applied sciences

MDPI

Review

Re-Identification in Urban Scenarios: A Review of Tools and Methods

Hugo S. Oliveira [1], José J. M. Machado [2] and João Manuel R. S. Tavares [2,*]

[1] Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias, s/n, 4200-465 Porto, Portugal; hugo.soares@fe.up.pt
[2] Departamento de Engenharia Mecânica, Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias, s/n, 4200-465 Porto, Portugal; jjmm@fe.up.pt
* Correspondence: tavares@fe.up.pt; Tel.: +351-22-041-3472

Abstract: With the widespread use of surveillance image cameras and enhanced awareness of public security, objects, and persons Re-Identification (ReID), the task of recognizing objects in non-overlapping camera networks has attracted particular attention in computer vision and pattern recognition communities. Given an image or video of an object-of-interest (query), object identification aims to identify the object from images or video feed taken from different cameras. After many years of great effort, object ReID remains a notably challenging task. The main reason is that an object's appearance may dramatically change across camera views due to significant variations in illumination, poses or viewpoints, or even cluttered backgrounds. With the advent of Deep Neural Networks (DNN), there have been many proposals for different network architectures achieving high-performance levels. With the aim of identifying the most promising methods for ReID for future robust implementations, a review study is presented, mainly focusing on the person and multi-object ReID and auxiliary methods for image enhancement. Such methods are crucial for robust object ReID, while highlighting limitations of the identified methods. This is a very active field, evidenced by the dates of the publications found. However, most works use data from very different datasets and genres, which presents an obstacle to wide generalized DNN model training and usage. Although the model's performance has achieved satisfactory results on particular datasets, a particular trend was observed in the use of 3D Convolutional Neural Networks (CNN), attention mechanisms to capture object-relevant features, and generative adversarial training to overcome data limitations. However, there is still room for improvement, namely in using images from urban scenarios among anonymized images to comply with public privacy legislation. The main challenges that remain in the ReID field, and prospects for future research directions towards ReID in dense urban scenarios, are also discussed.

Keywords: person ReID; computer vision; deep neural networks; image enhancement

Citation: Oliveira, H.S.; Machado, J.J.M.; Tavares, J.M.R.S. Re-Identification in Urban Scenarios: A Review of Tools and Methods. *Appl. Sci.* **2021**, *11*, 10809. https://doi.org/10.3390/app112210809

Academic Editors: João M. F. Rodrigues, Pedro J. S. Cardoso and Marta Chinnici

Received: 7 September 2021
Accepted: 10 November 2021
Published: 16 November 2021

1. Introduction

The task of object ReID on image cameras has been studied for several years by the computer vision and pattern recognition communities [1], with the primary goal to ReID a query object among different cameras.

Multi-object ReID, based on a wide range of surveillance cameras, is nowadays a vital aspect in modern cities, to better understand city movement patterns among the different infrastructures [2], with the primary intention of rapidly mitigate abnormal situations, such as tracking car thieves, wanted persons, or even lost children.

This is still a challenging task, since an object's appearance may dramatically change across camera views due to the significant variations in illumination, poses or viewpoints, or even cluttered backgrounds [2] (Figure 1). According to the state-of-the-art research studies, existing object ReID methodologies can be divided into two main categories: image-based and video-based object ReID.

(**a**) ID 1 (**b**) ID 1 (occluded) (**c**) ID 2 (**d**) ID 2 (occluded)

Figure 1. Some common problems found in object ReID.

The former category focuses on matching a probe image of one object, with an image of the object with the same ID among gallery sets, which is mainly established based on image content analysis and matching. In contrast, the latter category focuses on matching two videos, exploiting different information, such as temporal, and motion-based information. A gallery corresponds to a collection of object images gathered from different perspectives over time. In both approaches, the pairs of objects to be matched are analogous. However, in real scenarios, object ReID needs to be conducted between the image and video. For example, given a picture of a criminal suspect, the police would like to quickly locate and track the suspect from hundreds of city surveillance videos. The ReID under this scenario is called image-to-video person ReID, where a probe image is searched in a gallery of videos acquired from different surveillance cameras.

Although videos contain more information, image-to-video ReID share the same challenges with image-based and video-based objects ReID, namely, similar appearance, low resolution, substantial variation in poses, occlusion, and different viewpoints. In addition, an extra difficulty resides on the match between two different datasets, one static and another dynamic, i.e., image and video, respectively.

Image and video are usually represented using different features. While only visual features can be obtained from a single image, both visual features and spatial–temporal features can be extracted from a video. Recently, CNN has shown potential for learning state-of-the-art image feature embedding [3,4] and Recurrent Neural Network (RNN) yields a promising performance in obtaining spatial–temporal features from videos [5,6].

In general, there are two major types of deep learning structures for object ReID; namely, verification models and identification models. Verification models take a pair of data as input and determine whether they belong to the same object or not, by leveraging weak ReID labels that can be regarded as a binary-class classification or similarity task [7]. In contrast, identification models aim at feature learning by treating object ReID as a multi-class classification task [4], but lack direct similarity measurement between input pairs. Due to their complementary advantages and limitations, the two models have been combined to improve the ReID performance [8]. However, in the image-to-video object ReID task, a cross-modal system, directly using DNN and the information provided by the target task still cannot perfectly bridge the "media gap", which means that representations of different datasets are inconsistent. Therefore, most of the current works directly rely on weights from pre-trained deep networks as the backbone to obtain initial values for the target model and initiate the pre-trained network structure to facilitate the training of the new deep model. Figure 2 depicts the common base architecture for person ReID.

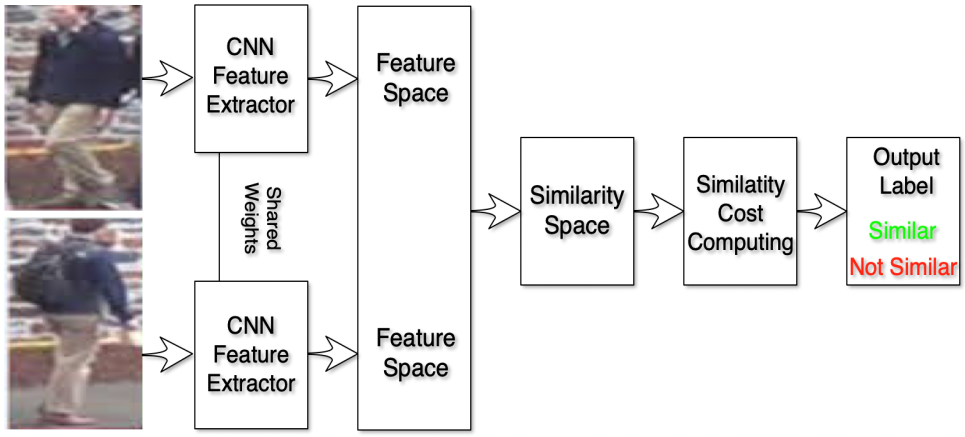

Figure 2. Common base architecture for person ReID.

With this problem in mind, we focus on identifying the promising techniques and methods for ReID that lead to a unified use in public urban scenarios, in adversarial and challenging conditions. This research article is organized toward identifying the main methodologies and key aspects. The following section presents an overview of the deep learning method for object ReID, and the most widely employed models proposed for ReID tasks. The main goal is to improve existing methods and derive new approaches to the ReID tasks. In the methods section, we detail the search used for this review article. The results section presents the main findings achieved with the selected works grouped by the person ReID and the multi-object ReID, with temporal constraints in consideration. The final section provides a critical discussion of the results and draws the conclusions.

2. Deep Learning for Object Re-Identification

A turning point in the history of machine learning and computer vision was reached by researchers of the University of Toronto, who proposed a new image classification approach and achieved excellent results in the ImageNet [9] competition [10]. The winning proposal, defined as AlexNet, consisted of a CNN composed of a set of stacked deep layers and dense layers that enabled the reduction of the error drastically. However, the first appearances of CNN dated from 1990, when Lecun et al. [11] proposed a CNN method addressing the task of hand-written digit recognition to alleviate the work of the postal office.

A CNN multi-layer contains at least one layer to perform convolution operations from image inputs, by using filters of kernels that are translated across and down the input matrix, to generate a feature representation map of the original image input. The characteristics of these filters can be widely different, and each one of them is composed of learnable parameters, updated through a gradient descent optimization scheme. The same layer can employ other filters. For the same part of the image, each filter produces a set of local responses, enabling correlation of specific pixel information with the content of the adjacent pixels. After the proposal by [10], many different network architectures were developed to address such a problem, each one with its inner characteristics, to name a few, the VGG [12], Inception [13], and ResNet [13] architectures. Many other network architectures were proposed, but in summary, they all share some building blocks.

