ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2023, Vol. 84, No. 5, pp. 558–593.
Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, 2023.
Russian Text © The Author(s), 2023, published in Avtomatika i Telemekhanika, 2023, No. 5, pp. 61–112.

Person Re-identification in Video Surveillance Systems Using Deep Learning: Analysis of the Existing Methods

H. Chen[*,**,a], **S. A. Ihnatsyeva**[***,b], **R. P. Bohush**[***,c], **and S. V. Ablameyko**[****,d]

[*] *Zhejiang Shuren University, Hangzhou, Zhejiang, China*
[**] *International Science and Technology Cooperation Base of Zhejiang Province: Remote Sensing Image Processing and Application, Hangzhou, 310000 China*
[***] *Euphrosyne Polotskaya State University of Polotsk, Polotsk, Belarus*
[****] *Belarusian State University, Minsk, Belarus*
e-mail: [a] *eric.hf.chen@hotman.com,* [b] *s.ignatieva@psu.by,* [c] *r.bogush@psu.by,* [d] *ablameyko@bsu.by*

Received May 11, 2022
Revised December 10, 2022
Accepted January 26, 2023

Abstract—This paper is devoted to a multifaceted analysis of person re-identification (ReID) in video surveillance systems and modern solution methods using deep learning. The general principles and application of convolutional neural networks for this problem are considered. A classification of person ReID systems is proposed. The existing datasets for training deep neural architectures are studied and approaches to increasing the number of images in databases are described. Approaches to forming human image features are considered. The backbone models of convolutional neural network architectures used for person ReID are analyzed and their modifications as well as training methods are presented. The effectiveness of person ReID is examined on different datasets. Finally, the effectiveness of the existing approaches is estimated in different metrics and the corresponding results are given.

Keywords: person re-identification, video data, convolutional neural networks, accuracy estimation metrics, image descriptors

DOI: 10.25728/arcRAS.2023.58.18.001

1. INTRODUCTION

The wide implementation of video surveillance systems allows solving many practical tasks, in particular, increasing the level of public safety. For example, it is important to determine the presence of a given person by his or her video data images in another place or at different times in spatially distributed video surveillance systems. This problem is called person re-identification (ReID). To solve it, one needs to identify distinctive features and compare them with the features from an available sample of images for a set of persons (gallery) through a query to the database. Note that the composition of features, to a large extent, determines the efficiency of person ReID. The search and extraction of the most distinctive features of the objects in the images, including people, are not formalized. Consequently, an empirical approach is used, representing in most cases a long and time-consuming process. Unreasonably high computational cost is required to re-identify persons due to the ambiguous appearance from different angles, variations in lighting, different camera resolutions, and occlusions. Therefore, significant progress has not been achieved for person ReID for a long time. Advances in computing tools and discoveries in deep learning, particularly the development of convolutional neural networks (CNNs), allowed automating the extraction of human image features and improving considerably the accuracy of person ReID. However, despite that many researchers and engineers in the world are working on this problem using deep learning methods, it is not completely solved so far. The development of person ReID systems still faces a lot of challenges and a rich variety of applications, such as pass systems at secure enterprises, search for missing people or offenders, and collection of statistical information about the visitors of malls and other social objects, lead to numerous approaches and algorithms with different qualitative characteristics.

2. ORGANIZING AND ESTIMATING THE EFFECTIVENESS OF PERSON RE-ID IN DISTRIBUTED VIDEO SURVEILLANCE SYSTEMS

2.1. The General Scheme of a Person Re-ID System

A spatially distributed video surveillance system consists of geographically dispersed IP cameras and usually has a single data center. Figure 1 shows the simplified person ReID structure in such a system, which includes three IP cameras, C_1, C_2, and C_3. In each frame F^k, where k denotes the video camera number, a detector finds all persons in the field of view of cameras and forms bounding boxes for them. The human images I_i, where $i = 1, \ldots, N_{img}$ and N_{img} denotes the total number of images, are placed in the gallery. For each of them, the vectors f_i^{gen} (CNN descriptors) are determined using CNNs to form a common CNN feature space $\chi_{Ii} = \{f_i^{gen}\}$. This space is represented as a table, with each row being a CNN descriptor f_i^{gen} for one image.

A composite feature vector is used to describe a person during ReID. This vector can be written as

$$P_{ID} = (p_n^{ID}, f_i^{gen}, f_i^{add}) \tag{1}$$

with the following notations: p_n^{ID} is the identifier (label) of the person; n is the number of possible identifiers equal to the total number of unique persons; f_i^{gen} is the CNN feature vector for the

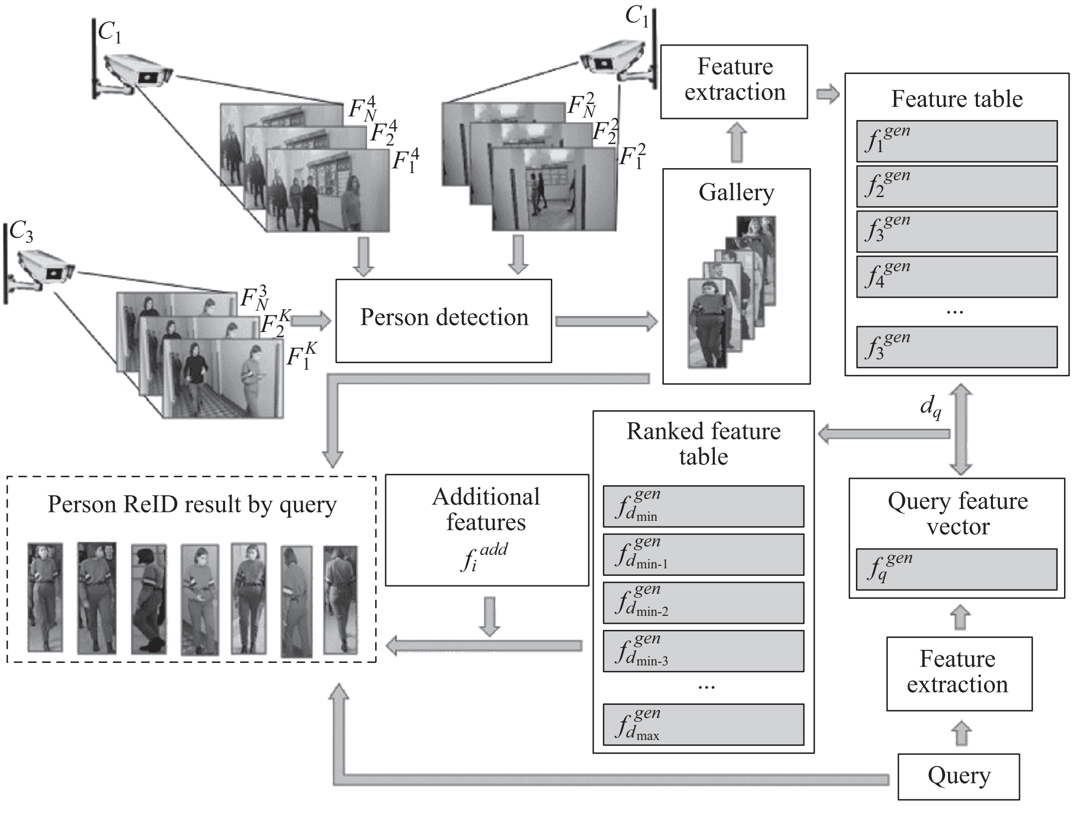

Fig. 1. The general scheme of a person ReID system.

ith image of the person (it may include global features f_i^{global}, characterizing the entire image, and the local ones $f_{i,j}^{local}$, obtained by dividing the image into j parts); finally, f_i^{add} is additional features (they may contain information to improve the effectiveness of the person ReID system, such as the identifier C_{ID}, the frame number from the $k$the video camera F_m^k, etc. [1]).

When a person ReID query arrives, its feature vector f_q^{gen} is calculated and then used to find the distance d_q determining the degree of similarity between this query and the descriptors of the gallery images. The resulting distances are used in the table $\chi_{I i}$ for ranking from $d_{\min}$ to $d_{\max}$. Based on additional features, candidate images are excluded when they do not match the sought person by some criteria even despite the similarity of visual characteristics. For example, imagine that an object of interest with similar visual features is simultaneously in the images from two non-overlapping cameras. In this case, the matter unambiguously concerns different persons: the same person cannot be present in two places at the same time. After excluding all unsuitable candidates, the images of persons whose f_i^{gen} are at the top of the ranked table are displayed as the intermediate results. The first person from this list is assumed the ReID result as the most similar one to the query.

2.2. Classification of Person Re-ID Systems

The wide area of application of person ReID systems leads to numerous algorithms and approaches to solve the problem and, accordingly, various classifications of such systems (Fig. 2). For example, depending on the interaction with the environment, it is possible to distinguish Close-world person ReID systems (with ready datasets for training and testing) and Open-world ones (with an image gallery constantly replenished with new frames) [2]. Close-world systems are typically used for research purposes and the dataset consists of a limited number of video sequences or images taken from multiple surveillance cameras. The data in such sets are annotated and prepared in advance and the query is present in the gallery. Open-world systems involve a dataset that changes over time with the arrival of new video recordings from surveillance cameras; in this case, bounding boxes need to be generated in real time. The new images have to be annotated by forming pseudo-labels for the ability to train CNNs during video surveillance. The organization of such systems is much more complicated. They require high-performance hardware but are closest to the real conditions.

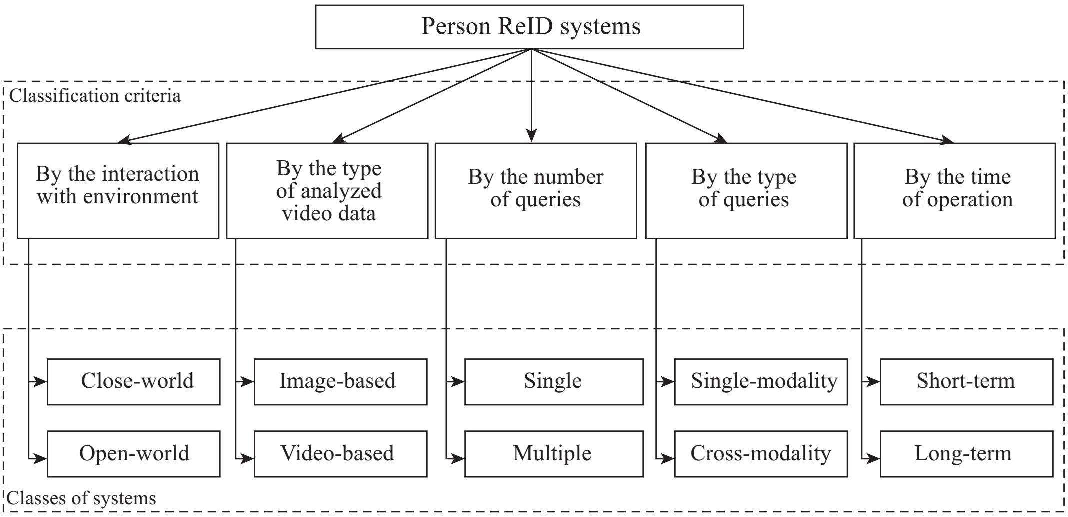

Fig. 2. Classification of person ReID systems.

Depending on the type of analyzed video data, there are static (image-based) and dynamic (video-based) person ReID systems. The former process individual frames in certain intervals, whereas the latter consider a sequence of video frames [3]. In video-based systems, features are formed by spatial domain analysis in combination with consideration of the person's temporal component (e.g., information about the gait, direction of movement, and other additional features).

Depending on the number of queries [4], person ReID systems can be divided into single (for one person) and multiple (for all persons in the field of view of cameras). In the first case, a person has to be found in the dataset by query and ReID comes to searching or checking whether the person sought is present in the gallery. In the second case, a unique identifier has to be established for each person and it is determined in which frames each of these persons occurs again; this problem comes to classification [5].

Depending on the type of queries, experts single out homogeneous (single-modality) and heterogeneous (cross-modality) person ReID systems [2]. For queries with homogeneous data, images or video from visible-range surveillance cameras are used. If a query contains a text description of a person sought, an image from an infrared camera, or a picture or a sketch, such systems are called cross-modality.

Depending on the time of operation, there are short- and long-term person ReID systems [6]. For example, if each person in images of the dataset is in the same clothes, the changes in appearance are insignificant (e.g., up to accessories or things in hands), and shooting was performed during a limited time interval (the person could not significantly change the image), then such system is short-term. Long-term ReID focuses on the ability to re-identify people even if a significant amount of time has already passed (the person may have changed appearance) [7].

Any of the systems discussed above can encounter the domain shift problem when training and testing are performed on data from different domains. A domain is understood as a set of images that have been acquired under the same conditions in the same surveillance system. Each image in the dataset is influenced by a combination of factors including camera resolution, background, lighting conditions, and even the appearance of people. (For example, statistically, Europeans differ from Asians, summer clothes differs from winter clothes, etc.) A system trained on a dataset from indoor surveillance cameras may have extremely low performance on a test sample consisting of human images from outdoor surveillance cameras. Algorithms aimed at solving this problem are called Cross-domain person ReID. They ensure domain adaptation or shiftability.

2.3. Accuracy Estimation Metrics

The choice of metrics is one of the most important issues for estimating person ReID results. Metrics allow numerically evaluating the effectiveness of algorithms and comparing the results for different person ReID approaches. RankN is the most common group of metrics, including Rank1, Rank5, Rank10, and mAP. This group characterizes the quality of ranking and shows the percentage of queries with the correct result among the first N results. Rank1 shows the percentage of queries for which the identifier of the first candidate image coincides with the query identifier. For N = 5, Rank5 shows the percentage of queries with the correct solution among the first five candidate images; in Rank10, the first ten candidate images are considered. RankN is calculated by dividing the sum of the number of queries with the correct solution among the first results by the total number of queries Q:

$$\mathrm{RankN} = \frac{\sum K_{i,\mathrm{N}}}{Q}, \tag{2}$$

where i denotes the query number and $K_{i,\mathrm{N}}$ is the ith query with the correct solution among the first N results.

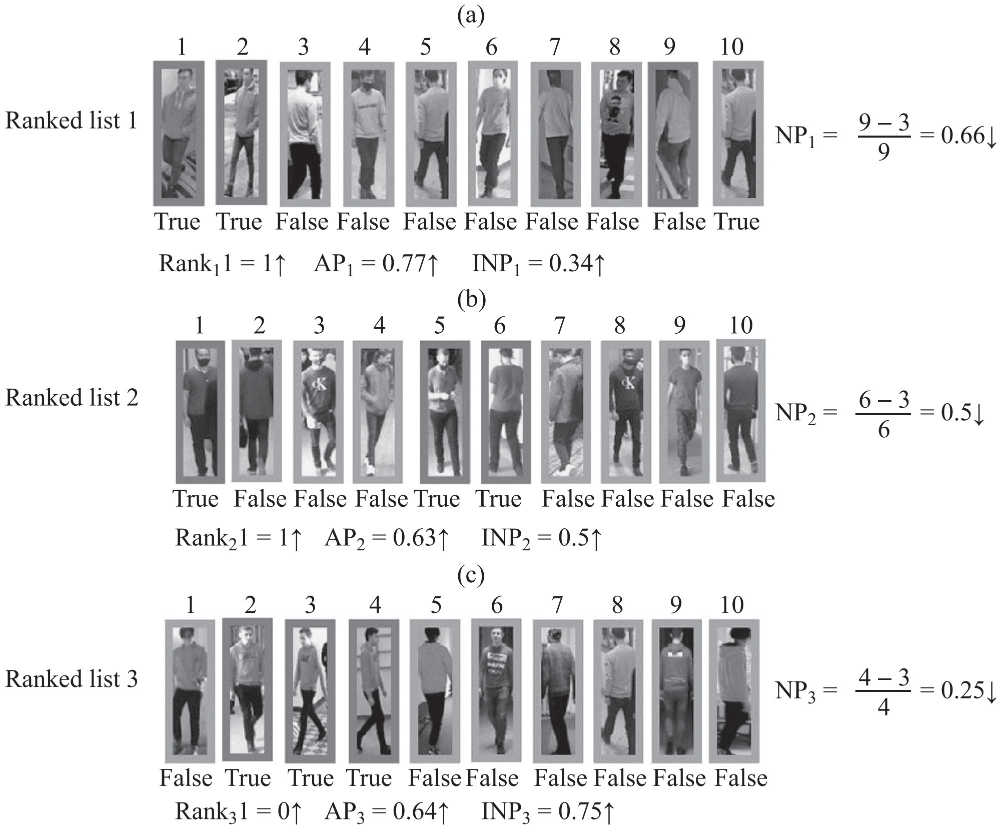
(a)
1 2 3 4 5 6 7 8 9 10
Ranked list 1
NP₁ = (9 – 3)/9 = 0.66↓
True True False False False False False False False True
Rank₁1 = 1↑ AP₁ = 0.77↑ INP₁ = 0.34↑
(b)
1 2 3 4 5 6 7 8 9 10
Ranked list 2
NP₂ = (6 – 3)/6 = 0.5↓
True False False False True True False False False False
Rank₂1 = 1↑ AP₂ = 0.63↑ INP₂ = 0.5↑
(c)
1 2 3 4 5 6 7 8 9 10
Ranked list 3
NP₃ = (4 – 3)/4 = 0.25↓
False True True True False False False False False False
Rank₃1 = 0↑ AP₃ = 0.64↑ INP₃ = 0.75↑

Fig. 3. The difference in Rank, AP, NP, and INP depending on the position of true and false predictions.

mAP is an accuracy estimate for person ReID algorithms that reflects the mean of the average precisions for all queries. It has the form

$$\mathrm{mAP} = \frac{1}{Q}\sum_{i=1}^{Q} AP_i, \tag{3}$$

where AP is the average precision (the domain below the *precision–recall* curve). Here, $\mathit{precision} = \frac{TP}{TP+FP}$, TP is the number of true positive query predictions, FP is the number of false positive query predictions, $\mathit{recall} = \frac{TP}{TP+FN}$ is sensitivity, and FN is the number of false negative query predictions.

