. Available from: http://dx.doi.org/10.1109/CVPR.2019.00524. 29, 30, 62, 63, 69

[84] David G. Lowe. **Distinctive Image Features from Scale-Invariant Keypoints**. *International Journal of Computer Vision*, **60**(2):91–110, 2004. Available from: http://dx.doi.org/10.1023/B:VISI.0000029664.99615.94. 15

[85] Chen Change Loy, Ke Chen, Shaogang Gong, and Tao Xiang. **Crowd counting and profiling: Methodology and evaluation**. In *Modeling, simulation and visual analysis of crowds*, pages 347–382. Springer, 2013. Available from: http://dx.doi.org/10.1007/978-1-4614-8483-7_14. 7, 8, 26, 27, 28, 58

[86] Ruihua Ma, Liyuan Li, Weimin Huang, and Qi Tian. **On pixel count based crowd density estimation for visual surveillance**. In *IEEE Computational Intelligence Society (CIS)*, **1**, pages 170–173, 2004. Available from: http://dx.doi.org/10.1109/ICCIS.2004.1460406. 28, 61

[87] Zhiheng Ma, Xing Wei, Xiaopeng Hong, and Yihong Gong. **Bayesian Loss for Crowd Count Estimation With Point Supervision**. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6141–6150, 2019. Available from: http://dx.doi.org/10.1109/ICCV.2019.00624. 29, 30, 62, 63, 69

[88] DI Mansur, MK Haque, K Sharma, DK Mehta, and R Shakya. **Use of head circumference as a predictor of height of individual**. *Kathmandu University Medical Journal (KUMJ)*, **12**(2):89–92, 2014. 60

[89] Alexis Mignon and Frédéric Jurie. **PCCA: A new approach for distance learning from sparse pairwise constraints**. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2666–2672, 2012. Available from: http://dx.doi.org/10.1109/CVPR.2012.6247987. 15, 16

[90] Chikahito Nakajima, Massimiliano Pontil, Bernd Heisele, and Tomaso A. Poggio. **Full-body person recognition system**. *Pattern Recognition*, **36**(9):1997–2006, 2003. Available from: http://dx.doi.org/10.1016/S0031-3203(03)00061-X. 15

[91] Bac Nguyen and Bernard De Baets. **Kernel Distance Metric Learning Using Pairwise Constraints for Person Re-Identification**. *IEEE Transactions Image Processing*, **28**(2):589–600, 2019. Available from: http://dx.doi.org/10.1109/TIP.2018.2870941. 16

[92] Timo Ojala, Matti Pietikäinen, and Topi Mäenpää. **Multiresolution gray-scale and rotation invariant texture classification with local binary patterns**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24**(7):971–987, 2002. Available from: http://dx.doi.org/10.1109/TPAMI.2002.1017623. 28, 61

[93] Timo Ojala, Matti Pietikäinen, and Topi Mäenpää. **Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24**(7):971–987, 2002. Available from: http://dx.doi.org/10.1109/TPAMI.2002.1017623. 28

[94] Daniel Onoro-Rubio and Roberto J López-Sastre. **Towards perspective-free object counting with deep learning**. In *European Conference on Computer Vision (ECCV)*, pages 615–629. Springer, 2016. Available from: http://dx.doi.org/10.1007/978-3-319-46478-7_38. 29, 30

[95] Vishal M. Patel, Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. **Visual Domain Adaptation: A survey of recent advances**. *IEEE Signal Processing Magazine*, **32**(3):53–69, 2015. Available from: http://dx.doi.org/10.1109/MSP.2014.2347059. 8, 19

[96] Sateesh Pedagadi, James Orwell, Sergio A. Velastin, and Boghos A. Boghossian. **Local Fisher Discriminant Analysis for Pedestrian Re-identification**. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3318–3325, 2013. Available from: http://dx.doi.org/10.1109/CVPR.2013.426. 15

[97] Luca Piras, Giorgio Giacinto, and Roberto Paredes. **Passive-aggressive Online Learning for Relevance Feedback in Content based Image Retrieval**. In *International Conference on Pattern Recognition Applications and Methods (ICPRAM)*, pages 182–187, 2013. Available from: https://doi.org/10.1016/10.5220/0004265401820187. 38