In regard to the ReID task, the most used backbone model architecture is ResNet [14], due to its flexibility and ease of reusing and implementation to solve new problems. Most of the ReID tasks explore the use of pre-trained ResNet as backbones for feature extraction for the object ReID task.

ResNets can address the vanishing gradient problem when the networks are too deep, making the gradients quickly shrink to zero after several chain rule applications, leading toward "not updating" the weights and, therefore, harming of the learning process. With

ResNets, the gradients can flow directly through the skip connections backward from later layers to the initial filters.

ResNets can have different sizes, depending on how big each model layer is, and how many layers it contains. As an example, ResNet 34 [15] contains one convolution and pooling layer step, followed by four similar layers (Figure 3). Each layer follows the same pattern by performing 3×3 convolutions with a fixed feature map dimension, and by bypassing the input every two convolutions. A concern with the special characteristics of ResNet 34 is the fact that the width W and height H dimensions remain constant during the entire single layer. The size reduction is achieved by the stride size used in the convolution kernels, instead of the polling layers commonly used in other models.

Every layer of a ResNet is composed of several blocks, enabling it to go deeper. This deepness is achieved by increasing the number of operations within a block, while maintaining the total number of layers. Each operation comprises a convolution step, batch normalization, and a ReLU activation to a particular input; except for the last operation of the block, which does not contain a ReLU activation function.

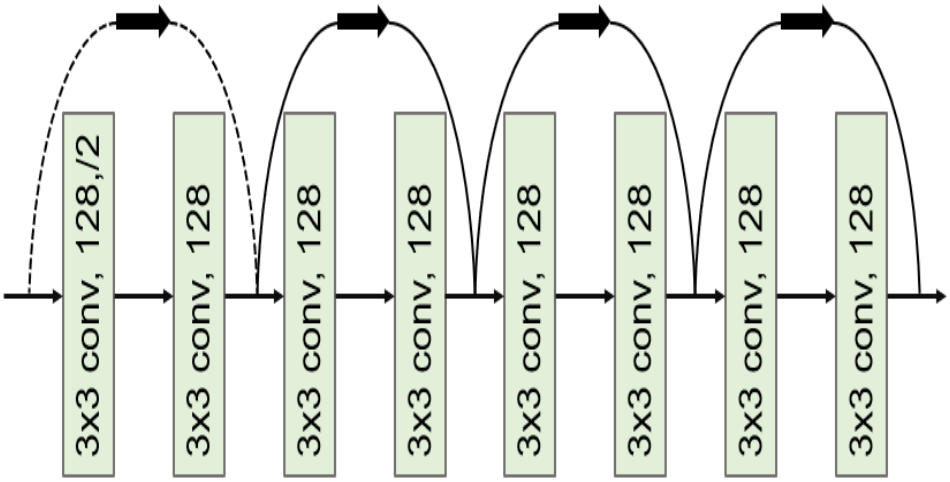

Figure 3. A common ResNet block architecture.

The aforementioned characteristics make ResNets particularly suitable for object ReID, since it enables shallow lower-level features to be reused at higher-level stages. This scheme allows exploring relevant information for ReID task from all layer feature maps, instead of relying only on more abstract summarized features provided by the higher layers.

3. Evaluation Metrics

Cumulative Matching Characteristic curve (CMC) is a common evaluation metric for person or object ReID methods. It can be considered a simple single-gallery-shot setting, where each gallery identity only has one instance. Given a probe image, an algorithm will rank the entire gallery sample according to the distances to the probe, with the CMC top-k accuracy given as:

$$Acc_k = \begin{cases} 1 & \text{if top-}k\text{ ranked gallery samples contain the query identity,} \\ 0 & \text{otherwise,} \end{cases}, \quad (1)$$

which is a shifted step function. The final Cumulative Matching Characteristics (CMC) curve is built by averaging the shifted step functions over all the queries.

Another commonly used metric is the mean Average Precision (mAP), which is very often employed on each image query, and defined as:

$$MAP = \frac{\sum_{q=1}^{Q} AveP(q)}{Q}, \quad (2)$$

where Q is the number of queries

4. Methodology

The current bibliographic analysis involved two steps: (a) collecting related work and (b) a detailed review and analysis of the gathered work.

The research methodology consisted of a keyword-based search of conference papers or journal articles from scientific databases; namely, IEEE Xplore and Science Direct, and from web scientific indexing services: Web of Science, Google Scholar, and arXiv. As search keywords, one performed the following query: ["deep learning"] AND ["reid" OR "re-identification" OR "person reid" OR "object reid"] The gathered information excluded filtered out articles referring to DNN, but did not apply to the ReID domain. Articles were initially identified from this process.

Restricting the search to articles in combination with connected papers, a web tool, the initial number of articles was lessened to 47. In the second step, the 21 articles selected from the previous step were analyzed one-by-one for the task of a person ReID, considering the following research questions:

1. What was the ReID- or multi-object ReID problem addressed?
2. What was the general approach and type of DNN-based models employed?
3. What were the datasets and models proposed by the authors? Were there any variations observed by the authors?
4. Was any pre-processing of data or data augmentation technique used?
5. What was the overall performance in (depending on the adopted metric)?
6. Did the authors test their model performances on different datasets?
7. Did the authors compare their approaches with other techniques? If yes, what was the difference in performance?

5. Person Re-Identification

Person ReID is the problem of matching the same individuals across multiple image cameras or across time within a single image camera. The computer vision and pattern recognition research communities have paid particular attention to it due to its relevance in many applications, such as video surveillance, human–computer interactions, robotics, and content-based video retrieval. However, despite years of effort, person ReID remains a challenging task for several reasons [16], such as variations in visual appearance and the ambient environment caused by different viewpoints from different cameras.

Significant changes in humans pose—across time and space—background clutter and occlusions; different individuals with similar appearances present difficulties to the ReID tasks. Moreover, with little or no visible image faces due to low image resolution, the exploitation of biometric and soft-biometric features for person ReID is limited. For the person ReID task, databases and different approaches have been proposed by several authors, which are summarized in the following sections.

5.1. Person Re-Identification Databases

The recognition of human attributes, such as gender and clothing types, has excellent prospects in real applications. However, the development of suitable benchmark datasets for attribute recognition remains lagged. Existing human attribute datasets are collected from various sources or from integrating pedestrian ReID datasets. Such heterogeneous collections pose a significant challenge in developing high-quality fine-grained attribute recognition algorithms.

Among the public databases that have been proposed for person ReID, some examples can be found in the open domain, such as the Richly Annotated Pedestrian (RAP) [17], which contains images gathered from real multi-camera surveillance scenarios with long-term collections, where data samples are annotated, not only with fine-grained human attributes, but also with environmental and contextual factors. RAP contains a total of

41,585 pedestrian image samples, each with 72 annotated attributes, as well as viewpoints, occlusions, and body part information.

Another example is the VIPeR [18] database, which contains 632 identities acquired from 2 cameras, forming a total of 1264 images. All images were manually annotated, with each image having a resolution of 128 × 48 pixels.

A summary of the public available databases for person ReID, with main characteristics, is presented in Table 1.

Table 1. Global overview of the found public available databases for person ReID.

Dataset	# Identities	# Cameras	# Images	Label Method	Size	Tracking Sequences
VIPeR	632	2	1264	Hand	128 × 48	NO
ETH1,2,3	853,528	1	8580	Hand	Vary	YES
QMUL iLIDS	119	2	476	Hand	Vary	NO
GRID	1025	8	1275	Hand	Vary	NO
CAVIAR4reid	72	2	1220	Hand	Vary	NO
3DPeS	192	8	1011	Hand	Vary	NO
PRID2011	934	2	24,541	Hand	128 × 64	YES
WARD	70	3	4786	Hand	128 × 48	YES
SAIVT-Softbio	152	8	64,472	Hand	Vary	YES
CUHK01	971	2	3884	Hand	160 × 60	NO
CUHK02	1816	10 (5 pairs)	7264	Hand	160 × 60	NO
CUHK03	1467	10 (5 pairs)	13,164	Hand	Hand/DPM	NO
RAiD	43	4	6920	Hand	128 × 64	NO
iLIDS-VID	300	2	42,495	Hand	Vary	YES
MPR Drone	84	1	-	ACF	Vary	NO
HDA Person Dataset	53	13	2976	Hand/ACF	Vary	YES
Shinpuhkan Dataset	24	16	-	Hand/ACF	128 × 48	YES
CASIA Gait Database B	124	11	-	Background subtraction	Vary	YES
Market-1501	1501	6	32,217	Hand/DPM	128 × 64	NO
PKU-reid	114	2	1824	Hand	128 × 64	NO
PRW	932	6	34,304	Hand	Vary	NO
Large scale person search	11,934	-	34,574	Hand	Vary	NO
MARS	1261	6	1,191,003	DPM+GMMCP	256 × 128	YES
DukeMTMC-reid	1812	8	36,441	Hand	Vary	NO
DukeMTMC4reid	1852	8	346,261	Doppia	Vary	NO
Airport	9651	6	39,902	ACF	128 × 64	NO
MSMT17	4101	15	126,441	Faster RCNN	Vary	NO
RPIfield	112	12	1,601,581	ACF	Vary	NO

DPM—deformable part models, ACF—pyramid features, GMMCP—generalized maximum multi clique.