In person ReID systems, a priority is to place true predictions at the beginning of the ranked list with as few false predictions as possible. Note that RankN and mAP do not reflect the difficulty of finding correctly identified human images for an incoming query. In addition, with the same Rank metrics, the accuracy of AP may differ. The mINP (mean Inverse Negative Penalty) metric, originally proposed in [2], was used in [8] to find the hardest correct predictions. This metric excludes the dominance of easy matches that affect Rank and mAP. It is calculated by introducing additional metrics: NP (Negative Penalty) and INP (Inverse Negative Penalty). NP is imposed for wrong predictions for the ith query; it reduces the probability of correct person ReID when wrongly finding the hardest match. INP is the inverse of NP; its growth indicates an increase in system efficiency. mINP characterizes the mean of INP values for all queries:

$$\mathrm{mINP} = \frac{1}{Q}\sum_{i}(1 - \mathrm{NP}_i) = \frac{1}{Q}\sum_{i}\left(1 - \frac{R_i^{hard} - |G_i|}{R_i^{hard}}\right) = \frac{1}{Q}\sum_{i}\frac{|G_i|}{R_i^{hard}}, \tag{4}$$

where $\text{NP}_i = \frac{R_i^{hard} - |G_i|}{R_i^{hard}}$ is the negative penalty, R_i^{hard} is the position of the hardest true prediction, and $|G_i|$ is the total number of true predictions for the query.

Figure 3 shows an example with only three true images in the gallery for each query. For the first two ranked lists in Fig. 3, the AP metrics are different under the same Rank1 value: $\text{AP}_1 = 0.77$ (Fig. 3a) and $AP_2 = 0.63$ (Fig. 3b). The reason is that the first ranked list contains two correct matches at the beginning whereas the second list only one. Moreover, the nearest correct match occupies the fifth position. A direct comparison of the lists in Figs. 3a and 3c indicates that $\text{AP}_3 = 0.64$ in the third ranked list (i.e., more than in the second one) but Rank1 in this example is 0. This is also explained by that all possible correct answers were obtained at the top of the ranked table (the second, third, and fourth positions), except for the first, incorrect, prediction. All correctly identified human images are preferable to obtain as early as possible. However, system evaluation based on the AP and Rank metrics will not determine this with maximum accuracy.

Analysis of Fig. 3c shows that only the first four candidate images need to be obtained to have all possible correct answers. In this case, the negative penalty will be $\text{NP} = 0.25$, the minimum value for the examples in Fig. 3. The hardest prediction in Fig. 3b corresponds to the sixth position in the ranked table; the wrong person detection in Fig. 3a is characteristic of the ninth position. Therefore, for the examples in Figs. 3a and 3b, the NP values increase and the INP values decrease accordingly. Thus, INP allows estimating the influence of the difficulty of finding all correct matches. The higher this value is, the better the system will search for all persons with the same identifier. Hence, it is necessary to decrease NP and reduce the number of positions from the beginning of the ranking list to the hardest one, which can be misidentified when searching for an image.

3. DATASETS AND THEIR PREPARATION FOR CNN TRAINING

Using CNNs for feature extraction requires training the deep neural network model. For this purpose, one commonly adopts an annotated dataset containing a unique identifier for each person $S = \{(I_i, p_1^{ID}), \ldots, (I_m, p_n^{ID})\}$, where I_i is the image, $1 \leqslant i \leqslant m$, m denotes the number of images, and p_n^{ID} is the person identifier. Often images are accompanied by information about the source camera number and the frame number in the video sequence. For the effective operation of the system, it is necessary to extract from the annotated dataset a feature vector $f^{gen}(I_i)$ such that in the entire feature space χ_{I_i}, the distance between them for the same identifiers will be smaller than for persons with different labels. In other words, one has to reduce as much as possible the error E of identity prediction in S:

$$\min E(I_i, p_n^{ID}) \in [p_n^{ID} - g(f^{gen}(I_i))], \tag{5}$$

where g is a classifier. The quality of the extracted features depends on the distribution and diversity of the data in S [9].

When training CNNs, common recommendations to improve the accuracy of person ReID are as follows: select optimal hyperparameters, such as learning rate, batch size, and the number of epochs; increase the training sample; use data augmentation, find the most effective loss function and CNN architecture, or divide the image into fragments.

For an already trained model, the algorithm can be improved by selecting the most effective way to rank the feature table, using re-ranking, considering additional information about the time and place of video surveillance and its attributes. Attributes are semantic information about a person that is important for his or her identification. They include the color and type of clothes, the length of a person's hair, and the presence and peculiarities of a bag, backpack, glasses, or other significant details.

3.1. Analysis of Datasets

The accuracy of person ReID is significantly affected by the size and composition of the training sample. However, the person ReID algorithm largely determines the requirements for the dataset. Forming an image bank for training and testing is a laborious and time-consuming process. In addition, there exists the domain shift problem [10, 11], i.e., a significant decrease in the accuracy of person ReID when using the system in conditions stylistically different from the training sample. A partial solution of this problem is to combine different datasets, see the discussion in [12, 13], particularly for the required domain [12, 14].

With the existing datasets used for CNN training, in addition to the domain shift problem, one faces the problem of protecting personal data. Some image databases are closed: their developers provide only extracted features for research [15]. Other datasets can be used with restrictions [16–18], i.e., authors are asked to respect the privacy of the students' images when publishing their research; the distribution of such image databases is possible only under agreement with the authors. The ability to use some datasets is restricted. For example, MSMT17 [19] is no more publicly available, and DukeMTMC-ReID [20] has been withdrawn and is not recommended for use [21].

The existing image sets differ in the number of video scenes and persons as well as the number of images for each person. Such databases may contain individual frames in their entirety (e.g., PRW [22] and CUHK-SYSU [23]) or rectangular fragments cut out from these frames based on bounding boxes (only the image of a person). Some datasets include sets of bounding boxes obtained from several consecutive frames, called tracklets (e.g., MARS [24] and LPW [25]). Also, there may be bounding boxes obtained from individual frames with some time interval (e.g., Market-1501 [26], CUHN01 [16], CUHN02 [17], CUHN03 [18], VIPer [27], and others).

As a rule, the images for datasets were provided in different outdoor (Market-1501 [26], LPW [25], and PRID [28]) or indoor (QMUL iLIDS [29] and Airport [30]) conditions. When forming the PolReID [31] image database, 856 indoor and outdoor scenes were used. In the CUHN01 dataset, images for each person were obtained from two cameras with nonoverlapping fields of view. CUHN02 involved five such camera pairs; in CUHN03, images were generated by six cameras, with bounding boxes from only two cameras for each person. The VIPeR dataset was formed from the images of two outdoor cameras, with only one image from each camera for each person. Three different locations were used to create LPW: the first location with three cameras and the other two with four cameras. The PRW, Market-1501, and MARS datasets were generated at the same location near a supermarket in Tsinghua University from six cameras and differ only in the data presentation: entire frames, bounding boxes with human images, and tracklets, respectively.

Cross-modality person ReID systems are trained and tested using special datasets with queries in the form of text (CUHK-PEDES [32] and ICFG-PEDES [33]), low resolution images (LR-PRID [34] and LR-VIPeR [35]), infrared camera images (SYSU-MM01 [36] and RegDB [37]), or sketches (PKU-Sketch [38]).

The CUHK-PEDES [32] dataset combines five others, namely, CUHK03 [18], Market-1501 [26], SSM [39], VIPeR [27], and CUHK01 [16]; each image is annotated with two text descriptions in English. A text description consists of 23.5 words on average and contains information about the person's appearance, actions, and poses. ICFG-PEDES [33], another dataset for cross-modality person ReID systems, contains on average 37.2 words with a more detailed description of appearance than CUHK-PEDES and is based on MSMT17 [19].

The LR-PRID [34] and LR-VIPeR [35] datasets were obtained using PRID [28] and VIPeR [27], respectively; for each person, they have a pair of images, one with low resolution and the other with high resolution, to be used in person ReID systems with video cameras of different resolutions.

Table 1. The comparative characteristics of datasets for person ReID systems

Dataset	The number of cameras	The number of persons	The number of bounding boxes	Image size
PRW [22]	6	932	34 304	Different
CUHK-SYSU [23]	6	8432	96 143	–
MARS [24]	6	1261	1 191 003	256×128
LPW [25]	3, 4, 4	2731	592 438	256×128
Market-1501 [26]	6	1501	32 217	128×64
CUHN01 [16]	2	971	3884	160×60
CUHN02 [17]	10 (5 pairs)	1816	7264	160×60
CUHN03 [18]	6	1360	13 164	Different
MSMT17 [19]	15	4101	126 441	Different
VIPeR [27]	2	632	1264	128×48
PRID [28]	2	934	24 541	128×64
QMUL iLIDS [29]	2	119	476	Different
Airport [30]	6	9651	39 902	128×64
PolReID [31]	856	657	52 035	Different
CUHK-PEDES [32]	–	13 003	80 412	–
ICFG-PEDES [33]	–	4102	52 522	–
LR-PRID [34]	2	100	200	–
LR-VIPeR [35]	2	632	1264	128×48 and 64×24
SYSU-MM01 [36]	6	491	38 271	–
RegDB [37]	2	412	8240	–
PKU-Sketch [38]	2	200	400	–

SYSU-MM01 [36] was obtained from two infrared and four RGB cameras. It consists of 15 712 infrared images and 22 559 color images for 491 persons. The RegDB dataset [37] contains 10 color images, taken during the day, and 10 thermal images from the night IR camera for 412 persons each. They can be used in cross-modality person ReID systems with infrared and RGB video cameras.

In [38], a dataset for two hundred persons was proposed, including two images from different cameras and a sketch for each person. The sketches were created with the participation of volunteers, who described the appearance of people to five different painters to train an open-world cross-modality person ReID system. In the case of no photograph of a person, a sketch drawn from the description is used.

MPR Drone [40] is another dataset for open-world person ReID systems, where one flying drone video camera captured images. This dataset consists of two parts: the first part is labeled for 113 610 detected bounding boxes, whereas the second part contains the raw frames for the first part.

LUPerson, a large unlabeled dataset, was presented in [41]. It includes over four million images for two hundred thousand persons and can be used for the unsupervised training of person ReID systems. It was generated from video data provided by over seventy thousand street videos from various cities.

Table 1 shows the comparative characteristics of the datasets considered.

Since explicit consent of all participants is required to create a dataset, some researchers use generated images to form a training sample. MOTSynth, a synthetic person ReID dataset, was proposed in [42]. It was created using video sequences from Grand Theft Auto V (GTA-V), a popular game simulating a city with inhabitants in 3D space. The authors manually labeled camera viewpoints, planned the routes and movements of pedestrians, and established the parameters of typical human behavior in crowded places. In total, 597 different pedestrian models were analyzed with randomly changing clothes, backpacks, bags, masks, hairstyles, and beards. In total, over 9519 unique pedestrians were obtained. According to the authors' results, training on the synthetic set improves ReID accuracy by 6.9% in the mAP metric compared to training based on Market-1501 [26] and by 2.5% in the mAP metric compared to training on the combined dataset from Market-1501 [26] and CUHK03 [18].

The paper [9] considered an algorithm for generating synthetic images to improve the system's stability to domain shift. MakeHuman [43] was used to create 3D realistic images of people and the Unreal Engine 4 (UE4) platform [44] was employed for video surveillance modeling with the ability to adjust the shooting conditions (night, indoor, and outdoor), the number of person occlusions, and walking speed. A large number of appearance details were engaged, such as masks, glasses, headphones, and hats. The resulting images of people contain real fragments of clothes, which distinguishes this approach from the existing ones. During generation, persons with similar appearances and small distinctive features were intentionally added. According to [9], this set yields a greater accuracy in the Rank1 metric in cross-domain testing using MSMT17 compared to other synthetic image databases such as SOMAset [45], SyRI [46], PersonX [47], and RandPerson [48]. These results were confirmed in the course of testing on Market-1501 and DukeMTMC-ReID.

In [49], the synthetic ClonedPerson dataset containing 3D images of people was proposed. Note that the clothes of all generated characters was cloned from real images to strengthen the similarity between a virtual person and his or her prototype. In total, the dataset includes 887 766 images for 5621 persons. The Unity3D platform [50] was used to generate the images, as for RandPerson [48]. This dataset was employed to train the CNN, achieving better testing results on images from a different domain in the mAP metric on CUHK03 [18], Market-1501 [26], and MSMT17 [19] compared to training with RandPerson [48] and UnrealPerson [9]. A significant advantage of synthetic datasets is the automatic generation of annotations.

3.2. Augmentation of the Training Sample

Augmentation is increasing the size of a training sample by modifying the images contained in it. Traditional approaches involve various transformations of images, such as rotation, reflection, resizing, changing contrast and brightness, color variations, and blurring. To improve stability to occlusions, random erasing [51] can be used: a rectangular fragment of the image, with randomly chosen size and shape, is filled with zeros or random values (Fig. 4). This augmentation method was tested for person ReID on the Market-1501, DukeMTMC-ReID, and CUHK03 datasets. According to the analysis results, in some cases (e.g., testing on CUHK03), random erasing improves accuracy by almost 9% in the Rank1 metric and by over 6% in the mAP metric. For Market-1501 and DukeMTMC-ReID with different algorithms, the accuracy in the Rank1 and mAP metrics is increased by 1–4%.

Note that in ReID algorithms, data augmentation is used to increase the training sample by randomly selecting an image for some transformation, but the mechanism of this impact is not considered. (In other words, supposedly, it positively affects the accuracy of the trained model through improving the generalization ability of the network.) In [52], rotations were used to increase the number of images, but the CNN was trained simultaneously for the original image and the transformed one; the losses due to the rotation were estimated and the RMS error between the feature vectors for the corresponding pair of images was minimized. Figure 5 compares the basic data augmentation algorithm in feature extraction training and the one proposed in [52]. The

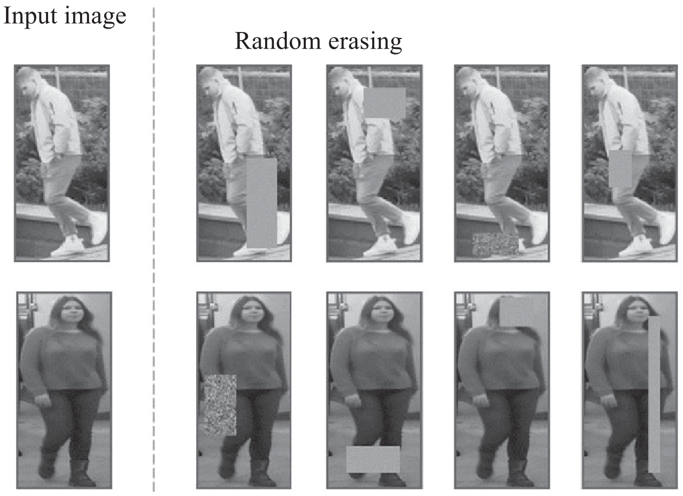

Fig. 4. Random erasing for an image fragment to augment data: some examples.