[98] Bryan James Prosser, Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. **Person Re-Identification by Support Vector Ranking**. In *British Machine Vision Conference, Proceedings (BMVC)*, pages 1–11, 2010. Available from: http://dx.doi.org/10.5244/C.24.21. 15

[99] Lorenzo Putzu, Luca Piras, and Giorgio Giacinto. **Ten Years of Relevance Score for Content Based Image Retrieval**. In *IAPR International Conference on Machine Learning and Data Mining in Pattern Recognition (MLDM)*, pages 117–131. Springer, 2018. Available from: http://dx.doi.org/10.1007/978-3-319-96133-0_9. 33, 37

[100] Lei Qi, Lei Wang, Jing Huo, Luping Zhou, Yinghuan Shi, and Yang Gao. **A Novel Unsupervised Camera-Aware Domain Adaptation Framework for Person Re-Identification**. In *IEEE International Conference on Computer Vision (ICCV)*, pages 8079–8088, 2019. Available from: http://dx.doi.org/10.1109/ICCV.2019.00817. 20

[101] Xuelin Qian, Yanwei Fu, Yu-Gang Jiang, Tao Xiang, and Xiangyang Xue. **Multi-scale Deep Learning Architectures for Person Re-identification**. In *IEEE International Conference on Computer Vision (ICCV)*, pages 5409–5418, 2017. Available from: http://dx.doi.org/10.1109/ICCV.2017.577. 16, 17, 18, 19

[102] Xuelin Qian, Yanwei Fu, Tao Xiang, Wenxuan Wang, Jie Qiu, Yang Wu, Yu-Gang Jiang, and Xiangyang Xue. **Pose-Normalized Image Generation for Person Re-identification**. In *European Conference on Computer Vision (ECCV)*, pages 661–678, 2018. Available from: http://dx.doi.org/10.1007/978-3-030-01240-3_40. 19

[103] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian processes for machine learning*. Adaptive computation and machine learning. MIT Press, 2006. 28

[104] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. **Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks**. In *Neural Information Processing Systems*, pages 91–99, 2015. 38, 39

[105] J. J. Rocchio. **Relevance Feedback in Information Retrieval**. In *The SMART Retrieval System: Experiments in Automatic Document Processing*, pages 313–323. 1971. 37

[106] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. **ImageNet Large Scale Visual Recognition Challenge**. *International Journal of Computer Vision (IJCV)*, **115**(3):211–252, 2015. Available from: http://dx.doi.org/10.1007/s11263-015-0816-y. 36

[107] Deepak Babu Sam, Neeraj N. Sajjan, R. Venkatesh Babu, and Mukundhan Srinivasan. **Divide and Grow: Capturing Huge Diversity in Crowd Images With Incrementally Growing CNN**. In *Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 3618–3626, 2018. Available from: http://dx.doi.org/10.1109/CVPR.2018.00381. 29

[108] Deepak Babu Sam, Neeraj N. Sajjan, Himanshu Maurya, and R. Venkatesh Babu. **Almost Unsupervised Learning for Dense Crowd Counting**. In *Association for the Advancement of Artificial Intelligence*, 2019. 27, 29

[109] Gregory Schröder, Tobias Senst, Erik Bochinski, and Thomas Sikora. **Optical Flow Dataset and Benchmark for Visual Crowd Analysis**. In *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, 2018. Available from: http://dx.doi.org/10.1109/AVSS.2018.8639113. 30, 56, 57

[110] William Robson Schwartz and Larry S. Davis. **Learning Discriminative Appearance-Based Models Using Partial Least Squares**. In *SIBGRAPI, Proceedings of the XXII Brazilian Symposium on Computer Graphics and Image Processing*, pages 322–329, 2009. Available from: http://dx.doi.org/10.1109/SIBGRAPI.2009.42. 15

[111] Chong Shang, Haizhou Ai, Zijie Zhuang, and Long Chen Rui Chen. **Improving Pedestrian Detection in Crowds With Synthetic Occlusion Images**. In *IEEE International Conference on Multimedia & Expo Workshops, ICME Workshops*, pages 1–4, 2018. Available from: http://dx.doi.org/10.1109/ICMEW.2018.8551575. 27, 30, 56, 57