Although many other datasets suitable for object ReID can be found, the ones listed in Table 1 are widely used by most of the authors as benchmarks for performance evaluation and comparison of the proposed works.

5.2. Person Re-Identification Methods

In this section, deep learning-based person ReID methods are grouped into four main categories, as represented in Figure 4, including methods for feature learning, sequence learning, generative learning, and deep learning metrics. These categories encompassed several methods, and they are discussed in the following in terms of their main aspects and experimental results.

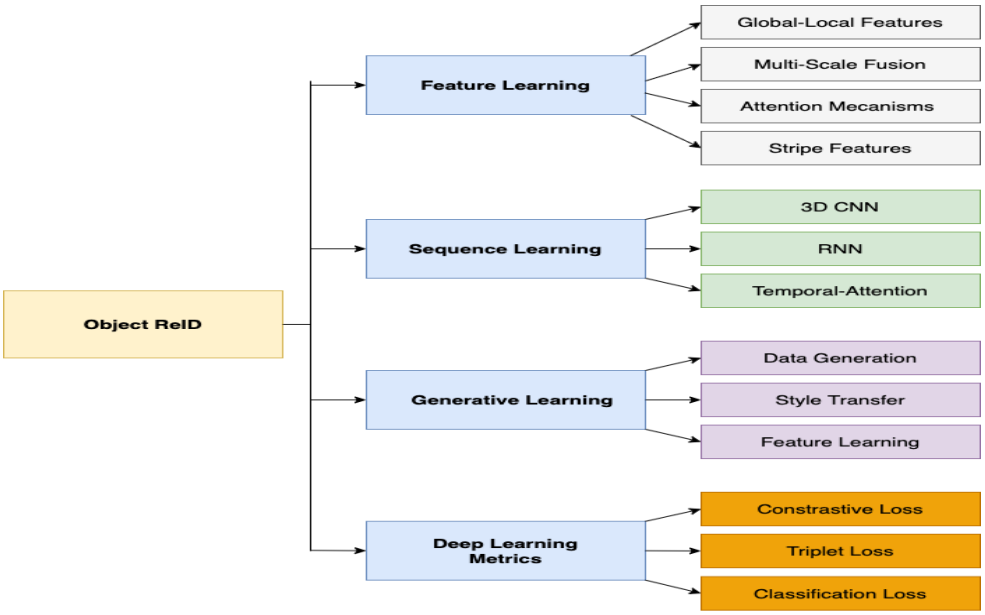

Figure 4. Deep learning-based person re-identification methods.

5.2.1. Feature Learning

Considering the features extracted from images, person ReID methods can explore the extracted feature from a global and local perspective, to better represent the object of interest. Global feature learning usually provides a single feature from the target image, making it difficult to capture detailed information of the person image. To overcome this problem, distinguishable local features are also used to capture subtle invariant features that are often combined in a fusion scheme.

An example of the use of a fusion scheme to combine global and local features for person ReID is presented by [19]. It consists of the formulation of an image-to-video person ReID as a classification-based information retrieval problem, where a model of "person appearance" is learned from the gallery images, and the identity of the interested person is determined by the probability that the corresponding probe image belongs to one of the gallery images.

To learn a model of person appearance, two kinds of features, Kernel Descriminator (KDES) [20] and CNN, are extracted from each person's image. Then, a Support Vector Machine (SVM) model is employed to learn the model. For ReID, three fusion schemes, early fusion, product rule, and query-adaptive late fusions, are proposed to aggregate the features. A ranking scheme in descending order of similarity is employed between the query image and the learned model to determine the most likely image pair. The work was evaluated in two benchmark datasets, CAVIAR4reid [21] and RAID [22], and by using CMC [23] as evaluation measures for person ReID. The model achieved a CMC of 96.11% when using the CAVIAR4reid dataset, and contained balanced cases combined with data augmentation during training, while employing late adaptive feature fusion. In contrast, when using the same dataset containing imbalanced cases and the same training and fusion methods, the model obtained an CMC of 93.33%. As for the RAID database, the model presented a CMC of 94.29% when using data augmentation during training and late feature fusion.

The same author refined in [24] its previous image-to-video person ReID framework [19] by adding two extra features: the Gaussian of Gaussian (GOG) [25] and learned features from Residual Neural Network (ResNet) [14]. The same feature fusion methodology employed in its previous work [19] was used, and the newly added features were evaluated using the same CAVIAR4reid and RAID databases. The model on CAVIAR4reid, containing balanced cases, and using late adaptive fusion combined with data augmentation during training attained an CMC of 86.39%. In contrast, on CAVIAR4reid and

with imbalanced cases, as well as the same fusion scheme and augmentation, the model achieved a CMC of 91.94%. Conducted experiments on the RAID database using the same fusion scheme and augmentation techniques achieved a CMC of 92.80%, proving the effectiveness of the newly added features. When compared to the previous work, the performance fell by approximately 2% on both datasets, mainly due to DNN co-adaptation that led to some degree of overfitting. While these works are somehow complementary, the suggested approach can overcome the difficulties in learning cross-scale features by learning multi-scale complementary features.

Jointly learning of local and global features using a CNN is proposed by [26], exploring advantages of jointly learning local and global features in a CNN, in order to obtain correlated local and global features in different contexts scenarios for person ReID. A deep two-branch CNN architecture was proposed, with one branch being responsible for learning localized features (local branch) and the second directed to learning global feature (global branch); the two components were not independent, but synergistically correlated and jointly learned, concurrently. The proposed joint learning multi-loss (JLML) CNN model consists of two branch CNN networks.

The local branch aims to learn the most discriminating local visual features of the surroundings of a people bounding box. In contrast, the second branch is responsible for learning the most discriminating global level features from the entire person's image. The joint learning scheme is employed for concurrently optimizing per-branch discriminative feature representations, and discovering correlated complementary information between local and global features by subjecting both local and global branches to the same identity label supervision. For sharing low-level features, a multi learning methodology is explored, as in the work proposed by [27]. An inter-branch common learning inter-permutation is to be shared on the first convolution layer, with the intuition that lower convolution layers capture low levels features, such as edged and corners that are common patterns to all images, while the complementary discriminative features from local and global representations are learned independently, and related to a given identity label. Moreover, a structure sparsity-induced regularization [28] is introduced to discourage the use of irrelevant features while encouraging discriminative features, to learn concurrently on both local and global contexts, and to maximize a shared identity matching objective. The final global feature representation corresponds to a sparsity measure with LASSO [29]. Cross-entropy is then used as the loss function for both global and local branches, in order to optimize person identity classification and a pairwise person ReID. Concerning the distance metrics for person ReID, a 1024-D deep feature representation is employed using only a generic distance metric without camera-pair specific distance learnable metrics. The models were evaluated on the CUHK01, VIPeR [30], CUHK03, Market-1501, and GRID datasets. On CHUK03, with the proposed method achieving a CMC of 83.2% Rank-1 using labeled objects and 80.6% with automatically detected objects. On Market-1501, the method achieved an CMC Rank-1 of 85.1% on single query, and 89.7% for multi-query. On CHUK01, the method achieved a CMC for Rank-1 of 91.2% when applying an 871/100 dataset split and 76.7% using a 486/485 split. For a GRID dataset, the method obtained a CMC Rank-1 of 37.5%. Finally, as to the VIPeR dataset, the method achieved a CMC Rank-1 of 50.2%. In most datasets, the proposed method surpassed the compared state-of-the-art approaches except for the VIPeR dataset. The combination of local and global features potentiates model generalization, avoiding overfitting to image-specific features.