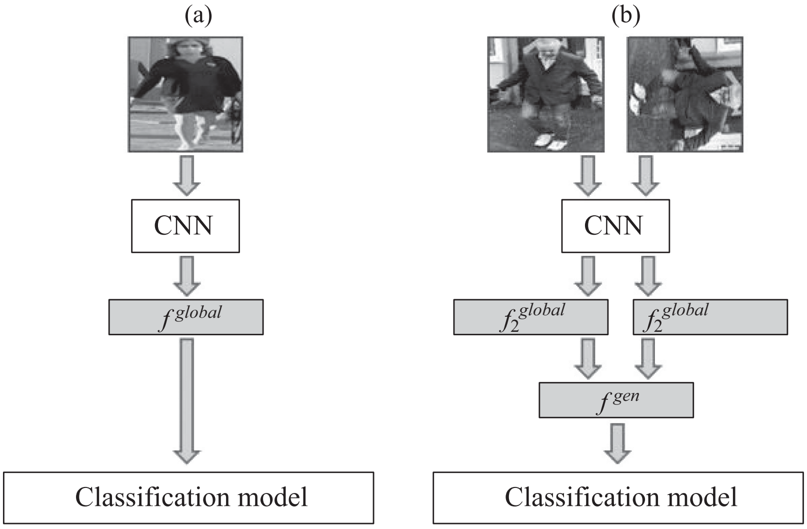

Fig. 5. The principle of feature extraction: (a) basic and (b) with image rotation [52].

basic algorithm involves one rotation for a random sample in one pass through the network during training. In the algorithm [52], each image is rotated by a random angle and supplied to the network input simultaneously with the original image. The CNN extracts features from a pair of images, and the features are then averaged. Compared to the basic counterpart, this algorithm improves the accuracy in the mAP metric: over 5% for Market-1501, over 10% for DukeMTMC-reID, and over 20% for MSMT-17 (Fig. 5a). Moreover, the maximum ReID accuracy in the mAP and Rank1 metrics, mAP = 81.3 and Rank1 = 87.5, were achieved for MSMT-17 at the issue time of [52]. The algorithm proposed in [53] increases the metric values to mAP = 84.4 and Rank1 = 89.9 for the MSMT-17 dataset.

A more sophisticated data augmentation method uses Generative Adversarial networks (GANs), which generate near-natural images from existing data. A generative adversarial network is a machine learning algorithm based on a combination of two neural networks: one generates images, and the other tries to determine their genuineness. For person ReID, GANs can be applied to improve the ability to extract effective features [54] or solve domain shift problems [55].

The authors [54] considered a typical person ReID problem arising in real conditions: various factors (exogenous perturbations) may deteriorate the quality of images obtained from surveillance cameras. For example, if it is raining at the time of surveillance, the system trained on other conditions-based data will not be able to interpret the extracted descriptors with high accuracy. In such cases, a large number of generated features will consider with a high probability the similarities of perturbations instead of the similarities of people. To solve this problem, it is necessary to study the features of different phenomena affecting image quality. However, in real conditions, it is difficult to obtain annotations to describe exogenous perturbations, and the training sample may have no reference examples. The images stable to the quality-deteriorating factors were extracted using GANs: these networks served to synthesize images with a preset degree of distortion.

In [55], GAN was applied for data augmentation. In contrast to similar systems, the authors proposed adding to the training sample only the images increasing the accuracy of person ReID. For this purpose, images with similar features as those previously obtained are discarded, as they can reduce the quality of training, increase the time, and even cause an imbalance during the generalization. As a result, the system assumes that the features extracted for similar images are more significant than those with insufficiently many examples. This problem was settled using the Local Outlier Factor (LOF) method. It controls the number of similar generated images and discards some of them randomly if their number increases. This approach improves the accuracy of person ReID and, furthermore, significantly enhances the system's stability to domain shift. Finally, it was compared with other algorithms for solving the domain shift problem; according to the results [55], LOF ranks first and second for different datasets among the current approaches in the Rank1, Rank5, and mAP accuracy metrics.

The paper [56] considered an approach to generate additional images of people when the number of images from one camera in a video surveillance system exceeds that from another camera, or the view from another camera is missing for a particular person. This approach was used to improve the robustness of algorithms when pairs of images from different cameras are needed for the same person. However, such patterns were generated not as images but in the feature space. The reason is that the generation of images requires a significantly higher computational cost of the generative model for the qualitative formation of background and illumination. However, this does not necessarily have a positive effect on the ReID model, while the generation of only features neglects the peculiarities of the entire image of the scene captured.

4. ANALYSIS OF THE FEATURES USED

Person ReID with CNNs involves the following features: global features (Fig. 6a), i.e., the ones formed for the entire human image; local features, when the image is divided into separate fragments (Fig. 6b); key points (Fig. 6c), which imply a separate feature vector for each part of the image; additional features (Fig. 6d), which include auxiliary annotations, information about the time and place of shooting, and attributes; personal features from the sequence of frames (Fig. 6e).

4.1. Global Features

In person ReID, the use of global features is the basic approach. They are applied together with local [3] or additional [57] features to increase the accuracy of person ReID or in algorithms where the effectiveness of ReID is improved by their acquisition [58] or processing [52].

When using global features, a ReID system may appear insufficiently stable to occlusions due to the fact that in the generated feature vector for a hidden image, part of the descriptors will characterize not the appearance of the person but the object overlapping him or her. In addition, such an approach may loose the features of small distinctive details of appearance, e.g., glasses and

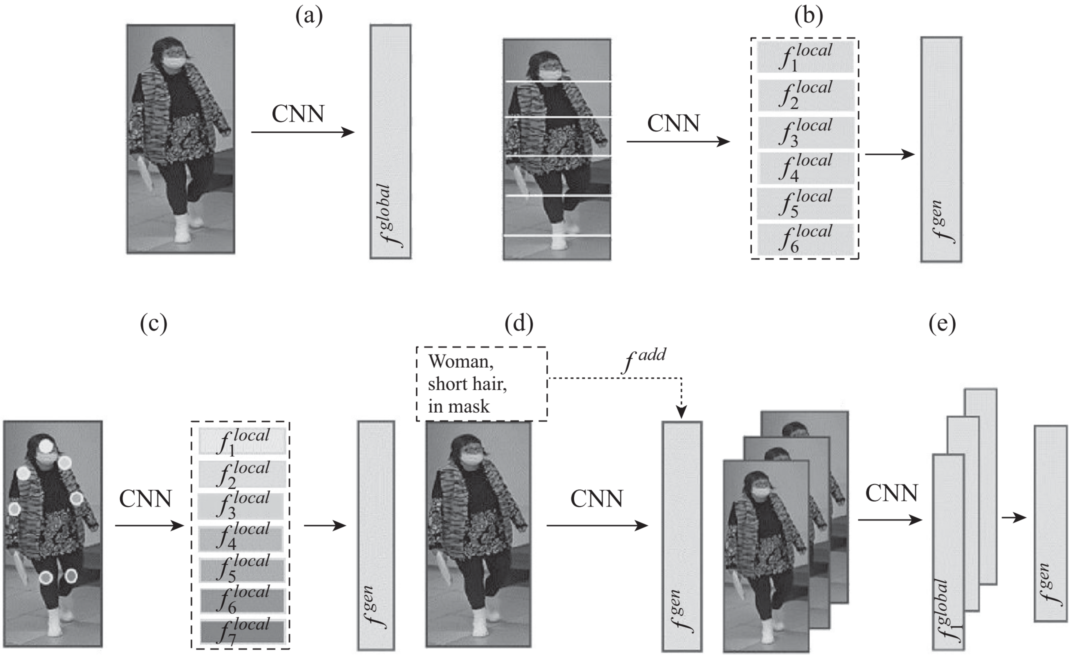

Fig. 6. Strategies to learn and use features.

clothes fittings or bags, which could be a determinative characteristic when deciding whether the detected person matches the query or not.

4.2. Local Features

Local features allow reducing the disadvantages of global features. They can be considered both independently and in conjunction with the global ones. For example, the authors [59] proposed a horizontal division of the image into six equal parts to be studied separately. This approach, called Part-based Convolutional Baseline (PCB), is a superstructure over CNNs, where the output data of the first convolutional layer are divided into parts. It improves the accuracy of person ReID by 1–2% in the Rank1 and mAP metrics. Its drawback consists in the requirement for the location and content of each part: the person must be in a strictly vertical position and the image fragments must be located in appropriate places. Detection errors, when part of the person is clipped by the bounding box, may cause ReID errors.

The paper [60] studied the effect of the number of image fragments on the accuracy of person ReID. The image was divided into two, three, four, six, eight, and twelve fragments, and the best ReID accuracy results in the Rank1 and mAP metrics were obtained for the six-parts division.

A ReID algorithm based on the key parts of the human body was presented in [61]. In this algorithm, key points are extracted using HR-Net [62], and then the features in the vicinity of each of them are examined. This approach aims to reduce the impact of occlusions. Therefore, when comparing the feature vectors, the descriptors of the hidden key points are neglected.

The algorithm proposed in [63] requires dividing the image of a human figure into 6 horizontal parts. The network tries to predict whether each of them has a visible part of the human figure. If the network decides positively, the key points of the person are determined using the AlphaPose estimator [64], and the features of the invisible parts are not considered when predicting whether the detected person is the desired one. This improves the accuracy of person ReID and increases the system's stability to occlusions.

4.3. Additional Features

Another approach to increase the accuracy of person ReID involves additional information provided with the dataset in the form of annotations. Such an approach was proposed in [57]: the visual features of objects are extracted using CNNs (DenseNet-121, ResNet-50, or PCB) and the camera and frame numbers are contained in the file names. After ranking the table of visual features, the descriptors of images irrelevant by the spatiotemporal characteristics of people (i.e., for those who could not physically present in a certain place or at a certain hour) are removed from it.

In most cases, ReID algorithms decide on the similarity or difference of the query and the images of people in the gallery using unobvious types of features. The authors [65] introduced an approach to identifying and visualizing the features that the system considered to decide, which ones were significant, and what contribution each feature made. The corresponding method, called Attribute-guided Metric Distillation (AMD), is an interpreter connected to the target model to evaluate the contribution of each feature and visualize the most significant details. The interpreter learns to separate the distance between the features of different people based on the attributes, and a loss function is introduced to focus on characteristic differences. According to the experiments, it is possible to visualize significant features and, furthermore, improve the accuracy of person ReID in the target models. In addition, the results presented indicate a higher ReID accuracy when testing the algorithm on cross-domain data.

In [66], it was proposed to improve the stability of person ReID systems to domain shift. As a rule, such systems assume the presence of a source domain (for training) and a target domain (for testing). They are supposed to be isolated from each other. The intermediate domains used in [66] as additional information allow reducing the difference between the source and target domains. Images of both the target and source domains are supplied to the input of the backbone CNN and descriptors are generated from them; then, they are combined with different mixing relations to produce a feature vector of the intermediate domain. The technique proposed in [67] was used for this purpose. When combining image descriptors from different domains, a side effect arises: the features of images of different people are mixed and an image of a new person is generated. As a result, the network may focus on a person with mixed descriptors during training instead of considering the diversity of styles across domains. This phenomenon was compensated using an additional module based on the AdaIN style transfer approach [68], which generates descriptors of the same person considering the features of the target or source domain. The generated features of the intermediate domains were used to train CNN and reduce the distance between the extracted descriptors from the source and target domains.

The researchers [69] applied the Wi-Fi technology to settle different ReID problems such as changes in lighting, occlusions, background noise, and possible changes in appearance. This technology allows counting and localizing people. The person detection procedure involves variations of Wi-Fi signals, which inform about the presence of a person. They can be monitored using the channel state information (CSI) of access points. Significant features are extracted from the Wi-Fi signal and then serve to form a radio-biometric signature for person ReID.

In [70], Label distribution learning (LDL) was proposed as additional information to improve the accuracy of person ReID in invisible domains. Here, multiple datasets are used to train the CNN, and the process aims at finding the relationship between the images of different people. Each person is treated as a separate class, and the search for matches between different classes from different datasets allows extracting domain-invariant features. Particular attention is paid to similar people from different domains to generate a descriptor characterizing the appearance of the person rather than the conditions of video surveillance. To reduce the gap between data from different domains, labels (identifiers) of images for training are assigned to consider inter-domain relations rather than the domain of the class itself.

In [71], information about the perspective of a person was used as additional features and the features associated with the viewing angle were considered during ReID. Within this approach, the CNN determines one of the three perspectives (front, side, and rear views), which improves the system's stability to domain shift.

4.4. Features That Use Temporal Peculiarities

Video-based person ReID algorithms take advantage of the temporal component of a video sequence has, as opposed to the analysis of individual frames [60]. The algorithm proposed in [3] combines both global and local features in a person's image to improve ReID accuracy on video. At different levels of the pyramid (Fig. 7), the image is separated by vertical or horizontal lines and a feature vector is extracted for each image fragment. For each ith person, the total feature vector [9] is defined as

$$f_i^{gen} = \left[f_i^{global}; f_{i,v}^{local-vertical}; f_{i,h}^{local-horizontal}; f_{i,patch}^{local-patch}\right], \tag{6}$$

where $v, h, patch$ is the number of image parts at each pyramid level.

For a sequence of K video frames, the feature vector of each person is given by

$$\overline{f}_i^{gen} = \left[\sum_{k=1}^{K} f_{i,k}^{global}; \sum_{k=1}^{K} f_{i,v,k}^{local-vertical}; \sum_{k=1}^{K} f_{i,h,k}^{local-horizontal}; \sum_{k=1}^{K} f_{i,patch,k}^{local-patch}\right]. \tag{7}$$

The authors [72] proposed extracting gait information for silhouettes of people using the background subtraction method. Despite that color images contain more information than the image of a person's figure, silhouette analysis allows identifying features characteristic of different people when they move. Within the approach described in [72], the first stage is to remove the background as well as the brightness and color differences of the person from the video, resulting in the image of the person's figure. After background subtraction, bounding boxes are generated for all people in every fifth frame of the video, and linear interpolation is used to calculate the remaining bounding boxes. The extracted silhouettes are normalized similarly to the method proposed in [73], and the top and bottom of the figure are considered in the first stage; then, the cumulative sum of pixels on the X axis with respect to the center of this object is analyzed. After that, all the images are reduced to the same size with preserving the aspect ratio and a height of 224 pixels. According to [72], for gait-based ReID, it is necessary to form Gait Energy Images (GEIs) reflecting the characteristic features of a person when walking by the analysis of the frame sequence. To form

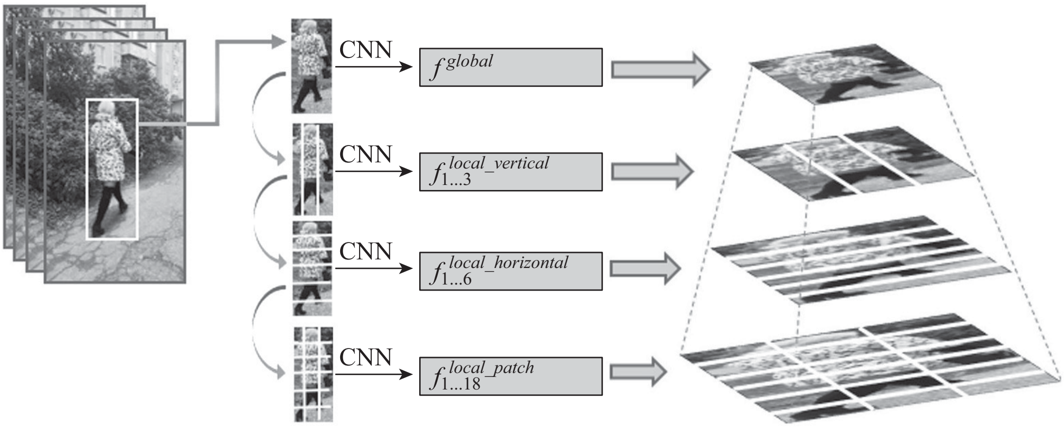
CNN
f^{global}
CNN
$f_{1...3}^{local\_vertical}$
CNN
$f_{1...6}^{local\_horizontal}$
CNN
$f_{1...18}^{local\_patch}$

Fig. 7. Global and local feature extraction based on part-aggregated images and a multilevel pyramid.