[112] Yantao Shen, Tong Xiao, Hongsheng Li, Shuai Yi, and Xiaogang Wang. **End-to-End Deep Kronecker-Product Matching for Person Re-Identification**. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6886–6895, 2018. Available from: http://dx.doi.org/10.1109/CVPR.2018.00720. 16, 17, 18, 19

[113] Karen Simonyan and Andrew Zisserman. **Very Deep Convolutional Networks for Large-Scale Image Recognition**. In *International Conference on Learning Representations (ICLR)*, 2015. 26

[114] Vishwanath Sindagi and Vishal M. Patel. **A Survey of Recent Advances in CNN-based Single Image Crowd Counting and Density Estimation**. *Pattern Recognition Letters*, **107**:3–16, 2017. Available from: http://dx.doi.org/10.1016/j.patrec.2017.07.007. 8, 27, 28

[115] Vishwanath Sindagi and Vishal M. Patel. **Multi-Level Bottom-Top and Top-Bottom Feature Fusion for Crowd Counting**. In *IEEE/CVF*

International Conference on Computer Vision (ICCV), pages 1002–1012, 2019. Available from: http://dx.doi.org/10.1109/ICCV.2019.00109. 29

[116] Vishwanath A Sindagi and Vishal M Patel. **Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting**. In *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2017. Available from: http://dx.doi.org/10.1109/AVSS.2017.8078491. 26, 29, 62, 63

[117] Vishwanath A. Sindagi and Vishal M. Patel. **HA-CCN: Hierarchical Attention-Based Crowd Counting Network**. *IEEE Transaction on Image Processing*, **29**:323–335, 2020. Available from: http://dx.doi.org/10.1109/TIP.2019.2928634. 30

[118] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. **Mask-Guided Contrastive Attention Model for Person Re-Identification**. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1179–1188, 2018. Available from: http://dx.doi.org/10.1109/CVPR.2018.00129. 19

[119] Liangchen Song, Cheng Wang, Lefei Zhang, Bo Du, Qian Zhang, Chang Huang, and Xinggang Wang. **Unsupervised Domain Adaptive Re-Identification: Theory and Practice**. *CoRR*, **abs/1807.11334**, 2018. 8, 19

[120] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. **Pose-Driven Deep Convolutional Model for Person Re-identification**. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3980–3989, 2017. Available from: http://dx.doi.org/10.1109/ICCV.2017.427. 16, 18, 19

[121] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. **Part-Aligned Bilinear Representations for Person Re-identification**. In *European Conference on Computer Vision (ECCV)*, pages 418–437, 2018. Available from: http://dx.doi.org/10.1007/978-3-030-01264-9_25. 16, 18, 19

[122] Xiaoxiao Sun and Liang Zheng. **Dissecting Person Re-Identification From the Viewpoint of Viewpoint**. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 608–617, 2019. Available from: http://dx.doi.org/10.1109/CVPR.2019.00070. 27, 30, 56, 57

[123] Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang. **SVDNet for Pedestrian Retrieval**. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3820–3828, 2017. Available from: http://dx.doi.org/10.1109/ICCV.2017.410. 16, 18

[124] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. **Beyond Part Models: Person Retrieval with Refined Part Pooling (and A Strong Convolutional Baseline)**. In *European Conference on Computer Vision (ECCV)*, pages 501–518, 2018. Available from: http://dx.doi.org/10.1007/978-3-030-01225-0_30. 16, 18, 19

[125] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. **Going deeper with convolutions**. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015. Available from: http://dx.doi.org/10.1109/CVPR.2015.7298594. 17

[126] Haotian Tang, Yiru Zhao, and Hongtao Lu. **Unsupervised Person Re-Identification With Iterative Self-Supervised Domain Adaptation**. In *Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019. 20

[127] Antti Tarvainen and Harri Valpola. **Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results**. In *Neural Information Processing Systems (NIPS)*, pages 1195–1204, 2017. 36

[128] Maoqing Tian, Shuai Yi, Hongsheng Li, Shihua Li, Xuesen Zhang, Jianping Shi, Junjie Yan, and Xiaogang Wang. **Eliminating Background-Bias for Robust Person Re-Identification**. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5794–5803, 2018. Available from: http://dx.doi.org/10.1109/CVPR.2018.00607. 16, 19