In [31], a novel deep ReID CNN is proposed for omni-scale feature learning (OSNet). The model is based on residual blocks composed of multiple convolutional streams, with each detecting feature at a certain scale. A novel unified aggregation gate is then introduced to dynamically fuse multi-scale features with output-dependent channel-wise weights. To efficiently learn spatial-channel correlations and to avoid over-fitting, pointwise and depth-wise convolutions are used. Depth-wise separable convolutions are also adopted to reduce the number of parameters. Person matching is based on the $\zeta 2$ distance from 512-D feature vectors extracted from the last layer. Two OSNet models were trained for

comparison proposes, with the first model being trained from scratch during 350 epochs using Stochastic Gradient Descent (SGD) as an optimizer. In contrast, a second model was fine-tuned using ImageNet [9] weights and AMSGrad optimizer [32]. The images were resized to 256 × 128, and common data augmentation techniques, such as random flip, random crop, and random image patch, and random erase [33], were applied. Model experiments were conducted on six widely used person ReID datasets: Market-1501, CUHK03, DukeMTMC-reid, MSMT17, VIPeR, and GRID datasets. Collected results showed that the model achieved overall supremacy when compared with most of the state-of-art methods, attaining a CMC Rank-1 of 94.8% on Market-1501, 72.3% on CUHK03, 88.6% on Duke, and 78.7% on MSMT17. The proposed fusion scheme shows the effectiveness of the omni-scale features in different scales to comply with a large range of possible viewpoints from image pairs.

To improve the capabilities of attention mechanisms and obtain fine detailed features for person ReID, ref [34] proposes a feature refinement process in combination with filter network, by weakening the high response features and eliminating the interference raised by the background information. The model includes a network formed by the weaken feature convolution blocks based on ResNet, in combination with a multi-branch scheme. An attention mechanism is also set in place to act as an attention feature map in the convolution module, with higher values corresponding to regions where the model has paid more attention.

Extensive experiments were conducted in the Market-1501, DukeMTMC-reID, CUHK03 and MSMT17 person ReID benchmarks datasets, with the proposed model achieving a mAP of 94.2% in Market-1501.

An automatic search for a CNN architecture, specifically suited for the ReID problematic was proposed by [35], and denoted as Auto-ID. The method is based on the neural architecture search (NAS) [36] to automate the process of architecture design without human effort, directed to the task of ReID, by using a retrieval search-based algorithm. This design enables to obtain more optimal architectures that make the best use of human body structure information for person ReID, eliminating human expert efforts in the manual design of CNN models for the task. The model starts by integrating structural body cues into the input tensors, and then by vertically splitting the input feature tensor into four body part features, averaging each tensor part into a vector, and transforming each of the tensors into a new part feature vector using a linear layer. The obtained part vectors interact between them via a self-attention mechanism, enabling each part of the vectors to incorporate more specific body part information. Each obtained part vector is repeated and concatenated to recover the original spatial shape as the input tensor. Finally, the formed global feature tensor is fused with the original input tensor, using a one-by-one convolutional layer. A class-balance data sampler to equal the sample batch data for the triplet loss is used to overcome the original sensitivity of triplet loss to the batch data. This sampler first samples some identities, uniformly, and then, for each identity, it randomly samples the same number of images. To better explore the benefits from the cross-entropy and triplet losses, a mixture retrieval loss between sample loss and triplet loss is considered. Experiments were conducted on the Market-1501, CUHK03, and MSMT17 public databases, with the proposed Auto-ID model achieving a CMC rank-1 of 95.4% on Market-1501, when using the re-ranking technique, 77.9% on CUHK03 with labeled examples, and 73.3% on detected ones. On MSMT17, the model achieved a CMC Rank-1 of 78.2%. The proposal work enables increasing the ReID performance by taking into consideration attention mechanisms combined with triplet loss methods.

A dropout technique was proposed by [4] for learning deep feature representations from multiple domains with CNN. Multi-domain learning is frequently directed toward solving the problem of using datasets across different domains simultaneously, by using all data they provide for the task of multi-domain learning, while robustly handling data domain discrepancies. The central problem in multiple learning relies on the fact that samples from the same domain follow the same underlying data distribution, which

degrades the model performance since some neurons may gain focus on some domain representation, while discarding the remainder domains, and leading to a bad model generalization. To overcome this problem, a domain guided dropout algorithm was proposed to avoid the model to learn only from domains where samples follow the same underlying data distribution. The CNN models are trained from scratch using all data domains using a single Softmax loss, creating a solid baseline. For each domain, a forward pass was performed on all data domains, and the impact that each neuron had on the objective function was quantified. After a few epochs of training, the standard network dropout layer was replaced by the proposed domain guided dropout layer, and the training process continued for several epochs to guide the neurons to the effective domain, enabling the CNN to learn more discriminative features for all of them. Experiments were conducted using the CUHK03, CUHK01, CUHK03, and PRID datasets, and compared with state-of-the-art methods using the CMC metrics. The proposed method outperformed the studied state-of-the-art methods, achieving a CMC Rank-1 of 75.3% on CUHK03, 66.6% on CUHK01, and 64.0% on PRID. The introduced domain dropout layer acts as regularized mechanism, avoiding the neurons co-adaptation to specific domains, reducing overfitting that degraded the ReID performance.

Attention mechanism has gained relevance in recent years, showing good performance in many fields, and it is often used as a local feature learning mechanism, which is useful for the task of ReID.

One example of an attention mechanism for deep learning networks was presented by [37] to provide simultaneously masks-free and foreground-focused samples for the inference phase. The main objective was to generate synthetic data that are composed of interleaved segments gathered from the original learning image set, while using class information only from specific segments. The proposed augmentation technique was evaluated using a baseline method proposed by [38], based on a deep learning-based classification framework using the ResNet-50 as a feature extractor, with weights initialized on ImageNet [9], along with a bag of tricks, known to be particularly effective for person ReID tasks. Since the richly annotated pedestrian (RAP) [17] dataset does not provide human body segmentation annotations, human binary segmentation masks were extracted using Mask-RCNN [39] to obtain the human body segmentation binary masks. Afterwards, fake images were generated to enlarge the dataset. For a matter of performance evaluation of the augmentation method, the default parameter settings detailed on the official project and the same weights were reused without modifications. The models were evaluated on two different loss schemes: Softmax and Triplet, with the results being slightly better when using triple loss on the RAP dataset. Accurately, the model presented a mAP Rank-1 of 62.9% when considering the upper body part, and 65.7% for the full body. The proposed augmentation technique enabled an increase in the performance of the baseline method by almost 20%.

Another attention mechanism based in Deep Learning (DL) models was proposed by [40] to overcome the problem of learning fine-grained pedestrian features that are useful for pedestrians ReID. A self-denominated HydraPlus-Net (HP-net), which multi-directionally feeds the multi-level attention maps to different feature maps and to different feature layers, is used. The method enables the model to capture multiple attention from a low-level to a semantic level by exploring the multi-scale selectiveness of attentive features, to enrich the final feature representations for the pedestrian image. The proposed approach was evaluated on three publicly standard datasets: CUHK03, VIPeR, and Market-1501 datasets. The model achieved a CMC Top-1 of 91.8 % on CUHK03 , 56.6% on ViPer, and 76.9% on Market-1501. The proposed method achieved top performance results, proving the effectiveness of attention mechanisms to capture relevant features maps for ReID tasks.

Attention mechanisms are also explored by [41], by introducing a bilateral complementary network (BiCnet), formed by a two-branch scheme; the first operating in the original image resolution, and the second called context branch, operating in downsampled resolution to capture long-range context. A specific attention mechanism, called diverse attention

operation, was added to enforce consecutive frames to focus on different body characteristics regarding each identity. The mining of spatial clues was carried by a temporal kernel selection to jointly combine the short- and long-term temporal relations. Exhaustive experiments were conducted in the MARS, DukeMTMC-VideoReID, and LS-VID datasets, with the model obtaining a mAP of 0.860 on MARS dataset.

Attention mechanisms used to obtain more robust salient features from images are proposed by [42]. Since complex backgrounds can generate salient features that can degrade the performance of the ReID task, a joint weak saliency, in combination with an aware attention mechanism, is set in place to obtain refined global features, while weakening some of the saliency features. Similar to [34], this model employs a ResNet scheme from the weekend saliency block, where an attention mechanism is set in place, and final results of both processes are fused together to form the final feature. The performance of the method is evaluated using the Market-1501 and DukeMTMC-ReID datasets, with the method achieving a mAP of 89.2% in the Market-1501 dataset.

A performance evaluation of the reviewed methods for person ReID that explore Feature learning is presented in Table 2.

Table 2. Performance evaluation of the reviewed person ReID using feature learning methods.