GEIs, a trajectory of movement is determined using the central coordinates of the bounding boxes. The resulting curvilinear trajectory of human movements is divided into several rectilinear sections using a piecewise regression algorithm. For each section, the clustering algorithm of k-means is applied to the corresponding frame sequence, and GEIs are formed.

The paper [74] considered the following person ReID approach: a 3D convolution operation combining the visual and temporal components is applied to certain consecutive frames to take into account changes in appearance during movement. In addition, a special SSN architecture is used to extract the features of individual body parts and divide descriptors into groups considering moving and static body parts in the video.

In [75], it was proposed to select the most effective spatiotemporal features based on the analysis of global and local descriptors for video sequences. Global features were constructed using the Relation-Based Global Feature Learning Module (RGL), which generates correlation maps of descriptors between frames to find the most significant ones. Local features were synthesized using the Relation-Based Partial Feature Learning Module (RPL), which determines the relationship between the features of the same fragment on different frames.

The approach proposed in [76] includes two modules for the more effective use of temporal information in video. The first one, Key Frame Screening with Index (KFSI), searches for similar frames and selects the most informative ones for person ReID to train the CNN. The second module, Feature Reorganization Based on Inter-Frame Relation (FRBIFR), is intended to identify the most significant features of people by analyzing their location on the frame sequence. As a result, the impact of noise factors (e.g., the overlapped images of people) is reduced.

4.5. Key Domain Features

To increase stability to background noise and changes in object's features during motion, several researchers proposed to search and select domains using attention modules (also called attention models or attention mechanisms) [77]. In [78], local and global analyses were performed for this purpose, and the Relation aware global attention (RGA) module was presented. This module covers the structural information of the entire image and studies fragmentary peculiarities. Finding the key domains allows localizing significant distinctive features. To find them, each descriptor is compared pairwise with all other descriptors, and the result is included in the total feature vector. Such an approach allows considering the relationship of global and local differences in the images of people.

The attention mechanism was used in the temporal domain [79]; the papers [60, 80] considered its application in the spatiotemporal domain. In [81], this mechanism was used in the spatial-local domain to assess human poses and predicting visible parts. The authors [82] proposed a self-attention mechanism to enhance the generalizability of CNNs by considering the relationship between features.

The pyramid multi-part features with multi-attention (PMP-MA) module for feature extraction was described in [60]. The features obtained in this way cover important peculiarities with different degrees of detail. According to [60], the accuracy of this module is Rank5 = 99.3% on the iLIDS-vid and DukeMTMC-VideoReID datasets and even Rank5 = 100% on the PRID dataset.

The idea suggested in [83] is adding attention modules between ResNet blocks to improve the extractability of features from video frames. As an image passes through the CNN, some significant information may be lost, but the generated feature vector will contain redundant information for person ReID. Therefore, in [83] it was proposed to embed spatial attention modules at different ResNet levels. Output feature maps from certain levels of the CNN are combined and form a descriptor for each individual frame of the video sequence. The attention module averages the values of the received feature maps and constructs the resulting vector.

4.6. Metrics for Determining the Distance between Features

To find a person's image x_p in a gallery $G = \{g_i | i = 1, \ \ldots, N\}$ of N images, it is necessary to calculate distances between the feature vectors of the pth query and images g_i. In this stage, the following metrics are most widespread:

1. The cosine distance [57, 14], given by

$$d(p, g_i) = \frac{x_p x_{g_i}}{\|x_p\| \|x_{g_i}\|}. \tag{8}$$

2. The Euclidean distance [7, 10, 13, 26, 84], given by

$$d(p, g_i) = \|x_p - x_{g_i}\|_2^2. \tag{9}$$

3. The Mahalanobis distance [85], given by

$$d(p, g_i) = \sqrt{(x_p - x_{g_i})^T M^{-1} (x_p - x_{g_i})}, \tag{10}$$

where M denotes the covariance matrix.

4. The Jaccard distance for k-nearest neighbors [85], given by

$$d(p, g_i) = 1 - \frac{|R^*(p, k) \cap R^*(g_i, k)|}{|R^*(p, k) \cup R^*(g_i, k)|}, \tag{11}$$

where $R^*(p, k)$ and $R^*(g_i, k)$ are the sets of nearest neighbors.

Note that some algorithms involve re-ranking after the first sorting to refine the result (and improve ReID accuracy). In [85], the Mahalanobis distance was used for initial sorting. The first k images were selected from the resulting table and included in $R(p, k)$, and then re-ranking was performed based on the Jaccard distance.

The following re-ranking approach was adopted in [26]. First, feature vectors are sorted based on the Euclidean distance. Then, during re-ranking, k-first results are selected from the table $S(p, g)$ and, for each of them, a match is searched in the gallery. This procedure yields new lists $S(r_i, g)$ with the weights $\frac{1}{i+1}$, where $i = 1, \ldots, k$. The final feature table has the form

$$S^*(p, g) = S(p, g) + \sum_{i=1}^{k} \frac{1}{i+1} S(r_i, g). \tag{12}$$

The authors [84] proposed considering the context information of descriptor ranking when training CNNs together with the features for person ReID. The corresponding algorithm uses a two-streamed architecture consisting of external and internal streams. The first stream performs sorting for each query to find the most effective visual differences at the top of the ranked gallery list and form a preliminary set for further processing. The second stream analyzes local features for the result of the previous step. This approach is supposed to create a hybrid ranking for matching people with a better person ReID accuracy than other methods with list post-processing. Note that other metrics can be used to estimate the similarity of features [86], but their effectiveness requires additional research.

5. CNNS DESCRIBING HUMAN IMAGES: MODELS AND TRAINING

5.1. Backbone CNNs

Nowadays, the most widespread backbone CNNs for feature extraction in person ReID are ResNet-50 [87] in the papers [12, 65, 88] and DenseNet-121 [89] in the papers [7, 57] as well as

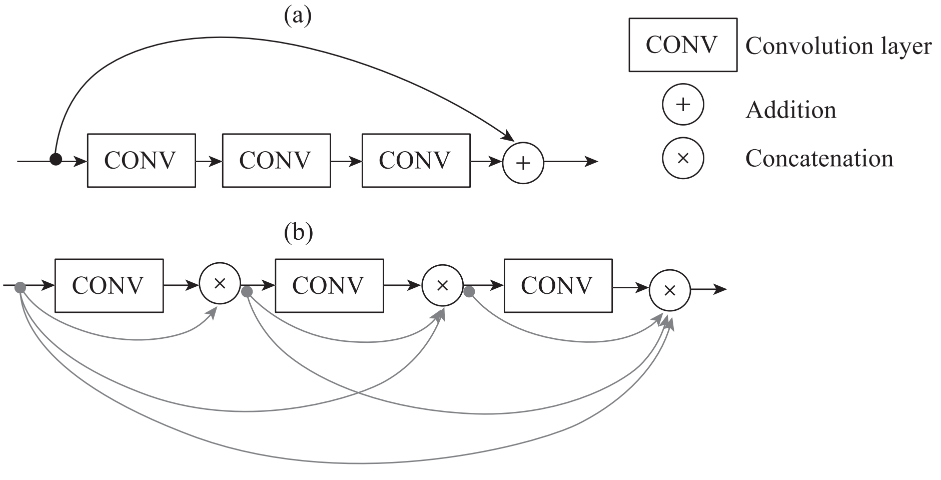

Fig. 8. The block diagrams of DenseNet and ResNet.

MobileNetV2 [88, 90], PCB [57, 84], GoogleNet [91], or original CNN architectures, such as [92]. The approach proposed in [93] improves the system's stability to occlusions. Person ReID is performed using the human head image, and the YOLOv3 CNN is used to detect the bounding boxes.

Architectures of the ResNet family are characterized by residual blocks (Fig. 8a), which use skip connections to reduce the probability of vanishing gradients during training. A residual block consists of two branches, one containing convolutional layers and the other passing information to the output without changes. At the output, the data from both branches are summed. During backpropagation learning, this approach prevents zeroing the gradients in the CNN.

The DenseNet-121 architecture (Fig. 8b) is characterized by connections between layers, in which the feature maps of all previous layers are used as input for all subsequent ones in the block. In contrast to ResNet, the feature maps are not summed up from layer to layer but concatenated. Some researchers compare the performance of their algorithms using different types of architectures as the backbone CNN for feature extraction. For example, the performance of ResNet-50 and DenseNet-121 was compared in [7]; according to the results, accuracy in the Rank1 and mAP metrics increases when using DenseNet-121. In [65], ResNet-34, ResNet-50, and ResNet-101 were studied for person ReID; as was established, deeper networks have a positive effect on ReID accuracy. The effectiveness of PCB [59] as a superstructure for ResNet-50 and DenseNet-121 was estimated in [57]. The experiments demonstrated the best values of the Rank1 and mAP metrics for PCB (Rank1 = 94.0 and mAP = 82.8); the intermediate position was occupied by DenseNet-121 (Rank1 = 90.8 and mAP = 76.9); the lowest metric values were observed for ResNet-50 (Rank1 = 87.7 and mAP = 72.2).

The authors [92] proposed a new CNN architecture for person ReID, the sparse graph wavelet convolution neural network (SGWCNN), based on the analysis of frame sequence features to consider the semantic relationship between local fragments of people in the video. This approach allows extracting additional information through the spatiotemporal analysis of video data. As expected, the proposed neural network will refine regional features for solving the problem of short-term occlusions in pedestrian movements effectively.

Note that the performance of a CNN is largely determined by the hyperparameters of its training: the number of epochs, the learning rate, and the batch size of the image.

The number of epochs determines how many times each image from the training sample passes through the network. If the values of this parameter are small, the model will not be fully trained and, consequently, ReID accuracy will be low. Too many epochs may cause overtraining, i.e., the network will remember all images considered and will not be able to process even the test cases effectively. For person ReID, the CNN is usually trained during 60–100 epochs. As a rule, batches of 16–64 images are supplied to the network input. Following the intention to parallelize the calculations, one often increases batch size, as it reduces the time to train the CNN, but the resulting accuracy of the trained neural network goes down. According to the approach presented in [94], batch size is gradually increased during CNN training, which allows minimizing accuracy drop while decreasing the training time. The paper [60] provided the most comprehensive study of the batch size effect on the accuracy of CNN training for person ReID. As was demonstrated therein, the highest accuracy can be achieved for a batch of 32 images on the DukeMTMC-VideoReID, MARS, iLIDS-vid, and PRID datasets.

As is well known, the learning rate shows how the weights change during each update. For person ReID, CNNs are trained using special schedulers to change the learning rate after a certain period or according to certain criteria. In [95], the ADEL rate reduction mechanism was considered. It monitors the values of the network weights; each time they stop changing in jumps, the learning rate is reduced. This approach ensures faster convergence in the CNN.

The researchers [96] introduced an approach involving three variation modes of the learning rate η depending on the curvature λ_0 of the loss function surface. The first mode assumes a lazy phase, in which the learning rate has a relatively small value $\eta < \frac{2}{\lambda_0}$, and the weights change with an almost constant step in the first learning stage. The second mode is characterized by a catapult phase, in which the learning rate takes values $\frac{2}{\lambda_0} < \eta < \eta_{\max}$. In this stage, there is an exponential increase in losses and a rapid decrease in the curvature of η until it stabilizes at a value $\lambda_{final} < \frac{\eta}{2}$. If this condition holds, a flat minimum is reached. The divergent phase is executed in the third mode. In this case, the learning rate exceeds the value $\eta_{\max}$ and the model stops learning. In [96], the following assumption was formulated and then confirmed by experiments: high learning rates allow finding flat minima, which generalize better than sharp minima. According to the viewpoint of the cited authors, the use of small training batches leads to the same outcome.

5.2. Modifications of CNNs

Changes in the backbone architectures provide opportunities to enhance the accuracy of person ReID systems. In [88], the impact of the data normalization method at the output of convolutional layers was investigated, and Meta Batch-Instance Normalization (MetaBIN) was proposed. This technology combines two approaches: batch normalization and individual image normalization [97]. The former approach provides information about different styles of images in a batch. However, this may decrease the accuracy of person ReID in invisible domains. The second approach ignores information about domain features, but the disadvantage is the possible reduction of useful information. The two problems are solved by introducing a trained parameter to balance between the approaches, increasing not only the effectiveness of person ReID but also system's stability when working in another domain. The paper [98] considered the impact of the activation function in ResNet-50, DenseNet-121, and DarkNet-53 CNNs on the accuracy of person ReID. ReLU [99], the most common activation function, is piecewise linear:

$$\phi(x) = \begin{cases} x, & x > 0, \\ 0, & x \leqslant 0, \end{cases} \tag{13}$$

where x denotes the input value of the neuron.

The main advantage is the low computational complexity of both feedforward and backpropagation (the forward and backward passes through the network). However, the values of the derivative on the positive (negative) part of the definitional domain of the activation function may cause exploding gradients in training (the loss of some information in training, respectively, since all neurons with negative values will not be activated). To avoid this, it is possible to apply the Leaky-ReLU [100] function

$$\phi(x) = \begin{cases} x, & x > 0, \\ \alpha x, & x \leqslant 0, \end{cases} \tag{14}$$

where α is the angular coefficient taking small values; traditionally, $\alpha = 0.01$.

The authors [101] empirically studied the influence of the slope angle of the negative part of the function in image classification using the ReLU and Leaky-ReLU activation functions and their modifications: Parametric Rectified Linear Unit (PReLU) and Randomized Leaky Rectified Linear Unit (RReLU). According to the experiments, the best results are obtained using PReLU. However, in this case, the CNN is overtrained with a high probability for a small dataset, so RReLU appears to be more effective in practice.

In addition to these modifications, the ELU, SeLU, and GeLU activation functions have a slight slope in the negative part of the definitional domain. Hence, they can be effective for person ReID.

The ELU (Exponential Linear Unit) [102] activation function is given by

$$\phi(x) = \begin{cases} x, & x \geqslant 0, \\ \alpha(e^x - 1), & x < 0, \end{cases} \tag{15}$$

where the coefficient $\alpha > 0$ restricts the output values in the negative part of the definitional domain.

The SELU (Scaled Exponential Linear Unit) activation function is a scaled version of ELU:

$$\phi(x) = \lambda \begin{cases} x, & x \geqslant 0, \\ \alpha(e^x - 1), & x < 0. \end{cases} \tag{16}$$

In [103], the coefficients were calculated as $\alpha = 1.67326$ and $\lambda = 1.0507$.

The GELU (Gaussian Error Linear Units) [104] activation function is given by

$$\phi(x) = \frac{1}{2}x\left[1 + erf\left(\frac{x}{\sqrt{2}}\right)\right] \approx 0.5x\left(1 + \tan\left(\sqrt{\frac{2}{\pi}}(x + 0.044715x^3)\right)\right) \tag{17}$$

or

$$\phi(x) = x\sigma(1.702x), \tag{18}$$

where $\sigma = \frac{1}{1+e^{-x}}$ denotes the sigmoid activation function.

In [105], an automatic generation approach was adopted to find the most effective function as follows: unary and binary functions were sequentially enumerated and combined alternately, and the result was assessed empirically. The resulting Swish function is given by

$$\phi(x) = x\sigma(\beta x), \tag{19}$$

where the coefficient β regulates the curvature of the function and σ is the sigmoid function.