[129] Peter Tu, Thomas Sebastian, Gianfranco Doretto, Nils Krahnstoever, Jens Rittscher, and Ting Yu. **Unified crowd segmentation**. In *European Conference on Computer Vision (ECCV)*, pages 691–704. Springer, 2008. Available from: http://dx.doi.org/10.1007/978-3-540-88693-8_51. 27

[130] Rahul Rama Varior, Bing Shuai, Jiwen Lu, Dong Xu, and Gang Wang. **A Siamese Long Short-Term Memory Architecture for Human Re-identification**. In *European Conference on Computer Vision (ECCV)*, pages 135–153, 2016. Available from: http://dx.doi.org/10.1007/978-3-319-46478-7_9. 16, 19

[131] Catherine Wah, Steve Branson, Pietro Perona, and Serge J. Belongie. **Multiclass recognition and part localization with humans in the loop**. In *International Conference on Computer Vision, ICCV*, pages 2524–2531, 2011. Available from: https://doi.org/10.1109/ICCV.2011.6126539. 21

[132] Cheng Wang, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. **Mancs: A Multi-task Attentional Network with Curriculum Sampling for Person Re-Identification**. In *European Conference on Computer Vision (ECCV)*, pages 384–400, 2018. Available from: http://dx.doi.org/10.1007/978-3-030-01225-0_23. 16, 18, 19

[133] Faqiang Wang, Wangmeng Zuo, Liang Lin, David Zhang, and Lei Zhang. **Joint Learning of Single-Image and Cross-Image Representations for Person Re-identification**. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1288–1296, 2016. Available from: http://dx.doi.org/10.1109/CVPR.2016.144. 16, 18

[134] Hanxiao Wang, Shaogang Gong, Xiatian Zhu, and Tao Xiang. **Human-in-the-Loop Person Re-identification**. In *European Conference on Computer Vision (ECCV)*, pages 405–422, 2016. Available from: http://dx.doi.org/10.1007/978-3-319-46493-0_25. v, 21, 23, 24, 33, 35, 37, 39, 40, 41, 53

[135] Qi Wang, Junyu Gao, Wei Lin, and Yuan Yuan. **Learning from synthetic data for crowd counting in the wild**. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8198–8207, 2019. Available from: http://dx.doi.org/10.1109/CVPR.2019.00839. 8, 27, 29, 30, 57, 60, 62, 63, 75

[136] Xiaogang Wang, Gianfranco Doretto, Thomas Sebastian, Jens Rittscher, and Peter H. Tu. **Shape and Appearance Context Modeling**. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1–8, 2007. Available from: http://dx.doi.org/10.1109/ICCV.2007.4409019. 15

[137] Yan Wang, Lequn Wang, Yurong You, Xu Zou, Vincent Chen, Serena Li, Gao Huang, Bharath Hariharan, and Kilian Q. Weinberger. **Resource Aware Person Re-Identification Across Multiple Resolutions**. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8042–8051, 2018. Available from: http://dx.doi.org/10.1109/CVPR.2018.00839. 16, 18, 19

[138] Yicheng Wang, Zhenzhong Chen, Feng Wu, and Gang Wang. **Person Re-Identification With Cascaded Pairwise Convolutions**. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1470–1478, 2018. Available from: http://dx.doi.org/10.1109/CVPR.2018.00159. 16

[139] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. **Person Transfer GAN to Bridge Domain Gap for Person Re-Identification**. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 79–88, 2018. Available from: http://dx.doi.org/10.1109/CVPR.2018.00016. 39

[140] Kilian Q. Weinberger, John Blitzer, and Lawrence K. Saul. **Distance Metric Learning for Large Margin Nearest Neighbor Classification**. In *Neural Information Processing Systems (NIPS)*, pages 1473–1480, 2005. 15

[141] Lin Wu, Chunhua Shen, and Anton van den Hengel. **PersonNet: Person Re-identification with Deep Convolutional Neural Networks**. *CoRR*, **abs/1601.07255**, 2016. Available from: http://arxiv.org/abs/1601.07255. 16, 19

[142] Qin Wu, Fangfang Yan, Zhilei Chai, and Guodong Guo. **Crowd counting by the dual-branch scale-aware network with ranking loss constraints**. *IET Computer Vision*, **14**(3):101–109, 2020. Available from: http://dx.doi.org/10.1049/iet-cvi.2019.0704. 29, 30, 62