Cat	Ref.	Main Technique(s)	# Data Success	Pros/Cons
Fusion	[19]	KDE and CNN features, late fusion, SVM model	CAVIAR4reid CMC 0.933	Robust, simple
	[24]	GOG and ResNet features, Data augmentation	CAVIAR4reid CMC 0.919	Simple, lack train data
	[26]	Joint learning multi-loss, two-branch CNN	CHUK03, CMC 0.832	Simple, efficient
	[31]	Residual blocks, multi-scale feature	Market-1501 CMC 0.948	Hard to train, can over fit
	[34]	Weaken feature convolution, ResNet	Market-1501 mAP 0.942	Robust, reusable
Strip	[35]	Neural architecture, search (NAS)	Market-1501 CMC 0.954	Hard to train, complex, not reusable
Drop	[4]	Modified dropout layer	CUHK03 CMC 0.666	Easy train, data domain, problematic
Attention	[37]	ResNet-50 as feature extractor, attention mechanism	RAP mAP 0.862	Simple to replicate, reusable
	[40]	ResNet-50 as feature extractor, multi-directional and level attention maps	CUHK03 CMC 0.918	Complex, not reusable state-of-the-art
	[41]	Two Branch, multi-scale and attention maps	MARS mAP 0.860	Complex, reusable state-of-the-art
	[42]	Attention, saliency maps ResNet	Market-1501 mAP 0.892	Complex, generalizes well, state-of-the-art

CMC—cumulative matching characteristic (higher the better), mAP—mean average precision (higher the better), all measures range: $[0.0, 1.0]$.

A performance comparison among the described works is depicted in Figure 5.

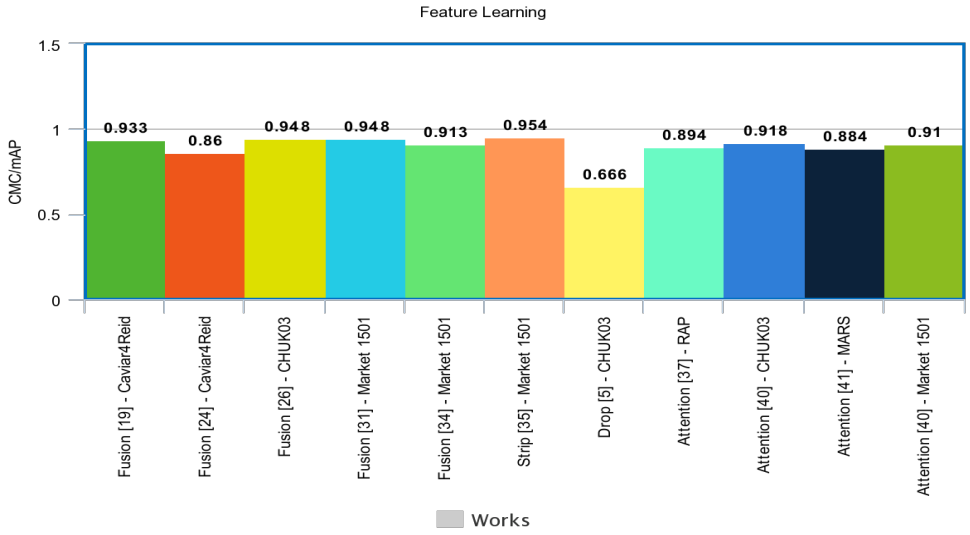

Figure 5. Performance summary of the evaluated feature learning methods.

Although the Fusion scheme of local and global features is a reasonable approach to explore, the use of attention mechanisms enables leveraging the performance of person ReID, by capturing important aspects that are relevant to objectives set-in place.

5.2.2. Deep Learning Metrics

Deep learning metrics is one commonly used strategy that aims to learn the dissimilarity or similarity between two given objects. The main objective is to learn a projection mapping from the original image into the embedding feature space, enabling to determine the degree of similarity between two-person images. This helps with learning the discriminate features by the design of loss-specific functions for the DNN model.

One standard metric corresponds to the contrast loss, which enables quantifying the similarity or dissimilarity between pairs of data, commonly used on the training of Siamese Networks, with the function expressed as:

$$L_c = yd(x_a - x_b)^2 + (1-y)\left[m - d(x_1 - x_b)^2\right]_+, \quad (3)$$

where $[\cdot]_+ = max(0, x)$, with x_a and x_b corresponds to the two image pairs of the Siamese Network, and a distance metric $d(x_a, x_b)$, usually the Euclidean distance, quantifies the degree of similarity among the pairs, m corresponds to a training parameter, and y is the corresponding matching label. When $y = 1$ the two input mages belong to the same ID (positive sample pair); on the other hand, when $y = 0$, it reflects the opposite case (negative sample pair).

One example of a Siamese Network for person ReID is explored by [43], by applying different distance metrics to corresponding feature maps. Defined as MSP-CNN, the approach starts by using image pairs as network input, with all images going through the same share-weighted deep CNN network, formed by small convolution filter layers followed by a simple inception module [44]. To attain the distinct characteristics from the diverse feature maps, similarity constraints are applied to both low-level and high-level feature maps during the training stage to effectively learn discriminative feature representations at different levels. The objective function was designed to emphasize low-level features that are frequently related to schoolbags, T-shirts, and higher-level special textures, which are shared among persons from the same personality to propagate those relevant features to the upper layers. At the higher-level feature maps, a Euclidean distance after L2normalization [45] is used to represent the abstract global similarities. The approach enables the CNN to extract robust feature representations without any complicated distance metric to be learned in the process, contrary to those found in more traditional hand-crafted systems. This enables easily incorporating constraints, forming a

unified multi-task network with similar constraints. Model evaluations were conducted using the CUHK03 [46] and Market-1501 [47] datasets, and the small CUHK01 dataset [48], being evaluated by the use of the CMC metric. Results on the CUHK03 dataset show that the proposed model achieved an CMC Rank-1 accuracy of 85.7% when using manual hand-labeled object boxes, and 83.6% when using a deformable parts model detector for object extraction. While the proposed model obtains competitive results, it still requires some degree of human annotation to achieve good performance.

A novel filter pairing neural network (FPNN) was proposed by [46], which is composed of six layers to jointly handle misalignment, photometric and geometric transforms, occlusions, and background clutter in person ReID tasks. The network was initially composed by a convolution and max-pooling layer that operated on two pair of RGB or Lab Color space (LAB) images from different cameras, generating the responses from local patches as local features. Each feature map is partitioned into $H_1 \times W_1$ stripe sub-regions, and only the maximum response in each sub-region is taken into account, with the max-pooling layer outputting a $H_1 \times W_1 \times K_1$ feature map. The computed feature maps are processed by a patch matching layer to match the filter responses from local patches across the different views. Considering that each input image contains M horizontal patches, these image patches are only compared with the corresponding stripe from the other pair images, forming displacement matrices to encode the spatial patterns of each matching patch under the different features representations. The patch matching is then further refined by dividing the patch displacement matrices into T groups, and within each group, a max out-grouping layer is used. Only prominent feature activations are passed to the next layer, allowing each feature to be represented by multiple redundant channels, enabling the modeling of a mixture of photometric transforms. For body parts, convolution and a max-polling layer are added to the patch displacement matrices to obtain the displacement matrices of body parts on a larger scale. For the final identity recognition, a Softmax function is used to measure the degree of similarity between two input person images, given the global geometric transforms detected on the previous layer. During the model training, several conventional techniques, such as dropout [49], data augmentation, and bootstrap were employed. The model was evaluated using the constructed CUHK03 [46], and the results were collected and compared with other state-of-the-art methods using the CMC metric, with the proposed method achieving a 20.65% when considering Rank-1 rate. The partial region patch makes the method suitable for partial pairs matching, enabling refining similarity metrics, according to the context of the image.

In contrast, an unsupervised learning approach is explored by [50], where an unsupervised incremental learning algorithm, denominated TFusion, aided by the transfer learning of pedestrian spatial–temporal patterns from an unlabeled target domain, is used for person ReID. The algorithm transfers the visual classifier trained on a small labeled source dataset to the unlabeled target dataset and learns pedestrian spatial–temporal patterns. A Bayesian fusion model is then used to combine the learned spatial–temporal patterns with extracted visual features to create an improved classifier. A learning-to-rank based on the mutual promotion procedure is used to optimize the classifiers based on the unlabeled data domain incrementally. The proposed framework explores a Siamese Network scheme [8] based on two ResNet50 [14] CNN pre-trained networks to extract visual features from different pair object images. The outputs of the Siamese Network are flattened into two one-dimensional vectors, with the model predicting the identities of each of the input pair images and their similarity score by using cosine similarity. A spatial–temporal pattern learning is formulated considering pedestrian patterns among different cameras and corresponding time intervals of objects that were previously considered similar by the model. In the last stage, the Bayesian fusion model combines the visual features with the spatial–temporal features to achieve a composite similarity score of the given pair of images. Exploring the fact that the Bayesian fusion model is based on the Bayes theorem, it is possible to access the likelihood of the scores of each image that belong to the same object. Model experiments were conducted using the GRID [51], Market-1501 [47], CUHK01 [48], and

VIPeR [30] datasets using a cross-validation strategy, where one of the datasets is selected as the source and another one as the target dataset to test, enabling performing cross-dataset person ReID evaluation. The results show that the proposed TFusion model achieved an CMC in Rank-1 of 64.10%, when using the GRID dataset, and 73.13% when using the Market-1501 dataset. The results, when compared to the work of [24] using fusion schemes, are much lower, mostly since the author relies only on ResNet50 as backbones, and uses simple cosine similarity metrics that may not capture other image similarity domains.