The Mish activation function was considered in [106]:

$$\phi(x) = x\tanh(softplus(x)) = x\tanh(ln(1 + e^x)). \tag{20}$$

The activation function affects both the training dynamics and the accuracy of the trained model. It follows from [98] that using GeLU, Swish, and Mish instead of ReLU can improve the accuracy of person ReID. According to additional studies, these functions increase the training time of the model without sufficiently stable results. GeLU and ReLU are preferable activation functions for person ReID CNNs.

A new architecture of CNNs, MCLNet (Modality Confusion Learning Network), was proposed to solve specific problems, e.g., in cross-modality person ReID systems [8] with images from infrared and visible cameras. MCLNet is based on a partially partitioned two-stream network. To improve the CNN stability to heterogeneous data, the features specific to each data type are sequentially extracted separately, followed by common descriptors. Visible and infrared patterns have different feature distributions and cannot be matched for comparison. Hence, the network is trained to ignore modality information and tries to extract common distinctive features for heterogeneous human images. To avoid missing significant features of different people, a learning entanglement mechanism is created. As a result, the mismatch between heterogeneous images is minimized whereas their similarity is maximized. The paper [7] proposed the CNN architecture called RCSANet (Clothing Status Awareness Network) for long-term person ReID. The corresponding methods take into account that after some period, the person changes clothes and falls into the camera's field of view again. However, such approaches are inefficient if the person remains in the same clothes, and the accuracy of long-term person ReID systems goes down significantly. RCSANet [7] arranges the features of pedestrians and includes the features of clothes in the common descriptor. RCSANet is a two-stream system based on DenseNet-121 and contains the ICE (Inter-Class Enforcement) stream to maximize the differences for each person and the ICR (Intra-Class appearance regularization) stream to arrange the features obtained in ICE considering clothes change information. For the test sample without clothes change, the proposed approach yielded Rank1 = 100% and mAP = 97.2%; in the presence of people in different clothes, the metrics values were Rank1 = 48.6% and mAP = 50.2%.

5.3. Siamese Networks

A Siamese neural network is an architecture containing two or more identical subnetworks with identical architectures, parameters, and weights. Such a network outputs the similarity index of the two images supplied to its input [107].

Siamese networks have paired models (Fig. 9a), consisting of two subnetworks [108, 69], and triplet models [91], which include three subnetworks (Fig. 9b).

In [108], a Siamese architecture was used to minimize the cosine distance between the features of two instances in contrastive learning to detect their similarity. In [69], a Siamese principle-based deep neural network with two branches was employed to process the amplitude and phase of Wi-Fi signals in order to extract significant features of a radio-biometric signature for person ReID.

The authors [109] applied Siamese networks to prevent overtraining and proposed an architecture with two Siamese networks. The first network is basic and receives the positive or negative pairs of images of people as the input data. Note that a positive pair consists of the images obtained for the same person at different times whereas a negative pair is formed by the images of two different people. The features extracted by each branch of the basic Siamese network are supplied to the inputs of another network used to extract deeper features. Each of the two Siamese networks predicts whether the input pair is images of the same person or not. A verification loss function is introduced to adjust the relative distance between the feature vectors yielded by each of the Siamese networks for people with the same or different identifiers, thereby improving the accuracy of person ReID.

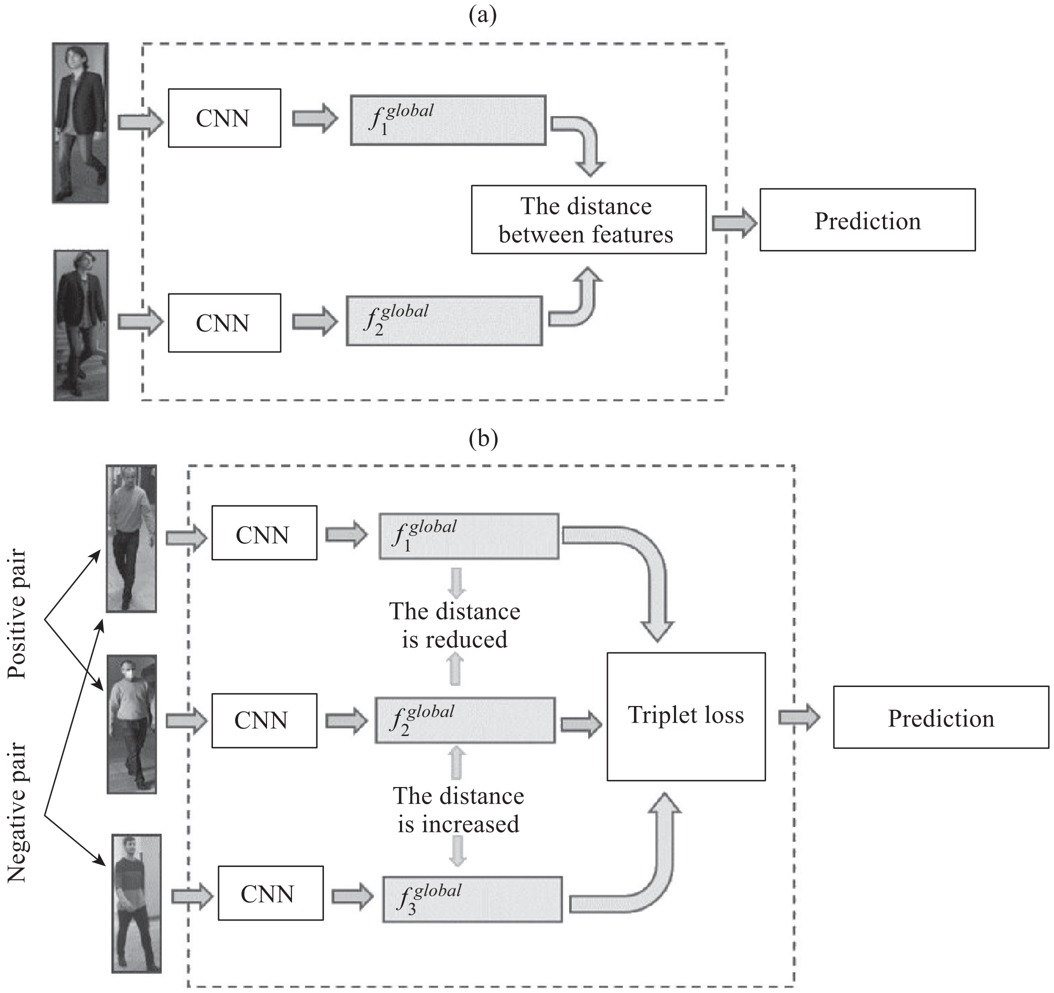

Fig. 9. Siamese neural network models: (a) the paired model and (b) the triplet model.

In [110], a deep person ReID architecture with an attention module in the Siamese network structure was suggested. This approach ensures the consistency of important appearance details from different frames and allows finding more significant distinctive features for different people. In addition, localizing distinctive features in an image is implemented during training; hence, the system can find key domains automatically.

The authors [91] presented a triplet-loss Siamese network with GoogleNet as the backbone subnetwork. In this architecture, human features are extracted from different levels of the network and are then combined to form a common descriptor map for each input image. With triplet loss, the network approaches the positive pairs of images in the feature space and distances from the negative ones.

5.4. Training CNNs

Generally speaking, training a neural network for effective feature extraction means finding weights in order to reduce the value of a loss function L. This function reflects the difference between the obtained result and the expected one. For person ReID, cross-entropy loss [11, 14, 42, 57] and triplet loss [7, 63, 111, 112] functions are common.

The cross-entropy loss function allows treating person ReID as a classification problem and is used after the softmax layer [113]. For a set $\{I_{ni}\}_{n_i=1,\ldots,n}$ of n training images that contains n_{id}

different people (classes) with corresponding identification labels $\{p_{n_i}^{ID}\}_{n_i=1,\ldots,n}$, $\{p_{n_i}^{ID}\} \in [1, \ldots, n_{id}]$, the cross-entropy loss [113] is given by

$$L_i = -\sum_{k=1}^{n_{id}} \{p_{n_i}^{ID} = k\} \log \frac{e^{\hat{p}_{n_i}^{ID_i}}}{\sum\limits_{l=1}^{n_{id}} e^{\hat{p}_{n_i}^{ID_i}}}, \tag{21}$$

where $\hat{p_{n_i}}^{ID_i}$ denotes the predicted value.

A peculiarity of triplet losses is the analysis of two pairs of images: in a positive pair, the images belong to the same person ($y_a = y_p$, where y_a denotes the person's image with an identifier (label) a and y_p is an image forming the positive pair, $p = a$); in a negative pair, the two images belong to different people ($y_a \neq y_n$, where y_n is an image forming the negative pair, i.e., their identifiers do not coincide, $n \neq a$). Thus, the distance $d_{a,p}$ between the features for a positive pair and the distance $d_{a,n}$ between the features of different people are considered. The regularization factor m is introduced for the CNN to increase $d_{a,n}$ for different classes and also to decrease it for the same classes. When training the network without this factor, it will increase the distance between the images of different people and ignore the distance between the same classes. The reason is that it is easier to find the difference between different people than the similarity between the same people. Thus, the factor m restricts the growth of $d_{a,n}$ and decreases $d_{a,p}$. The triplet loss function [111] is given by

$$L = \sum_{\substack{a,p,n \\ y_a = y_p \neq y_n}} \max([m + d_{a,p} - d_{a,n}], 0). \tag{22}$$

In [53], triplet losses were used to analyze a negative pair including the images of different but most similar people. This approach allows training the network to find differences for people with similar appearances. The following accuracy values were obtained on the MSMT17 dataset [19]: 84.4% in the mAP metric) and 89.9% in the Rank1 metric (the best results for MSMT17 at the time of analysis).

According to the authors [114], triplet losses are not an effective enough approach to image clustering. Therefore, a cluster loss function was developed by them. Compared to the triplet loss function, this function yields larger interclass and smaller intraclass differences at the model output. The cluster loss function is given by

$$L_C = \frac{\beta \sum\limits_{i}^{p} d_i^{intra}}{\gamma + \sum\limits_{i}^{p} d_i^{inter}}, \tag{23}$$

where $d_i^{intra} = \sum_k \|f(x) - f_i^m\|_2^2$ is the intraclass variation for each ith identifier representing the distance between the features $f(x)$ of a sample identifier and the average value for the ith identifier over K images, $f_i^m = \frac{\sum_K f(x)}{K}$; $d_i^{inter} = \sum_{\forall i_d \in P, i_d \neq i} \|f_i^m - f_{i_d}^m\|_2^2$ is the interclass variation representing the distance between the average value of the identifier's features and the average value for the features of all P identifiers.

Several loss functions are sometimes used to improve the accuracy of person ReID. For example, in [65], two components were proposed for the loss function in order to determine the most effective features and the most significant attributes, namely, the loss function of metric distillation, L_d, and the loss function of attribute prior, L_p:

$$L = L_d + \alpha L_{p1} + \beta L_{p2}. \tag{24}$$

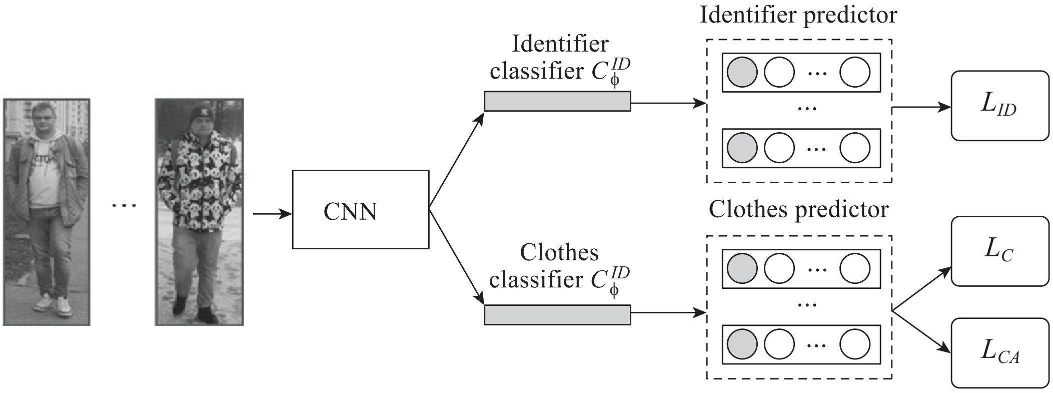
Identifier predictor
Identifier classifier C_ϕ^{ID}
CNN
L_{ID}
Clothes predictor
Clothes classifier C_ϕ^{ID}
L_C
L_{CA}

Fig. 10. The general scheme of the Clothes-based Adversarial Loss algorithm.

Here, $L_d = |d_{i,j} - \sum_{k=1}^{M} d_{i,j}^{k}|$ is the loss function specifying the distance $d_{i,j}$ between the feature vectors extracted by the ReID algorithm for the entire image and the features extracted by the modified algorithm intended to find the features of different attributes (thus, the contribution of each attribute to the total feature vector is estimated); d_{i,j^k} is the distance between x_i and x_j for the kth of M attributes.

Obviously, the component L_p of the loss function consists of two parts, namely, L_{p1} (the contribution of common attributes) and L_{p2} (the contribution of exclusive attributes). They have the following form:

$$L_{p1} = \max\left(0, \left(\frac{M_E}{M}\right)^{v} - \sum_{e=1}^{M_E} \frac{d_{i,j}^{e}}{\hat{d}_{i,j}}\right) + \max\left(0, \sum_{c=1}^{M-M_E} \frac{d_{i,j}^{c}}{\hat{d}_{i,j}} - 1 + \left(\frac{M_E}{M}\right)^{v}\right), \tag{25}$$

$$L_{p2} = \sum_{e=1}^{M_E} \max\left(0, e^{-\lambda}\frac{(\frac{M_E}{M})^{v}}{M_E} - \frac{d_{i,j}^{e}}{\hat{d}_{i,j}}\right) + \sum_{c=1}^{M-M_E} \max\left(0, \frac{d_{i,j}^{c}}{\hat{d}_{i,j}} - e^{\lambda}\frac{1-(\frac{M_E}{M})^{v}}{M-M_E}\right), \tag{26}$$

where $\hat{d}_{i,j} \approx \sum_{k=1}^{M} d_{i,j}^{k}$ is the predicted value of the distance between features and M_E is the number of exclusive attributes.

The researchers [6] suggested the Clothes-based Adversarial Loss (CAL) function to extract features without regard to person's clothes for long-term ReID. Figure 10 demonstrates the general scheme of this approach. It consists of two classifiers, C_ϕ^{ID} (identifier) and C_ϕ^{C} (clothes). Each classifier is trained separately. The first training stage is to minimize the clothes classification loss L_C based on the cross-entropy between the predicted clothes label $C_\phi^{C}(g_0(x_i))$ and y_i^C. The CNN is trained to ignore the clothes features by minimizing the loss function L_{CA} (Clothes-based Adversarial Loss), which determines the features not related to clothes. After training the clothes classifier, its weights are frozen, and the next stage aims at minimizing the loss function for the identifier classifier.

This algorithm was tested on the CCVID dataset [6]. According to the presented results, the CAL method increases the accuracy of person ReID by over 20% in the Rank1 and mAP metrics compared to the basic algorithm.

Currently, there is a growing emphasis on unsupervised or semi-supervised learning, in which the data have no labels and annotations prepared in advance. In some publications on person ReID, researchers suggest that information based on the existing labeled datasets with known identifiers be used in invisible domains. Invisible domains refer to datasets whose images were not used in training and which may have no labeled identifiers. Here, the matter concerns unsupervised domain adaptation (UDA). Such an approach was applied in [115] with the union of several datasets as input information. Market-1501 [26], DukeMTMC-ReID [20], CUHK03 [18], and MSMT17 [19], united in different combinations, were considered as source and target domains. In addition, two modules were introduced to study distinctive features for one domain and united domains. In the former case, RDSBN, the batch normalization module, reduces the impact of domain-specific features and improves the distinctiveness of facial features. In the latter case, the aggregation of domain information based on the Graph Convolution Network (GCN) decreases the distance between the features of different domains. GCN is used to construct a graph connecting all instances in a domain (a node generalizing the characteristic features for each person within a domain). This allows determining global descriptors for the entire domain.