[143] Xingjiao Wu, Yingbin Zheng, Hao Ye, Wenxin Hu, Jing Yang, and Liang He. **Adaptive Scenario Discovery for Crowd Counting**. In *IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP)*, pages 2382–2386, 2019. Available from: http://dx.doi.org/10.1109/ICASSP.2019.8683744. 29

[144] Xinyu Wu, Guoyuan Liang, Ka Keung Lee, and Yangsheng Xu. **Crowd density estimation using texture analysis and learning**. In *IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 214–219, 2006. Available from: http://dx.doi.org/10.1109/ROBIO.2006.340379. 28

[145] Fei Xiong, Mengran Gou, Octavia I. Camps, and Mario Sznaier. **Person Re-Identification Using Kernel-Based Metric Learning Methods**. In *European Conference on Computer Vision (ECCV)*, pages 1–16, 2014. Available from: http://dx.doi.org/10.1007/978-3-319-10584-0_1. 16

[146] Bin Xu, Jiajun Bu, Chun Chen, Deng Cai, Xiaofei He, Wei Liu, and Jiebo Luo. **Efficient manifold ranking for image retrieval**. In *International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 525–534, 2011. Available from: http://dx.doi.org/10.1145/2009916.2009988. 37

[147] Bin Xu, Jiajun Bu, Chun Chen, Can Wang, Deng Cai, and Xiaofei He. **EMR: A scalable graph-based ranking model for content-based image retrieval**. *IEEE Trans. on Knowledge and Data Eng.*, **27**(1):102–114, 2015. Available from: http://dx.doi.org/10.1109/TKDE.2013.70. 37

[148] Jiaolong Xu, David Vázquez, Krystian Mikolajczyk, and Antonio M. López. **Hierarchical online domain adaptation of deformable part-based models**. In *IEEE International Conference on Robotics and Automation (ICRA)*,

pages 5536–5541, 2016. Available from: http://dx.doi.org/10.1109/ICRA.2016.7487769. 32

[149] Biao Yang, Weiqin Zhan, Nan Wang, Xiaofeng Liu, and Jidong Lv. **Counting crowds using a scale-distribution-aware network and adaptive human-shaped kernel**. *Neurocomputing*, 2019. Available from: http://dx.doi.org/10.1016/j.neucom.2019.02.071. 29

[150] Wenjie Yang, Houjing Huang, Zhang Zhang, Xiaotang Chen, Kaiqi Huang, and Shu Zhang. **Towards Rich Feature Discovery With Class Activation Maps Augmentation for Person Re-Identification**. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1389–1398, 2019. Available from: http://dx.doi.org/10.1109/CVPR.2019.00148. 16, 18

[151] Yang Yang, Jimei Yang, Junjie Yan, Shengcai Liao, Dong Yi, and Stan Z. Li. **Salient Color Names for Person Re-identification**. In *European Conference on Computer Vision (ECCV)*, pages 536–551, 2014. Available from: http://dx.doi.org/10.1007/978-3-319-10590-1_35. 15

[152] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. **Deep Learning for Person Re-identification: A Survey and Outlook**. *CoRR*, **abs/2001.04193**, 2020. 7, 8, 9, 12

[153] Anran Zhang, Jiayi Shen, Zehao Xiao, Fan Zhu, Xiantong Zhen, Xianbin Cao, and Ling Shao. **Relational Attention Network for Crowd Counting**. In *IEEE International Conference on Computer Vision (ICCV)*, pages 6787–6796, 2019. Available from: http://dx.doi.org/10.1109/ICCV.2019.00689. 29, 30

[154] Cong Zhang, Hongsheng Li, Xiaogang Wang, and Xiaokang Yang. **Cross-scene crowd counting via deep convolutional neural networks**. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 833–841, 2015. Available from: http://dx.doi.org/10.1109/CVPR.2015.7298684. 8, 27, 30, 55

[155] Lu Zhang, Miaojing Shi, and Qiaobo Chen. **Crowd Counting via Scale-Adaptive Convolutional Neural Network**. In *IEEE Winter Conference on*

Applications of Computer Vision(WACV), pages 1113–1121, 2018. Available from: http://dx.doi.org/10.1109/WACV.2018.00127. 29