A deep convolutional network with layers, specially designed to address the problem of ReID, was proposed by [52], by outputting a similarity value, indicating whether the two input images are from the same person. The network encompasses two layers: the neighborhood difference layer for comparing convolutional image features from each patch and a subsequent layer where features are summarized. For the extracted features to be comparable across the two images in later layers, the first two layers are set to perform a tied convolution, with weights shared across the two views, ensuring that the same filters are used in both image pairs to compute the corresponding features. Two tied convolution layers enable providing a set of feature maps for each input image, from where relations between the two views are learned and supplied to a cross-input neighborhood difference layer, to compute the differences between the two views around a neighborhood of each feature location, generating a set of 25 neighborhood difference maps. Subsequently, a patch summary layer summarizes these neighborhood difference maps by producing a holistic representation of the differences in each view 5×5. The learning of the spatial relationships across neighborhood differences is achieved by employing a convolution layer using 25 filters of size 3×3 with stride 1, and the resulting features are passed through a max-pooling kernel to reduce the height and width. Finally, a fully connected layer captures the relations by combining information from patches that are far from each other and with a Softmax layer to output the similarity of both images. The proposed methods were evaluated using the CUHK03, CUHK01, and VIPeR datasets and the CMC curve. The model achieved a CMC Rank-1 accuracy of 54.74% on CUHK03-labeled and 44.96% on CUHK03-detected; in VIPeR, the method obtained a 34.81% Rank-1 accuracy, while in CUHK01, it achieved a Rank-1 recognition rate of 65%. The use of shared weights ensures fair feature selection; however, some specific image domains can be neglected, harming the ReID process.

Triplet loss is one of the most widely used deep learning metrics used in person ReID problems, aiming to minimize the intra-class distance while maximizing the intra- to intra-class distance of the given samples. The triplet loss can be expressed as:

$$L_{trip} = \left[m + d(x_a, x_p) - d(x_a, x_n) \right]_+. \quad (4)$$

When compared to contrast loss, the input of the triplet Loss consists of three images, with each triplet set containing a pair of a positive sample x_p, a negative sample x_n, with a corresponding anchor image x_a. x_a and x_p correspond to images with the same ID, while the pair x_a and x_n to images with different IDs. During model training, the distance between the same ID pairs x_a and x_p is minimized, while the distance between different ID pairs x_a and x_n is set apart. To increase the performance during training, a combination of classification loss and triplet loss [53] is used, enabling learning discriminatory features.

One example of the use of triplet loss is in the work by [53], which also explores a modification of the triplet loss, defined as TriNet, to perform end-to-end deep metric learning to tackle the person ReID problematic. Triplet loss has been proposed previously by [54], and vastly explored on FaceNet [45], where a CNN is used to learn an embedding for faces. Two approaches were explored in the proposed work, with the first being based on a ResNet-50 [14] architecture and the weights provided from the ImageNet pre-training procedure, with the last layer being discarded and replaced by two Fully Connected (FC) layers. The first contains 1024 units, followed by batch normalization [55] and Rectified Linear Unit (ReLU) [56], and the second goes down to 128 units, forming the final embedding dimension. The second approach consists of a network trained from scratch,

denoted as LuNet, which follows the style of [57], but uses leaky ReLU [58], nonlinearities, multiple 3×3 max-polling's with stride 2, and omits the final average pooling of feature-maps in favor of a channel-reducing final res-block. Distinct training parameters were used to train both networks, being the TriNet trained with the modified batch triplet loss, by setting the batch size to 72 to circumvent memory issues, due to the number of parameters (25.74 M); while in the second network, LuNet contains 5.00 Million parameters, and was trained using a large batch size (128). Model experiments were conducted on the CUHK03, Market-1501, and MARS [59] datasets, and several triplet variations and model comparisons were evaluated. One advantage of the use of triplet loss is that it allows performing end-to-end learning between the input image and the target embedding space, directly optimizing the network for the final task. Person comparison is performed by computing the Euclidean distance of their embeddings. The proposed pre-trained TriNet achieved a CMC of 89.63% when using CUHK03 and labeled box sets, and of 87.58% when automatically detecting box sets. Additionally, the proposed LuNet achieved competitive performance. One of the main disadvantages of using pre-trained networks is the flexibility to try out new advances in deep learning or to make task-specific changes in a network.

Once traditional triplet loss randomly selects three images from the training set during training, on many occasions, the sample combinations may evidence the lack of complex sample combinations that correspond to the more difficult cases, degrading the generalization capabilities of the model. To overcome this, many researchers improved the triplet loss to mine hard samples.

A multi-channel parts-based on CNN under a modified triplet framework for person ReID was proposed by [60]. The CNN network consists of multiple channels to jointly learn both the global full-body and local body-part features of the input image. The person ReID modeled network is trained using a modified triplet loss function to pull the feature instances of the same person together, while setting those instances further, corresponding to different persons in the learned feature space. Three CNN with the same sets of weights and biases are used, with the triplets from image I_1 space being mapped into a learning feature space from I_i. The multi-channel CNN model is composed of the following distinct layers: one global convolution layer, one full-body convolution layer, four body-part convolution layers, five channel-wise full connection layers, and one network-wise full connection layer. A global convolution layer acts as the first layer of the CNN network. It is split into four equal parts, with each part forming the first layer of the independent body-part channel, responsible for learning features of the corresponding body parts. Moreover, a full-body channel that considers the entire global convolution layer as its first layer is added to learn the global full-body features of the persons. The four body-part channels, together with the full-body channel, constitute five independent channels that are trained separately from each other. The final outputs of the channel-wise full connection layers, from the five separate channels, are concatenated into one vector and fed into the final fully connected layer. For model training, a novel data augmentation technique is performed by cropping the center of each image region of 80×230 pixels and introducing a small random perturbation to augment the training data, a technique close to [61]. Experiments were conducted in the VIPeR, i-LIDS [62] and PRID2011 [63] datasets. The models were assessed using the CMC metric for quantitative evaluation on each of the referred datasets, and several model variations were also evaluated, with the best-proposed model variation formed by the full version of the proposed multi-channel CNN model trained with the modified triplet loss function achieving a CMC Rank-1 accuracy of 60.4% on i-LIDS, a 22.0% CMC Rank-1 on PRID2011, 47.8% on VIPeR, and 53.7% on CUHK01. The proposed method showed promising performances in competitive scenarios and positive ReID person in partially occluded environments.

A performance evaluation of the reviewed methods for person ReID using deep learning metrics is given in Table 3.

Table 3. Performance evaluation of the reviewed person ReID using deep metric methods.

Cat	Ref.	Main Technique(s)	# Data Success	Pros/Cons
Contrast Loss	[43]	Unsupervised, ResNet50 features, Siamese networks, Bayesian fusion	CUHK03, CMC 0.857	Good dataset generalization, Cross domains, Complex
	[46]	Filter pairing neural network	CHUK03, CMC 0.206	Bad performance Complex, Not robust
	[50]	Unsupervised, ResNet50 features, Siamese networks, Bayesian fusion, spatial–temporal model	CUHK03, CMC 0.857	Good dataset generalization, Cross domains, Complex
	[52]	Siamese networks, Tied convolution	CUHK03 CMC 0.547	Simple, reusable
Triplet Loss	[53]	Triplet loss, pre-trained ResNet	CUHK03 CMC 0.876	Simple to replicate, architecture poses restraints
	[60]	Three CNN with shared weights, modified triplet loss	i-LIDS CMC 0.604	Simple train Scalable Efficient

CMC—Cumulative matching characteristic (higher, the better), mAP—-mean average precision (higher, the better), all measures range: [0.0, 1.0].

A performance comparison among the described works is depicted in Figure 6.

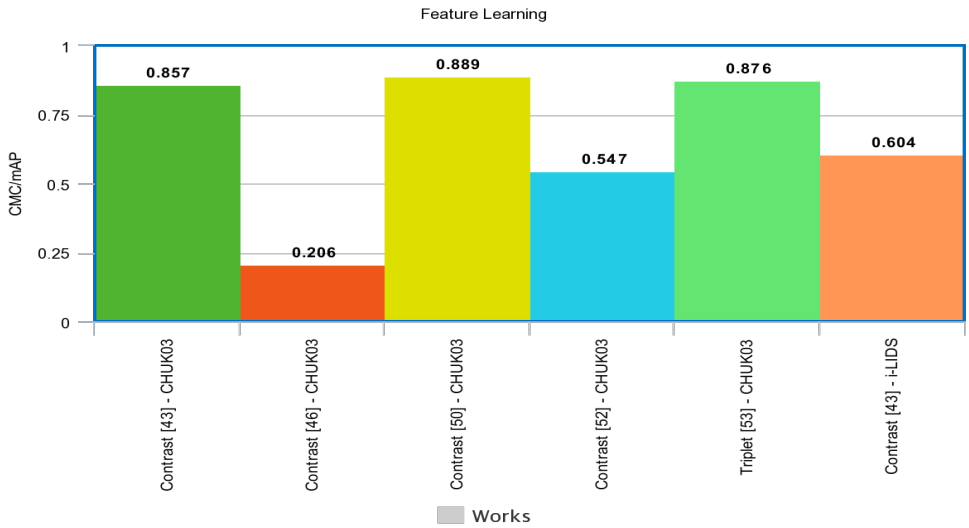

Figure 6. Performance summary of the evaluated deep learning metric methods.