Unsupervised person ReID methods based on domain adaptation for the target domain often work well on only one domain they have been adapted to. One solution of this problem was discussed in [116]. The idea proposed therein is to perform unsupervised lifelong adaptation to new data (continuous learning) with preserving the knowledge acquired for the previous domains. This is crucial for systems operating in real conditions with new data appearing regularly. The system may include additional video cameras installed in other locations, and adaptation to the new data must retain person ReID skills accumulated in the already known domains. For this purpose, a small number of samples from the existing domains are stored in long-term memory buffers for the clusters formed previously. When adapting the model to the new domain, the old patterns are also added to the sample for contrastive learning. The main principle of such CNN training is to maximize the similarity between a positive pair of images obtained under different conditions.

In unsupervised learning, some researchers employ clustering techniques to create pseudo labels during model training. In this case, images of different people may be combined in one cluster, whereas the cluster for the same person may be divided into two groups. This significantly reduces the accuracy of the CNN trained on such data. The authors [117] assumed that some of the information may be missing due to the limited number of samples for each person. Therefore, they proposed the Implicit Sample Extension (ISE) method to create support patterns on the boundaries of clusters based on real images of the current and neighboring clusters through a progressive linear interpolation (PLI) strategy. This strategy unites two clusters if they contain images of the same person and separates the cluster if it contains images of different people.

Self-supervised learning (SSL) methods focus on learning distinctive features from large unlabeled datasets. In [118], self-supervised pre-learning with unlabeled images of people was proposed to improve the accuracy of person ReID. Note that this method shows better results compared to traditional pre-learning on ImageNet. Its main idea is to extract global and local features. The Part-Aware SelfSupervised pre-training (PASS) system proposed in [118] consists of two parts having the same architecture: a student network and a teacher network. PASS trains the student network to match the output of the teacher. Global as well as local features from randomly chosen domains are supplied to its input. The teacher network analyzes only the global features. The similarity between the results is estimated based on cross-entropy. After pre-training, PASS is able to learn global features while automatically focusing on different local features of the images.

In [119], pre-training for person ReID was considered and a person tracking algorithm was applied to the video from the LUperson dataset. Each tracked person is assigned labels and these labels serve to form a new training sample, LUperson-NL, and pre-train the CNN. This approach introduces noisy labels into LUperson-NL, which possibly contain errors due to assigning identifiers in unlabeled datasets. For example, an error is assigning different identifiers to the same person by his or her images taken from different cameras or at different times. Another example is assigning the same identifiers to similar but different people. It is supposed that erroneous labels will be corrected later. The approach suggested in [119] involves three stages as follows. The first stage is supervised training for person ReID using the resulting noisy labels. The second stage consists in contrastive learning to correct the noisy labels: a prototype image is selected, and the new data are compared with the selected prototype; in the case of similarity, the new images are added to the cluster, and the feature vector averaged for all images in the cluster is calculated and dynamically updated. In the third stage, contrastive learning is applied based on the corrected labels. As a result, similar examples are united into one prototype, and the noisy labels are corrected.

The authors [120] proposed Part-based Pseudo Label Refinement (PPLR) to reduce the impact of noisy labels in unsupervised learning through a complementary relationship between global and local human features in the image. Irrelevant parts of the image (e.g., occlusions) may occur at different times, distort the composition of local and global human features, and finally cause incorrect predictions. To eliminate the effect of irrelevant parts, labels are refined based on the cross agreement score of similarity of k-nearest neighbors between the spaces of global and local human image features.

The paper [108] considered a skeleton-based approach to unsupervised person ReID; a contrastive learning scheme was presented to extract the features of unlabeled 3D human skeletons. Within this approach, skeleton masks are overlaid on raw data sequences, a prototype mask is selected, and clustering is performed using the most distinctive skeletal features. Similar skeletal features are compared to find distinctive features for different prototypes without any labels. Note that a correlation arises within the same video sequence due to changes in human movement. This correlation is considered using a Siamese architecture, which captures the most characteristic features of each person based on his or her skeleton.

6. THE EFFECTIVENESS OF PERSON REID ALGORITHMS: A COMPARATIVE ANALYSIS ON DIFFERENT DATASETS

Table 2 compares the effectiveness of different person ReID algorithms for single images from the large-scale and most widespread databases: Market-1501, DukeMTMC-ReID, and MSMT17.

Experimental results for CUHK03 were provided in a few papers [58, 59, 110]. However, this set includes no more than five human images obtained from each of the two cameras used, which is insufficient for an effective estimation of ReID accuracy. According to Table 2, the same algorithm yields different accuracy results for different datasets: the highest values for Market-1501 and the lowest values for MSMT17. These results are due to the fact that MSMT17 has much more images (see Table 1) than the others and a more complex video surveillance scenario was used to form this dataset (different times of day and weather conditions, images from both indoor and outdoor surveillance cameras). Although the accuracy figures for this dataset are significantly lower compared to the others, they can be considered more objective since the MSMT17 formation scenario is closer to real-life person ReID situations; hence, it more correctly reflects the effectiveness of algorithms for open-world person ReID systems. At the time of analysis, the best result in the mAP metric on the MSMT17 dataset was obtained for HBReID [53]; for this algorithm, Rank1 = 89.9%.

This result was reached by carefully selecting positive and negative pairs of images for triplet losses and using adversarial losses (studying the background and neglecting it when classifying people, i.e., considering human features only). The closest result in terms of accuracy, Rank1 = 87.5, was demonstrated by FlipReID [52]. This algorithm involves the averaged features of the input image and a randomly rotated one.

For the Market-1501 and DukeMTMC-ReID datasets, the most efficient person ReID algorithm among those considered was proposed in [57].

Besides visual features, this algorithm operates additional information about the time (frame number) and location (camera identifier) of shooting, which is provided with the datasets as an image file name from Market-1501 and DukeMTMC-ReID. For person ReID, the authors utilized a natural condition that a person cannot be in the field of view of several non-overlapping cameras

Table 2. The accuracy of person ReID for single images from Market-1501, DukeMTMC-ReID, and MSMT17 datasets

Algorithm	Year	Metrics	Market-1501	DukeMTMC-ReID	MSMT17
PCP+RPP [59]	2018	mAP	81.6	69.2	–
		Rank1	93.8	83.3	–
CASN [110]	2019	mAP	82.8	73.7	–
		Rank1	94.4	87.7	–
MGN+PTL [58]	2019	mAP	87.34	79.16	–
		Rank1	94.83	89.36	–
st-ReID [57]	2019	mAP	**95.5**	**92.7**	–
		Rank1	**98.0**	**94.5**	–
HOReID [61]	2020	mAP	84.9	75.6	–
		Rank1	94.2	86.9	–
AGW[2]	2021	mAP	–	–	49.3
		Rank1	–	–	68.3
		mINP	–	–	14.7
CIL [121]	2021	mAP	84.04	–	52.4
		Rank1	93.38	–	76.1
		mINP	57.9	–	12.45
SBS [65]	2021	mAP	88.29	78.26	–
		Rank1	95.55	89.21	–
Algorithm [55]	2021	mAP	89.2	79.6	57.2
		Rank1	96.2	91.0	81.9
FlipReID [52]	2021	mAP	94.7	90.7	81.3
		Rank1	95.8	93.0	87.5
HBReID [53]	2021	mAP	–	–	**84.4**
		Rank1	–	–	**89.9**
RANGEv2 [84]	2022	mAP	86.8	78.2	51.3
		Rank1	94.7	87.0	76.4
RANGEv2+K-reciprocal [84]	2022	mAP	91.3	84.2	–
		Rank1	95.1	88.7	–
CAL [6]	2022	mAP	87.5	–	57.3
		Rank1	94.7	–	79.9

simultaneously: he or she needs some time for transition. Thus, all images with visual similarity to the query must be checked (whether these people could be in a certain place at a certain time relative to the previous one), and all irrelevant images by these criteria are ignored during person ReID. Although this approach was proposed in 2019, its Rank1 and Rank5 metric values are currently among the best results on the Market-1501 and DukeMTMC-ReID datasets. Note that the authors of the algorithm [57] carried out no experiments on other datasets, probably due to insufficient spatiotemporal information provided therein. Despite accuracy variations for different datasets, most modern algorithms have the following common property (see Table 2): if an improvement in accuracy is observed for one dataset, it will be observed for other datasets as well.

Table 3 compares the effectiveness of different person ReID algorithms for video sequences. According to this table, an improvement in accuracy on one dataset not necessarily implies an improvement on another dataset.

The datasets used in training and testing contain sequences of human images from several frames (tracklets), and the number of tracklets in separate datasets for each person differs in the number of images. The application of tracklets allows covering temporal features to exclude the impact of short-term occlusions, consider gait information, or average visual features for dynamic body parts

Table 3. The accuracy of person ReID algorithms for video sequences from MARS, DukeVideo, PRID, QMUL iLIDS, iLIDS-VID, and VIPer datasets

Algorithm	Year	Metrics	MARS	DukeVideo	PRID	iLIDS-VID
AGW [2]	2021	mAP	83.0	94.9	–	–
		Rank1	87.6	95.4	94.4	–
		mINP	63.9	91.9	95.4	–
PiT [3]	2022	mAP	**97.23**	–	–	86.80
		Rank1	**90.22**	–	–	92.07
MetaBin [88]	2021	mAP	–	–	81.0	87.0
		Rank1	–	–	74.2	81.3
SSN3D [74]	2021	mAP	86.2	**96.3**	–	–
		Rank1	90.1	96.8	–	88.9
PMP–MA [60]	2022	mAP	88.1	**96.3**	99.3	95.3
		Rank1	90.6	**97.2**	**98.9**	92.8
		Rank5	99.6	99.3	**100**	99.3
Algorithm [83]	2022	mAP	82.6	94.2	–	–
		Rank1	88.2	95.4	96.6	89.3
		Rank5	96.5	99.3	**100**	98.7

on several frames. Consequently, the number of human images in a tracklet appreciably affects the estimation of a particular algorithm.

The PiT algorithm [3] demonstrated the best accuracy in the mAP and Rank1 metrics on the MARS dataset (see Table 3). This algorithm uses local feature pyramids with varying degrees of detail and averages the descriptors over several frames. The PMP-MA algorithm [60] demonstrated the best accuracy in the mAP metric on the Duke-Video and iLIDS-VID datasets, the best results were obtained in the mAP metric. PMP-MA involves a pyramidal representation of multi-attention schemes and considers the results of fine-tuning the CNN, including the selection of batch size during training. PMP-MA and the algorithm [83] yielded Rank5 = 100% on the PRID dataset, which is ensured by the relative ease of the dataset for this metric (only two cameras, a fairly uniform background, and rare person's occlusions with other people).

It is topical to compare the accuracy of person ReID algorithms during training and testing on different databases (i.e., estimate their effectiveness when changing domains). Table 4 presents the accuracy of person ReID of the algorithms based on feature-finding approaches with domain shiftability.

Note that regardless of the algorithm used, increasing the training sample by uniting the existing datasets can improve the accuracy of person ReID, as confirmed by the studies in [12, 48]. For example [48], adding images from MSMT17 to the synthetic set used as a training sample improved the mAP value from 47.6% to 61% on the DukeMTMC-ReID dataset.

Also, see [48], including data from the target domain in the training sample increased the Rank1 value for MSMT17 from 6.3 to 36.8%. Similarly, in [12], the use of images from the target domain in training improved the mAP value from 33.9 to 82.3% on Market-1501 and 33.6 to 73.2% on DukeMTMC-ReID.

Among modern cross-domain person ReID algorithms (see Table 4), the highest accuracy in the mAP metric was achieved for IDM [66] on the Market-1501 and MSMT17 datasets. In this algorithm, intermediate domains combining the features of the source and target domains are generated to improve stability to domain shift. When using DukeMTMC-ReID as the target domain, the JVTC approach [9] using the synthetic UnrealPerson dataset as the training sample [9] proved to be the most effective.

Table 4. The accuracy of cross-domain person ReID algorithms on Market-1501, DukeMTMC-ReID, MSMT17, VIPeR, PRID, and GRID datasets

Algorithm (Year)	Dataset	Metric	Test sample					
			Market -1501	DukeMTMC-ReID	MSMT17	VIPeR	PRID	GRID
Algorithm [48]	RandPerson [48]	mAP	28.8	27.1	6.3	–	–	–
		Rank1	55,6	47,6	20,1	–	–	–
	RandPerson [48] + MSMT17	mAP	35.8	39.8	36.8	–	–	–
(2020)		Rank1	62.3	61.0	65.0	–	–	–
SNR [12]	Market–1501	mAP	84.7	33.6	–	42.3	42.2	36.7
		Rank1	**94.4**	55.1	–	32.3	30.0	29.0
(2020)	DukeMTMC–ReID	mAP	33.9	72.9	–	41.2	45.4	35.3
		Rank1	66.7	84.4	–	32.6	35.0	26.0
	Market–1501+DukeMTMC–ReID+CUHK+MSMT17	mAP	82.3	73.2	–	65.0	60.0	41.3
		Rank1	93.4	85.5	–	55.1	49.0	30.4
NRMT [10]	Market–1501	mAP	–	62.3	19.8	–	–	–
		Rank1	–	78.1	43.7	–	–	–
(2020)	DukeMTMC–ReID	mAP	72.2	–	20.6	–	–	–
		Rank1	88.0	–	45.2	–	–	–
Algorithm [11]	Market–1501	* mAP	–	65.2	20.4	–	–	–
		Rank1	–	79.5	43.7	–	–	–
	DukeMTMC–ReID	mAP	71.5	–	24.3	–	–	–
(2020)		Rank1	88.1	–	51.7	–	–	–
CBN [9] (2021)	UnrealPerson [9]	mAP	54.3	49.4	15.3	–	–	–
		Rank1	79.0	69.7	38.5	–	–	–
JVTC [9] (2021)		mAP	80.2	**75.2**	34.8	–	–	–
		Rank1	93.0	**88.3**	68.2	–	–	–
MetaBin [88] MobileNetV2 (2021)	CUHK02+ CUHK03+ Market–1501 + +DukeMTMC–ReID+CUHK–SYSU	mAP	–	–	–	66.0	79.8	**58.1**
		Rank1	–	–	–	56.9	72.5	**49.7**
MetaBin [88] ResNet–50 (2021)		mAP	–	–	–	**68.6**	**81.0**	57.9
		Rank1	–	–	–	**59.9**	**74.2**	48.4
QAConv [49] (2022)	ClonedPerson [49]	mAP	21.8	–	18.5	–	–	–
		Rank1	22.6	–	49.1	–	–	–
IDM [66] (2022)	Market–1501	mAP	–	73.2	40.2	–	–	–
		Rank1	–	85.5	**69.9**	–	–	–
	DukeMTMC–ReID	mAP	**85.3**	–	**40.5**	–	–	–
		Rank1	94.2	–	69.5	–	–	–
	MSMT17	mAP	85.2	73.6	–	–	–	–
		Rank1	94.1	84.6	–	–	–	–

Video-based person ReID algorithms aimed at increasing stability to domain shift often use the VIPeR, PRID, and GRID datasets. Among the analyzed approaches, the best Rank1 and mAP values were obtained for the MetaBIN algorithm [88] (see Table 4). The main idea of this algorithm is to generalize the normalization layers and reduce the impact of features inherent to the source domain.

Table 5. The accuracy of person ReID algorithms with unsupervised and semi-supervised learning on Market-1501, DukeMTMC-ReID, and MSMT17 datasets

Algorithm	Year	Metrics	Market-1501	DukeMTMC -ReID	MSMT17
[115] trained on Market–1501	2021	mAP Rank1	– –	66.6 80.3	34.9 64.7
[115] trained on Duke–MTMC–ReID	2021	mAP Rank1	81.5 92.9	– –	33.6 64.0
[116] with lifelong domain adaptation	2022	mAP Rank1	59.3 82.7	– –	40.8 67.5
ISE[117] + GeM–pooling	2022	mAP Rank1	85.3 94.3	– –	37.0 67.6
PASS [118]	2022	mAP Rank1	**93.3** **96.9**	– –	**74.4** **89.7**
PNL[119]+MGN pre-trained on LuPerson	2022	mAP Rank1	91.9 96.6	**84.3** **92.0**	68.0 86.0
PPLR [120] without camera labels	2022	mAP Rank1	81.5 92.8	– –	31.4 61.1
PPLR [120] with camera labels	2022	mAP Rank1	84.4 94.3	– –	42.2 73.3

Unsupervised or semi-supervised learning on unlabeled data is another approach for adapting to domain shift during person ReID. According to Table 5, the algorithms [118, 119] are the best ones among those considered. The algorithm [118] involves pre-training on unlabeled human images, a two-streamed architecture, and global and local features. The algorithm [119] involves pre-training on the unlabeled LUPerson dataset and noisy labels for it are generated and corrected during training. In most publications, the ImageNet dataset was employed for pre-training. However, as has been shown by recent studies [118, 119], the most efficient approach is to use human images in this stage.