[156] Qi Zhang and Antoni B Chan. **Wide-Area Crowd Counting via Ground-Plane Density Maps and Multi-View Fusion CNNs**. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, page 8297–8306, 2019. Available from: http://dx.doi.org/10.1109/CVPR.2019.00849. 64

[157] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. **Single-image crowd counting via multi-column convolutional neural network**. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 589–597, 2016. Available from: http://dx.doi.org/10.1109/CVPR.2016.70. 29, 30, 61, 63, 64

[158] Youmei Zhang, Chunluan Zhou, Faliang Chang, and Alex C. Kot. **A scale adaptive network for crowd counting**. *Neurocomputing*, **362**:139–146, 2019. Available from: http://dx.doi.org/10.1016/j.neucom.2019.07.032. 29, 30

[159] Zhimeng Zhang, Jianan Wu, Xuan Zhang, and Chi Zhang. **Multi-Target, Multi-Camera Tracking by Hierarchical Clustering: Recent Progress on DukeMTMC Project**. *CoRR*, **abs/1712.09531**, 2017. 38

[160] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. **Deeply-Learned Part-Aligned Representations for Person Re-identification**. In *IEEE International Conference on Computer Vision, ICCV*, pages 3239–3248, 2017. Available from: http://dx.doi.org/10.1109/ICCV.2017.349. 19

[161] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. **Unsupervised Salience Learning for Person Re-identification**. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3586–3593, 2013. Available from: http://dx.doi.org/10.1109/CVPR.2013.460. 15

[162] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. **Scalable Person Re-identification: A Benchmark**. In *International Conference on Computer Vision (ICCV)*, pages 1116–1124, 2015. Available from: http://dx.doi.org/10.1109/ICCV.2015.133. 38

[163] Liang Zheng, Yi Yang, and Alexander G. Hauptmann. **Person Re-identification: Past, Present and Future**. *CoRR*, **abs/1610.02984**, 2016. 7, 9, 10, 15, 16, 33

[164] Meng Zheng, Srikrishna Karanam, Ziyan Wu, and Richard J. Radke. **Re-Identification With Consistent Attentive Siamese Networks**. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5735–5744, 2019. Available from: http://dx.doi.org/10.1109/CVPR.2019.00588. 16, 17, 19

[165] Zhedong Zheng, Liang Zheng, and Yi Yang. **Unlabeled Samples Generated by GAN Improve the Person Re-identification Baseline in Vitro**. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3774–3782, 2017. Available from: http://dx.doi.org/10.1109/ICCV.2017.405. 19

[166] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. **Invariance Matters: Exemplar Memory for Domain Adaptive Person Re-Identification**. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 598–607, 2019. Available from: http://dx.doi.org/10.1109/CVPR.2019.00069. 19, 36

[167] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. **Camera Style Adaptation for Person Re-Identification**. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5157–5166, 2018. Available from: http://dx.doi.org/10.1109/CVPR.2018.00541. 20

[168] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. **Omni-Scale Feature Learning for Person Re-Identification**. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3701–3711, 2019. Available from: http://dx.doi.org/10.1109/ICCV.2019.00380. 16, 19

[169] Sanping Zhou, Fei Wang, Zeyi Huang, and Jinjun Wang. **Discriminative Feature Learning With Consistent Attention Regularization for Person Re-Identification**. In *IEEE International Conference on Computer Vision (ICCV)*, pages 8039–8048, 2019. Available from: http://dx.doi.org/10.1109/ICCV.2019.00813. 16, 18, 19

[170] Xiaofeng Zhu, Xuelong Li, Shichao Zhang, Chunhua Ju, and Xindong Wu. **Robust Joint Graph Sparse Coding for Unsupervised Spectral Feature Selection**. *IEEE Transactions on Neural Networks and Learning Systems*, **28**(6):1263–1275, 2017. Available from: http://dx.doi.org/10.1109/TNNLS.2016.2521602. 21

[171] Zhikang Zou, Xinxing Su, Xiaoye Qu, and Pan Zhou. **DA-net: Learning the fine-grained density distribution with deformation aggregation network**. *IEEE Access*, **6**:60745–60756, 2018. Available from: http://dx.doi.org/10.1109/ACCESS.2018.2875495. 29, 62, 63