Concerning the use of deep learning metrics, the use of contrast loss and triplet loss are the most common and usual methods employed in the person ReID task. This preference is mainly related to the simplicity of the methods, without major modifications to the existing pre-trained backbone, with the Siamese scheme being extremely suitable for pair image comparison.

5.2.3. Sequence Learning for ReID

One common approach to capture the spatial–temporal cues for the task of ReID is to explore a sequence of videos or a small set of images to train RNN models that can be

employed in the person ReID task. Many approaches explore 3D CNN to capture temporal and spatial features simultaneously [64].

The 3D CNN, in combination with a non-local attention mechanism, was proposed by [64] for person ReID, inspired by video action recognition models that involve the identification of different actions from video tracks. For the task, 3D convolutions on video volume, instead of using 2D convolutions across frames, are used to extract spatial and temporal features simultaneously. To handle misalignments, a non-local block is employed to capture spatial–temporal long-range dependencies, resulting in a network being able to learn useful spatial–temporal information as a weighted sum of the features in all space and temporal positions from the input feature map. Triplet loss function with hard mining proposed by [53] and a Softmax cross-entropy loss function with label smoothing regularization are employed to train the network. As for the network, 3D convolutions are replaced with two consecutive convolution layers, one one-dimensional (1D) convolution layer acting purely on the temporal axis, followed by a two-dimensional (2D) convolution layer to learn spatial features on the residual block. The modified 3D ResNet-50 is pre-trained on kinetics [65] to enhance the generalization performance of the model, and the final classification layer is replaced to output person identity. Experiments were performed using three datasets, namely, the iLIDS-VID, PRID-2011, and MARS datasets, and the results were compared with the ones of several state-of-the-art methods and of an established baseline model, which corresponds to a ResNet50 trained with Softmax cross-entropy loss and triplet with hard mining on an image-based person ReID. The proposed framework showed competitive results, outperforming several state-of-the-art approaches by a large margin on multiple metrics, attaining a mAP of 84.3% on the MARS dataset.

In [66] is proposed a two-stream convolution network to extract spatial and temporal cues for video-based person ReID. A temporal stream network was built by inserting several multi-scale 3D (M3D) convolution layers into a 2D CNN network. The M3D convolution network introduces a fraction of parameters into the 2D CNN to gain the ability of multi-scale temporal feature learning. In addition, a temporal stream was included using residual attention layers to refine the temporal features further. The jointly learning of spatial–temporal attention masks in a residual manner enables the identification of the discriminative spatial regions and temporal cues. Model evaluations were performed on three widely used benchmarks datasets: the MARS, PRID2011, and iLIDS-VID datasets, with the proposed model obtaining a mAP on 0.740 on the MARS dataset.

RNN in the form of Long Short-Term Memory (LSTM) of Gated Recurrent Unit (GRU) are commonly employed to capture temporal or spatial features. In ReID tasks, often the use of RNN is applied into sequences of images or video frames to capture spatial features extracted from CNN.

One example of use of LSTM is presented by [6], to progressive aggregate frame-wise human region representation at each frame extracted from the Local Binary Patterns (LBPs) detector, yielding a sequence feature representation. LSTM enables remembering and propagating previously accumulated representative features while forgetting irrelevant ones. The proposed RNN acts as a feature aggregation, generating highly discriminating sequence-level object representations. The evaluations of the models were conducted using the iLIDS-VID and PRID 2011 datasets, obtaining a Rank 1 of 49.3 on iLIDS-VID.

A RNN to jointly use spatial and temporal features is presented by [67], enabling exploring all relevant information useful for the person ReID task. The method encompasses a temporal attention mechanism to automatically pick the most discriminating features in a specific frame obtained from a CNN, while integrating surrounding information. Experiments were carried out using the iLIDS-VID, PRID 2011, and MARS datasets, with the proposed model achieving a Rank-1 of 70.6 on MARS.

A performance evaluation of the reviewed methods for person ReID using sequence learning is given in Table 4.

Table 4. Performance evaluation of the reviewed person ReID Sequence methods.

Cat	Ref.	Main Technique(s)	# Data Success	Pros/Cons
3D CNN	[64]	3D CNN, Attention, triple loss	MARS mAP 0.834	Simple to replicate, reusable state-of-the-art
	[66]	3D- Two stream CNN, Residual attention	MARS mAP 0.740	Replicable, SOTA
RNN	[6]	LSTM, LBP features	iLIDS-VID Acc1 0.493	Not robust, Old
	[67]	LSTM, CNN features	MARS Rank1 0.706	Simple, replicable

CMC—cumulative matching characteristic (higher, the better), mAP—mean average precision (higher, the better), all measures range: [0.0, 1.0].

A performance comparison among the described works is presented, as depicted in Figure 7.

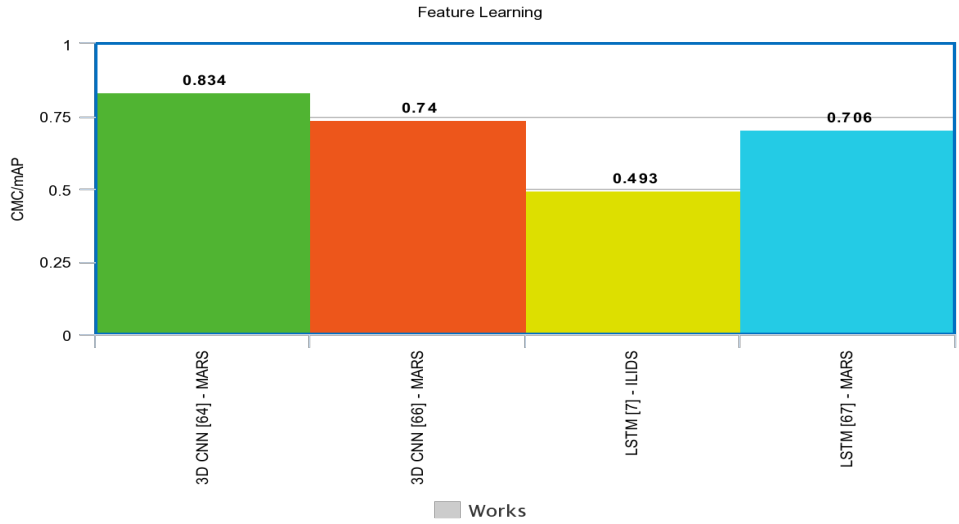

Figure 7. Performance summary of the evaluated sequence model methods.

While the use of RNN networks seems to e a natural choice to capture relevant features for the person ReID task, recently, authors have relied on 3D CNN to automatically capture these dependencies, obtaining models that are less complex to train and achieve superior performance when compared with traditional RNN, such as LSTM or GRU.

5.2.4. Generative Learning for ReID

One of the main difficulties in ReID tasks is the small diversity of images from the same object with different surroundings and conditions, which poses difficulties for models to generalize well to unseen image contexts. Among the identified datasets, one of their main limitations concerns the uniform illumination, and similar image poses. The use of generative learning, mainly by Generative Adversarial Network (GAN) to increase the amount of training data while presenting the model with more complicated cases, is one common practice in the field of computer vision, and was recently used in ReID tasks. GAN commonly employ the use of two neural networks that compete against each other to become more accurate and output more precise predictions. They are composed of the generator responsible for generating artificial data that can be mapped into the unknown training data distribution. In contrast, a discriminator tries to identify with the generator outputs that correspond to the real examples or the generated ones. The GAN

training is performed in an adversarial way, improving the capabilities of the generator, simultaneously representing the natural data distribution, and the discriminator to identify the artificially generated images, overcoming the training data limitations. One of the main problems in person ReID concerns the reduced number of images from different poses.