7. CONCLUSIONS

Person ReID in a distributed video surveillance system is a fairly new topical problem, successful solutions based on deep learning technologies have been recently proposed. This paper has considered the general principles of person ReID using convolutional neural networks during video surveillance. Person ReID systems have been classified. The existing datasets for training deep neural network architectures have been analyzed; the approaches to increasing the number of images in databases have been described; the types of human image features have been discussed. The backbone models of convolutional neural network architectures used for person ReID, as well as their modifications and training methods, have been examined. The effectiveness of person ReID on different datasets, including cross-domain person ReID, has been studied. The effectiveness of the existing person ReID approaches has been estimated in different metrics; the advantages and drawbacks of these metrics have been outlined. Although deep learning and neural networks have demonstrated their great capabilities in analyzing video images, there are still problems to be solved for person ReID. One of the most important shortcomings is that deep learning needs a huge amount of accurately annotated datasets, which requires tedious work and often causes distortions. Many researchers are sharing their data on publicly available platforms, which is useful for developing a unified estimation index. However, some ReID datasets were eliminated from public access, e.g., DukeMTMC-ReID [20]. MTMC17 [19] is no more publicly available and can only be obtained after signing an agreement with the authors (the use for research purposes only without transfer to third parties [122]).

Deep learning has achieved satisfactory results in image classification and segmentation. For person ReID, especially from image sequences, the performance of deep learning is still insufficiently high. Therefore, it is of interest to develop new solutions using CNNs with higher accuracy and speed, especially for cross-domain person ReID.

FUNDING

This work is supported by the National High-end Foreign Experts Program (G2021016028L) and Zhejiang Shuren University Basic Scientific Research Special Funds.

REFERENCES

1. Ye, S., Bohush, R.P., and Chen, H., Person Tracking and Re-identification for Multicamera Indoor Video Surveillance Systems, *Pattern Recognit. Image Anal.*, 2020, no. 30, pp. 827–837. https://doi.org/10.1134/S1054661820040136
2. Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., and Hoi, S.C., Deep Learning for Person Re-identification: A Survey and Outlook, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. https://doi.org/10.1109/TPAMI.2021.3054775
3. Zang, X., Li, G., and Gao, W., Multi-direction and Multi-scale Pyramid in Transformer for Video-based Pedestrian Retrieval, *arXiv: abs/2202.06014*, 2022. https://doi.org/10.1109/TII.2022.3151766
4. Mihaescu, R., Chindea, M., Paleologu, C., Carata, S., and Ghenescu, M., Person Re-Identification across Data Distributions Based on General Purpose DNN Object Detector, *Algorithms*, 2020, no. 13, art. no. 343. https://doi.org/10.3390/a13120343
5. Liu, H., Qin, L., Cheng, Z., and Huang, Q., Set-based Classification for Person Re-identification Utilizing Mutual-information, *2013 IEEE International Conference on Image Processing*, 2013, pp. 3078–3082. https://doi.org/10.1109/ICIP.2013.6738634
6. Gu, X., Chang, H., Ma, B., Bai, S., Shan, S., and Chen, X., Clothes-Changing Person Re-identification with RGB Modality, *arXiv: abs/2204.06890*, 2022. https://doi.org/10.48550/arXiv.2204.06890
7. Huang, Y., Wu, Q., Zhong, Y., and Zhang, Z., Clothing Status Awareness for Long-Term Person Re-Idenification, *2021 IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11895–11904. https://doi.org/10.1109/ICCV48922.2021.01168
8. Hao, X., Zhao, S., Ye, M., and Shen, J., Cross-Modality Person Re-identification via Modality Confusion and Center Aggregation, *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 16383–16392. https://doi.org/10.1109/ICCV48922.2021.0160
9. Zhang, T., Xie, L., Wei, L., Zhuang, Z., Zhang, Y., Li, B., and Tian, Q., UnrealPerson: An Adaptive Pipeline towards Costless Person Re-identification, *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 11501–11510. https://doi.org/10.1109/CVPR46437.2021.01134
10. Zhao, F., Liao, S., Xie, G., Zhao, J., Zhang, K., and Shao, L., Unsupervised Domain Adaptation with Noise Resistible Mutual-Training for Person Re-identification, *ECCV 2020. Lecture Notes in Computer Science*, vol. 12356, Cham: Springer, 2020, pp. 526–544. https://doi.org/10.1007/978-3-030-58621-8_31
11. Luo, C., Song, C., and Zhang, Z., Generalizing Person Re-identification by Camera-Aware Invariance Learning and Cross-Domain Mixup, *ECCV 2020. Lecture Notes in Computer Science*, vol. 12356, Cham: Springer, 2020, pp. 224–241. https://doi.org/10.1007/978-3-030-58555-6_14
12. Jin, X., Lan, C., Zeng, W., Chen, Z., and Zhang, L., Style Normalization and Restitution for Generalizable Person Re-identification, *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3140–3149. https://doi.org/10.1109/cvpr42600.2020.00321
13. Song, J., Yang, Y., Song, Y., Xiang, T., and Hospedales, T.M., Generalizable Person Re-identification by Domain-Invariant Mapping Network, *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 719–728. https://doi.org/10.1109/CVPR.2019.00081

14. Ihnatsyeva, S., Bohush, R., and Ablameyko, S., Joint Dataset for CNN-based Person Re-identification, *Proceedings of the 15th International Conference on Pattern Recognition and Information Processing (PRIP'2021)*, September 21–24, 2021, Minsk: United Institute of Informatics Problems of the National Academy of Sciences of Belarus, 2021, pp. 33–37.

15. Liao, S., Mo, Z., Hu, Y., and Li, S., Open-set Person Re-identification, *arXiv: abs/1408.0872*, 2014. https://doi.org/10.48550/arXiv.1408.0872

16. Li, W., Zhao, R., and Wang, X., Human Reidentification with Transferred Metric Learning, *Proceedings of the 11th Asian Conference on Computer Vision (ACCV)*, 2012. https://doi.org/10.1007/978-3-642-37331-2_3

17. Li, W. and Wang, X., Locally Aligned Feature Transforms across Views, *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3594–3601. https://doi.org/10.1109/CVPR.2013.461

18. Li, W., Zhao, R., Xiao, T., and Wang, X., DeepReID: Deep Filter Pairing Neural Network for Person Re-identification, *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 152–159. https://doi.org/10.1109/CVPR.2014.27

19. Wei, L., Zhang, S., Gao, W., and Tian, Q., Person Transfer GAN to Bridge Domain Gap for Person Re-identification, *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 79–88. https://doi.org/10.1109/CVPR.2018.00016

20. Ristani, E., Solera, F., Zou, R.S., Cucchiara, R., and Tomasi, C., Performance Measures and a Data Set for Multi-target, Multi-camera Tracking, *arXiv: abs/1609.01775*, 2016. https://doi.org/10.1007/978-3-319-48881-3_2

21. Exposing.ai. Duke MTMC. URL: https://exposing.ai/duke_mtmc.

22. Zheng, L., Zhang, H., Sun, S., Chandraker, M., Yang, Y., and Tian, Q., Person Re-identification in the Wild, *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3346–3355. https://doi.org/10.1109/CVPR.2017.357

23. Xiao, T., Li, S., Wang, B., Lin, L., and Wang, X., Joint Detection and Identification Feature Learning for Person Search, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3376–3385. https://doi.org/10.1109/CVPR.2017.360

24. Zheng, L., Bie, Z., Sun, Y., Wang, J., Su, C., Wang, S., and Tian, Q., MARS: A Video Benchmark for Large-Scale Person Re-identification, *ECCV 2016. Lecture Notes in Computer Science*, vol. 9910, Cham: Springer, 2016, pp. 863–884. https://doi.org/10.1007/978-3-319-46466-4_52

25. Song, G., Leng, B., Liu, Y., Hetang, C., and Cai, S., Region-based Quality Estimation Network for Large-scale Person Re-identification, *arXiv: abs/1711.08766*, 2018. https://doi.org/10.48550/arXiv.1711.08766

26. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q., Scalable Person Re-identification: A Benchmark, *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1116–1124. https://doi.org/10.1109/ICCV.2015.133

27. Gray, D., Brennan, S., and Tao, H., Evaluating Appearance Models for Recognition, Reacquisition, and Tracking, *Proceedings of the IEEE Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, 2007.

28. Hirzer, M., Beleznai, C., Roth, P.M., and Bischof, H., Person Re-identification by Descriptive and Discriminative Classification, *SCIA. Lecture Notes in Computer Science*, vol. 6688, Berlin–Heidelberg: Springer, 2011, pp. 91–102. https://doi.org/10.1007/978-3-642-21227-7_9

29. Zheng, W., Gong, S., and Xiang, T., UnrealPerson: An Adaptive Associating Groups of People, *BMVC*, 2009. https://doi.org/10.5244/C.23.23

30. Karanam, S., Gou, M., Wu, Z., Rates-Borras, A., Camps, O.I., and Radke, R.J., A Systematic Evaluation and Benchmark for Person Re-identification: Features, Metrics, and Datasets, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019, no. 41, pp. 523–536. https://doi.org/10.1109/TPAMI.2018.2807450

31. Ihnatsyeva, S. and Bohush, R., PolReID, 2021. URL: https://github.com/SvetlanaIgn/PolReID

32. Li, S., Xiao, T., Li, H., Zhou, B., Yue, D., and Wang, X., Person Search with Natural Language Description, *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5187–5196. https://doi.org/10.1109/CVPR.2017.551

33. Ding, Z., Ding, C., Shao, Z., and Tao, D., Semantically Self-Aligned Network for Text-to-Image Part-aware Person Re-identification, *arXiv: abs/2107.12666*, 2021.

34. Li, X., Zheng, W., Wang, X., Xiang, T., and Gong, S., Multi-Scale Learning for Low-Resolution Person Re-identification, *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 3765–3773. https://doi.org/10.1109/ICCV.2015.429

35. Jing, X., Zhu, X., Wu, F., Hu, R., You, X., Wang, Y., Feng, H., and Yang, J., Super-Resolution Person Re-identification with Semi-Coupled Low-Rank Discriminant Dictionary Learning, *IEEE Transactions on Image Processing*, 2015, no. 26, pp. 1363–1378. https://doi.org/10.1109/TIP.2017.2651364

36. Wu, A., Zheng, W., Yu, H., Gong, S., and Lai, J., RGB-Infrared Cross-Modality Person Re-identification, *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 5390–5399. https://doi.org/10.1109/ICCV.2017.575

37. Nguyen, T.D., Hong, H.G., Kim, K., and Park, K.R., Person Recognition System Based on a Combination of Body Images from Visible Light and Thermal Cameras, *Sensors*, 2017, no. 17. https://doi.org/10.3390/s17030605

38. Pang, L., Wang, Y., Song, Y., Huang, T., and Tian, Y., Cross-Domain Adversarial Feature Learning for Sketch Re-identification, *Proceedings of the 26th ACM International Conference on Multimedia*, 2018. https://doi.org/10.1145/3240508.3240606

39. Xiao, T., Li, S., Wang, B., Lin, L., and Wang, X., End-to-end Deep Learning for Person Search, *arXiv: abs/1604.01850*, 2016.

40. Layne, R., Hospedales, T.M., and Gong, S., Investigating Open-World Person Re-identification Using a Drone, *ECCV Workshops*, 2014. https://doi.org/10.1007/978-3-319-16199-0_16

41. Fu, D., Chen, D., Bao, J., Yang, H., Yuan, L., Zhang, L., Li, H., and Chen, D., Unsupervised Pre-training for Person Re-identification, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 14745–14754. https://doi.org/10.1109/CVPR46437.2021.01451

42. Fabbri, M., Brasó, G., Maugeri, G., Cetintas, O., Gasparini, R., Osep, A., Calderara, S., Leal-Taixe, L., and Cucchiara, R., MOTSynth: How Can Synthetic Data Help Pedestrian Detection and Tracking, *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 10829–10839. https://doi.org/10.1109/iccv48922.2021.01067

43. Makehuman community. Makehuman, 2020. URL: http://www.makehumancommunity.org.

44. Epic Games Incorporated. Unreal Engine, 2020. URL: https://www.unrealengine.com.

45. Barbosa, I.B., Cristani, M., Caputo, B., Rognhaugen, A., and Theoharis, T., Looking beyond Appearances: Synthetic Training Data for Deep CNNs in Re-identification, *arXiv: abs/1701.03153*, 2018. https://doi.org/10.1016/j.cviu.2017.12.002

46. Bak, S., Carr, P., and Lalonde, J., Domain Adaptation through Synthesis for Unsupervised Person Re-identification, *arXiv: abs/1804.10094*, 2018. https://doi.org/10.1007/978-3-030-01261-8_12

47. Sun, X. and Zheng, L., Dissecting Person Re-identification from the Viewpoint of Viewpoint, *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 608–617. https://doi.org/10.1109/CVPR.2019.00070

48. Wang, Y., Liao, S., and Shao, L., Surpassing Real-World Source Training Data: Random 3D Characters for Generalizable Person Re-identification, *Proceedings of the 28th ACM International Conference on Multimedia*, 2020. https://doi.org/10.1145/3394171.3413815

49. Wang, Y., Liang, X., and Liao, S., Cloning Outfits from Real-World Images to 3D Characters for Generalizable Person Re-identification, *arXiv: abs/2204.02611*, 2022. https://doi.org/10.48550/arXiv.2204.02611

50. Unity Technologies. 2020. Unity3D: Cross-platform Game Engine. URL: https://unity.com.

51. Zhong, Z., Zheng, L., Kang G., Li, S., and Yang, Y., Random Erasing Data Augmentation, *Proceedings of AAAI*, 2020. https://doi.org/10.1609/AAAI.V34I07.7000

52. Ni, X. and Rahtu, E., FlipReID: Closing the Gap between Training and Inference in Person Re-identification, *Proceedings of 2021 9th European Workshop on Visual Information Processing (EUVIP)*, 2021, pp. 1–6. https://doi.org/10.1109/EUVIP50544.2021.9484010

53. Li, W., Xu, F., Zhao, J., Zheng, R., Zou, C., Wang, M., and Cheng, Y., HBReID: Harder Batch for Re-identification, *arXiv: abs/2112.04761*, 2021. https://doi.org/10.48550/arXiv.2112.04761

54. Huang, Y., Zha, Z., Fu, X., Hong, R., and Li, L., Real-World Person Re-identification via Degradation Invariance Learning, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 14072–14082. https://doi.org/10.1109/cvpr42600.2020.01409

55. Jiang, Y., Chen, W., Sun, X., Shi, X., Wang, F., and Li, H., Exploring the Quality of GAN Generated Images for Person Re-identification, *Proceedings of the 29th ACM International Conference on Multimedia*, 2021. https://doi.org/10.1145/3474085.3475547

56. Wu, C., Ge, W., Wu, A., and Chang, X., Camera-Conditioned Stable Feature Generation for Isolated Camera Supervised Person Re-identification, *arXiv: abs/2203.15210*, 2022. https://doi.org/10.48550/arXiv.2203.15210

57. Wang, G., Lai, J., Huang, P., and Xie, X., Spatial-Temporal Person Re-identification, *arXiv: abs/1812.03282*, 2019. https://doi.org/10.1609/aaai.v33i01.33018933

58. Yu, Z., Jin, Z., Wei, L., Guo, J., Huang, J., Cai, D., He, X., and Hua, X., Progressive Transfer Learning for Person Re-identification, *IJCAI*, 2019. https://doi.org/10.24963/ijcai.2019/586

59. Sun, Y., Zheng, L., Yang, Y., Tian, Q., and Wang, S., Beyond Part Models: Person Retrieval with Refined Part Pooling, *Proceedings of ECCV*, 2018. https://doi.org/10.1007/978-3-030-01225-0_30

60. Bayoumi, R.M., Hemayed, E.E., Ragab, M.E., and Fayek, M.B., Person Re-identification via Pyramid Multipart Features and Multi-Attention Framework, *Big Data and Cognitive Computing*, 2022. https://doi.org/10.3390/bdcc6010020

61. Wang, G., Yang, S., Liu, H., Wang, Z., Yang, Y., Wang, S., Yu, G., Zhou, E., and Sun, J., High-Order Information Matters: Learning Relation and Topology for Occluded Person Re-identification, *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 6448–6457. https://doi.org/10.1109/CVPR42600.2020.00648

62. Sun, K., Xiao, B., Liu, D., and Wang, J., Deep High-Resolution Representation Learning for Human Pose Estimation, *Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5686–5696. https://doi.org/10.1109/CVPR.2019.00584

63. Yang, J., Zhang, J., Yu, F., Jiang, X., Zhang, M., Sun, X., Chen, Y., and Zheng, W.S., Learning to Know Where to See: A Visibility-Aware Approach for Occluded Person Re-identification, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11885–11894

64. Fang, H., Xie, S., Tai, Y., and Lu, C., RMPE: Regional Multi-person Pose Estimation, *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2353–2362. https://doi.org/10.1109/ICCV.2017.256

65. Chen, X., Liu, X., Liu, W., Zhang, X., Zhang, Y., and Mei, T., Explainable Person Re-identification with Attribute-guided Metric Distillation, *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2022, pp. 11793–11802. https://doi.org/10.1109/ICCV48922.2021.01160

66. Dai, Y., Sun, Y., Liu, J., Tong, Z., Yang, Y., and Duan, L., Bridging the Source-to-target Gap for Cross-domain Person Re-identification with Intermediate Domains, *arXiv: abs/2203.01682*, 2022. https://doi.org/10.48550/arXiv.2203.01682

67. Zhang, H., Cisse, M., Dauphin, Y., and Lopez-Paz, D., Mixup: Beyond Empirical Risk Minimization, *arXiv: abs/1710.09412*, 2018. https://doi.org/10.48550/arXiv.1710.09412

68. Huang, X. and Belongie, S.J., Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization, *Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1510–1519. https://doi.org/10.1109/ICCV.2017.167

69. Avola, D., Cascio, M., Cinque, L., Fagioli, A., and Petrioli, C., Person Re-identification Through Wi-Fi Extracted Radio Biometric Signatures, *IEEE Transactions on Information Forensics and Security*, 2022, vol. 17, pp. 1145–1158. https://doi.org/10.1109/TIFS.2022.3158058

70. Qi, L., Shen, J., Liu, J., Shi, Y., and Geng, X., Label Distribution Learning for Generalizable Multisource Person Re-identification, *arXiv: abs/2204.05903*, 2022. https://doi.org/10.48550/arXiv.2204.05903

71. Yang, X., Zhou, Z., Wang, Q., Wang, Z., Li, X., and Li, H., Cross-domain Unsupervised Pedestrian Re-identification Based on Multi-view Decomposition, *Multimed Tools Appl.*, 2022. https://doi.org/10.1007/s11042-021-11797-w

72. Elharrouss, O., Almaadeed, N., Al-Maadeed, S.A., and Bouridane, A., Gait Recognition for Person Re-identification, *J. Supercomput.*, 2021, no. 77, pp. 3653–3672. https://doi.org/10.1007/s11227-020-03409-5

73. Chao, H., He, Y., Zhang, J., and Feng, J., GaitSet: Regarding Gait as a Set for Cross-View Gait Recognition, *arXiv: abs/1811.06186*, 2019. https://doi.org/10.1609/aaai.v33i01.33018126

74. Jiang, X., Qiao, Y., Yan, J., Li, Q., Zheng, W., and Chen, D., SSN3D: Self-Separated Network to Align Parts for 3D Convolution in Video Person Re-identification, *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, no. 35(2), pp. 1691–1699. https://ojs.aaai.org/index.php/AAAI/article/view/16262

75. Yang, F., Wang, X., Zhu, X., Liang, B., and Li, W., Relation-Based Global-Partial Feature Learning Network for Video-Based Person Re-identification, *Neurocomputing*, 2022, vol. 488, pp. 424–435. https://doi.org/10.1016/j.neucom.2022.03.032

76. Lu, Z., Zhang, G., Huang, G., Yu, Z., Pun, C., and Ling, K., Video Person Re-identification Using Key Frame Screening with Index and Feature Reorganization Based on Inter-frame Relation, *Int. J. Mach. Learn. Cyber.*, 2022. https://doi.org/10.1007/s13042-022-01560-4

77. Yadav, A. and Vishwakarma, D.K., Person Re-identification Using Deep Learning Networks: A Systematic Review, *arXiv: abs/2012.13318*, 2020. https://doi.org/10.48550/arXiv.2012.13318

78. Zhang, Z., Lan, C., Zeng, W., Jin, X., and Chen, Z., Relation-Aware Global Attention for Person Re-identification, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3183–3192. https://doi.org/10.1109/CVPR42600.2020.00325

79. Pathak, P., Eshratifar, A.E., and Gormish, M.J., Video Person Re-ID: Fantastic Techniques and Where to Find Them, *Proceedings of AAAI*, 2020. https://doi.org/10.1609/aaai.v34i10.7219

80. Liu, X., Zhang, P., Yu, C., Lu, H., and Yang, X., Watching You: Global-guided Reciprocal Learning for Video-based Person Re-identification, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 13329–13338. https://doi.org/10.1109/CVPR46437.2021.01313

81. Gao, S., Wang, J., Lu, H., and Liu, Z., Pose-Guided Visible Part Matching for Occluded Person ReID, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11741–11749. https://doi.org/10.1109/cvpr42600.2020.01176

82. Zhang, S., Yin, Z., Wu, X., Wang, K., Zhou, Q., and Kang, B., FPB: Feature Pyramid Branch for Person Re-identification, *arXiv: abs/2108.01901*, 2021. https://doi.org/10.48550/arXiv.2108.01901

83. Yang, F., Li, W., Liang, B., Han, S., and Zhu, X., Multi-stage Attention Network for Video-Based Person Re-identification, *IET Comput. Vis.*, 2022, pp. 1–11. https://doi.org/10.1049/cvi2.1210

84. Wu, G., Zhu, X., and Gong, Sh., Learning Hybrid Ranking Representation for Person Re-identification, *Pattern Recognition*, 2022, vol. 121. https://doi.org/10.1016/j.patcog.2021.108239

85. Zhong, Z., Zheng, L., Cao, D., and Li, S., Re-ranking Person Re-identification with k-Reciprocal Encoding, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3652–3661. https://doi.org/10.1109/CVPR.2017.389

86. Bohush, R.P., Ablameyko, S.V., and Adamovskiy, E.R., Image Similarity Estimation Based on Ratio and Distance Calculation between Features, *Pattern Recognit. Image Anal.*, 2020, no. 30, pp. 147–159. https://doi.org/10.1134/S1054661820020030

87. He, K., Zhang, X., Ren, S., and Sun, J., Deep Residual Learning for Image Recognition, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. https://doi.org/10.1109/cvpr.2016.90

88. Choi, S., Kim, T., Jeong, M., Park, H., and Kim, C., Meta Batch-Instance Normalization for Generalizable Person Re-identification, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 3424–3434. https://doi.org/10.1109/CVPR46437.2021.00343

89. Huang, G., Liu, Z., and Weinberger, K.Q., Densely Connected Convolutional Networks, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2261–2269. https://doi.org/10.1109/CVPR.2017.243

90. Chen, P., Dai, P., Liu, J., Zheng, F., Tian, Q., and Ji, R., Dual Distribution Alignment Network for Generalizable Person Re-identification, *arXiv: abs/2007.13249*, 2021. https://doi.org/10.48550/arXiv.2007.13249

91. Zhao, C., Chen, K., Wei, Z., Chen, Y., Miao, D., and Wang, W., Multilevel Triplet Deep Learning Model for Person Re-identification, *Pattern Recognit. Lett.*, 2019, no. 117, pp. 161–168. https://doi.org/10.1016/j.patrec.2018.04.029

92. Yao, Y., Jiang, X., Fujita, H., and Fang, Z., A Sparse Graph Wavelet Convolution Neural Network for Video-Based Person Re-identification, *Pattern Recognition*, 2022, vol. 129. https://doi.org/10.1016/j.patcog.2022.108708

93. Lu, P., Lu, K., Wang, W., Zhang, J., Chen, P., and Wang, B., Real-Time Pedestrian Detection in Monitoring Scene Based on Head Model, *Intelligent Computing Theories and Application (ICIC 2019). Lecture Notes in Computer Science*, vol. 11644, Cham: Springer, 2019, pp. 558–568. https://doi.org/10.1007/978-3-030-26969-2_53

94. Lee, S., Kang, Q., Madireddy, S., Balaprakash, P., Agrawal, A., Choudhary, A.N., Archibald, R., and Liao, W., Improving Scalability of Parallel CNN Training by Adjusting Mini-Batch Size at Run-Time, *Proceedings of the 2019 IEEE International Conference on Big Data*, 2019, pp. 830–839. https://doi.org/10.1109/BigData47090.2019.9006550

95. Lewkowycz, A., How to Decay Your Learning Rate, *arXiv: abs/2103.12682*, 2021. https://doi.org/10.48550/arXiv.2103.12682

96. Lewkowycz, A., Bahri, Y., Dyer E., Sohl-Dickstein, J., and Gur-Ari, G., The Large Learning Rate Phase of Deep Learning: The Catapult Mechanism, *arXiv: abs/2003.02218*, 2020. https://doi.org/10.48550/arXiv.2003.02218

97. Ulyanov, D., Vedaldi, A., and Lempitsky, V.S., Instance Normalization: The Missing Ingredient for Fast Stylization, *arXiv: abs/1607.08022*, 2016. https://doi.org/10.48550/arXiv.1607.08022

98. Chen, H., Ihnatsyeva, S., Bohush, R., and Ablameyko, S., Choice of Activation Function in Convolution Neural Network in Video Surveillance Systems, *Programming and Computer Software*, 2022, no. 5, pp. 312–321. https://doi.org/10.1134/S0361768822050036

99. Nair, V. and Hinton, G.E., Rectified Linear Units Improve Restricted Boltzmann Machines, *Proceedings of the 27th International Conference on Machine Learning (ICML)*, Haifa, Israel, 2010, pp. 807–814.

100. Maas, A.L., Hannum, A.Y., and Ng, A.Y., Rectifier Nonlinearities Improve Neural Network Acoustic Models, *Proceedings of the 30th International Conference on Machine Learning (ICML)*, Atlanta, Georgia, 2013

101. Xu, B., Wang, N., Chen, T., and Li, M., Empirical Evaluation of Rectified Activations in Convolutional Network, *arXiv: abs/1505.00853*, 2015. https://doi.org/10.48550/arXiv.1505.00853

102. Clevert, D., Unterthiner, T., and Hochreiter, S., Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs), *arXiv: abs/1511.07289v5*, 2016. https://doi.org/10.48550/arXiv.1511.07289

103. Klambauer, G., Unterthiner, T., Mayr, A., and Hochreiter, S., Self-Normalizing Neural Networks, *arXiv: abs/1706.02515*, 2017. https://doi.org/10.48550/arXiv.1706.02515

104. Hendrycks, D. and Gimpel, K., Bridging Nonlinearities and Stochastic Regularizers with Gaussian Error Linear Units, *arXiv: abs/1606.08415*, 2016. https://doi.org/10.48550/arXiv.1606.08415

105. Ramachandran, P., Zoph, B., and Le, Q.V., Swish: a Self-Gated Activation Function, *arXiv: abs/1710.05941v2*, 2017. https://doi.org/10.48550/arXiv.1710.05941

106. Misra, D., Mish: A Self Regularized Non-Monotonic Neural Activation Function, *arXiv: abs/1908.08681*, 2019. https://doi.org/10.48550/arXiv.1908.08681

107. Lavi, B., Ullah, I., Fatan, M., and Rocha, A., Survey on Reliable Deep Learning-Based Person Re-identification Models: Are We There Yet?, *arXiv: abs/2005.00355*, 2020. https://doi.org/10.48550/arXiv.2005.00355

108. Rao, H. and Miao, C., SimMC: Simple Masked Contrastive Learning of Skeleton Representations for Unsupervised Person Re-identification, *arXiv: abs/ 2204.09826v1*, 2022. https://doi.org/10.48550/arXiv.2204.09826

109. Zheng, Y., Zhou, Y., Zhao, J., Jian, M., Yao, R., Liu, B., and Chen, Y., A Siamese Pedestrian Alignment Network for Person Re-identification, *Multim. Tools Appl.*, 2021, no. 80, pp. 33951–33970. https://doi.org/10.1007/s11042-021-11302-3

110. Zheng, M., Karanam, S., Wu, Z., and Radke, R.J., Re-identification with Consistent Attentive Siamese Networks, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5728–5737. https://doi.org/10.1109/CVPR.2019.00588

111. Hermans, A., Beyer, L., and Leibe, B., In Defense of the Triplet Loss for Person Re-Identification, *arXiv: abs/1703.07737*, 2017. https://doi.org/10.48550/arXiv.1703.07737

112. Organisciak, D., Riachy, C., Aslam, N., and Shum, H., Triplet Loss with Channel Attention for Person Re-identification, *J. WSCG*, 2019, no. 27. https://doi.org/10.24132/JWSCG.2019.27.2.9

113. Zhai, Y., Guo, X., Lu, Y., and Li, H., In Defense of the Classification Loss for Person Re-identification, *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019, pp. 1526–1535. https://doi.org/10.1109/CVPRW.2019.00194

114. Alex, D., Sami, Z., Banerjee, S., and Panda, S., Cluster Loss for Person Re-identification, *Proceedings of the 11th Indian Conference on Computer Vision, Graphics and Image Processing*, 2018. https://doi.org/10.1145/3293353.3293396

115. Bai, Z., Wang, Z., Wang, J., Hu, D., and Ding, E., Unsupervised Multi-Source Domain Adaptation for Person Re-identification, *Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 12909–12918. https://doi.org/10.1109/CVPR46437.2021.01272

116. Chen, H., Lagadec, B., and Bremond, F., Unsupervised Lifelong Person Re-identification via Contrastive Rehearsal, *arXiv: abs/2203.06468*, 2022. https://doi.org/10.48550/arXiv.2203.06468

117. Zhang, X., Li, D., Wang, Z., Wang, J., Ding, E., Shi, J., Zhang, Z., and Wang, J., Implicit Sample Extension for Unsupervised Person Re-identification, *arXiv: abs/2204.06892*, 2022. https://doi.org/10.48550/arXiv.2204.06892

118. Zhu, K., Guo, H., Yan, T., Zhu, Y., Wang, J., Tang, M., Part-Aware Self-Supervised Pre-Training for Person Re-identification, *arXiv: abs/2203.03931*, 2022. https://doi.org/10.48550/arXiv.2203.03931

119. Fu, D., Chen, D., Yang, H., Bao, J., Yuan, L., Zhang, L., Li, H., Wen, F., and Chen, D., Large-Scale Pre-training for Person Re-identification with Noisy Labels, *arXiv: abs/2203.16533*, 2022. https://doi.org/10.48550/arXiv.2203.16533

120. Cho, Y.H., Kim, W.J., Hong, S., and Yoon, S., Part-based Pseudo Label Refinement for Unsupervised Person Re-identification, *arXiv: abs/2203.14675*, 2022. https://doi.org/10.48550/arXiv.2203.14675

121. Chen, M., Wang, Z., and Zheng, F., Benchmarks for Corruption Invariant Person Re-identification, *arXiv: abs/2111.00880*, 2021. https://doi.org/10.48550/arXiv.2111.00880

122. Dataset and Code. URL: https://www.pkuvmc.com/dataset.html.

This paper was recommended for publication by O.P. Kuznetsov, a member of the Editorial Board