To overcome this limitation, a feature distilling generative adversarial network (FD-GAN) is proposed by [68] in combination with a Siamese CNN structure to learn identity-related and pose-unrelated representations. In addition, a novel same-pose loss was also formulated and integrated, requiring the appearance of the same person's generated images to be similar. The proposed FD-GAN explores the Siamese scheme, where an image encoder, an image generator, an identity verification classifier, and two adversarial discriminators are included. The corresponding branch of the network takes a person image and a target pose landmark map as inputs. The image encoder at each branch initially transforms the input person image into feature representations. Then, the identity verification classifier is used to supervise the feature learning for person ReID. The image generator starts by taking the encoded person features and target pose map as inputs, and outputs another image of the same person in a different target pose. The target pose map is represented by an 18-channel map, with each channel representing the location of one pose landmark's location, and with the one-dot landmark location being converted to a Gaussian-like heatmap. The encoding is performed by using a 5-block convolution-Batch Normalization (BN)-ReLU subnetwork, generating a 128-dimensional pose feature vector. The visual features, target pose features, and an additional 256-dimensional noise vector, sampled from standard Gaussian distribution, are then concatenated and input into a series of 5 convolution-BN-dropout-ReLU upsampling blocks to output the generated person images. Concerning training, it was performed in three main stages. Initially, the Siamese network baseline built on ResNet-50, using the weights from ImageNet [9], was established. The network was firstly optimized with SGD and trained during 80 epochs. In the second training stage, the encoder and validation classifier were fixed, and the generator was integrated. Adam optimizer [69] was employed to optimize the generator, while the identity discriminator and posed discriminator were optimized with SGD. Lastly, a global fine-tuning was done on the model through all blocks in an end-to-end fashion. For performance evaluation, Market-1501, CUHK03, and DukeMTMC-ReID datasets were used, with mAP and CMC Rank-1 accuracy metrics being adopted for performance evaluation on all the three datasets, and the proposed FD-GAN obtained a CMC of 90.5% on Market-1501, 92.6% on CUHK03, and 80.0% on DukeMTMC-ReID. The inclusion of the FD-GAN and pose encoder enables a substantial increase in model performance.

Another standard limitation concerns the lack of diversity in the image domain, namely, different images gathered and subjected to other illumination conditions. A style transfer GAN is proposed by [70] to serve as a data argumentation approach to smooth camera style disparities. The method employs CycleGAN [71] to style transfer trained labeled mages from different cameras aiming an increase of the diversity of training examples, avoiding model overfitting. Focusing on a better handling of noise, a smooth label regularization is introduced. The style transfer method is evaluated on the ReID task using the Market-1501 and DukeMTMC-ReID datasets, with the proposed model achieving a mAP of 71.55% in the Market-1501.

The vast domain range of images poses difficulties to the ReID tasks. To reduce the impact of image domain diversity, ref [72] proposes a joint learning scheme to improve domain adaptation, to disentangle ID-related/unrelated features, which enforces adaptation to focus on the ID-related features space only. The disentangle module is responsible for encoding cross-domain images into a shared appearance and two separated structure spaces, with the adversarial alignment being performed by the adaptation module. Extensive experiments using the Market-1501 and DukeMTMC-ReID datasets were performed, with the model achieving a Rank-1 of 83.1 in the Market-1501.

A careful evaluation of methodologies for person ReID was performed by [38]. The evaluation starts by setting a baseline backbone architecture based on ResNet-50 [14] with

weights initialized on ImageNet [9], and changing the dimension of the fully connected layer to the number of identities in the training set. Similar assumptions were made for all experiments for training as described in the article [38]. Several training tricks were evaluated, such as warm-up learning rate [73], to bootstrap the network for better performance; random erasing augmentation [33], where an image I in a mini-batch has the probability of undergoing random erasing of p_e; label smoothing [44] to prevent the model from overfitting the training ID, where a small constant ϵ is introduced to avoid the overfit of the training set; last stride [74] to obtain a higher spatial resolution, enriching the granularity of features. To embedded different features distances to accommodate different class distributions in different sub-spaces on the ID loss during inferring stage several strategies are employed, such as BNNeck, which adds only BN layer after features and before classifier FC layers; and, finally, a center loss [75], to simultaneously learning deep features of each class, while penalizing the distances between the deep features and their corresponding class centers, avoiding the drawbacks of the triplet loss. The performance and contribution of each of the tricks were evaluated, with the best model using all the described tricks used, achieving a CMC Rank-1 of 94.5% on Market-1501 and CMC Rank-1 of 86.9% on DukeMTMC-ReID. While the study shows the effectiveness of the DNN tricks to the ReID task, in [76] is proposed a k-reciprocal encoding method to re-rank the ReID results, to increase the accuracy. The main underline consideration concerns the fact that if a gallery image is on par with the probe in the k-reciprocal nearest neighbors, it is more likely to be a true match. In detail, given an image, a k-reciprocal feature is calculated by encoding its k-reciprocal nearest neighbors into a single vector used for re-ranking under the Jaccard distance.

A performance evaluation of the reviewed methods for person ReID using Generative learning and other complementary methods is given in Table 5.

Table 5. Performance evaluation of the reviewed person ReID using generative methods.

Cat	Ref.	Main Technique(s)	# Data Success	Pros/Cons
GAN	[68]	GAN, ResNet-50, pre-trained, pose generator	Market-1501 CMC 0.905	Uses GAN, hard to train state-of-the-art
	[70]	GAN, style transfer, smooth regularization	Market-1501 mAP 0.715	Uses GAN, simple replicable
	[72]	GAN, joint learning, domain adaptation	Market-1501 Rank 1 0.831	Uses GAN, complex replicable
Others	[38]	Evaluation of techniques, Pre-trained, Modified Triple loss	Market-1501 CMC 0.945	Simple to reuse, reusable explanatory

CMC—cumulative matching characteristic (higher, the better), mAP—mean average precision (higher, the better), all measures range: $[0.0, 1.0]$.

A performance comparison among the described works is depicted in Figure 8.

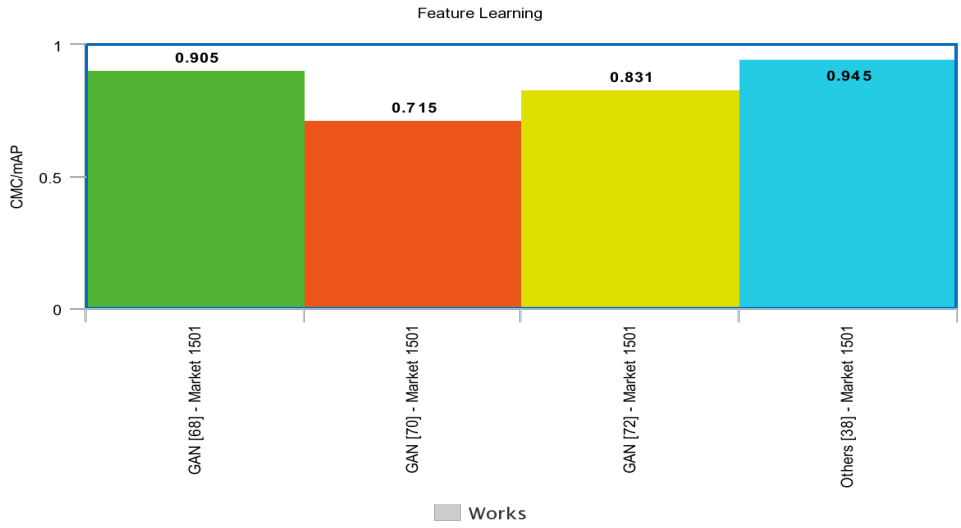

Figure 8. Performance summary of the evaluated generative model methods.

5.2.5. Summary of Person ReID

The different techniques reviewed in this section focus on solving the person ReID problematic. The study about the different methods, which researchers among the literature have proposed, reveals that very few methods can achieve accurate results on a wide range of datasets that contain different varying position, occlusions genres, shapes, and the illumination of the person in the scene. The performance of the enlisted methods is useful for comparison purposes, giving insight on how to devise a robust, yet simple, person ReID method that can achieve high accuracy. The best performing models are mostly based on pre-trained deep neural network models for feature extraction, combined with schemes and modified triplet loss for person ReID. New approaches are focusing on 3D CNN networks by transfer learning their embedding and reusing them into person ReID. Other methods explore baseline models and complement then using data augmentation and other tricks to improve their performance for ReID tasks. Moreover, a clear trend in the ReID research community relates to the use of GAN methods, in combination with ResNets, to increase model robustness against different object poses, overcoming the number of pair examples in the training dataset, and achieving superior results. A recent trend in person ReID is the use of attention mechanisms to capture relevant features, leading to a significant improvement in model performance.

An important issue regarding the person ReID is the use of biometric characteristics, such a human faces or people's skin. In [77], an important study was conducted using obfuscated and non-obfuscated person faces, and most of the case models performed in the majority of benchmark datasets were better when trained and used people's faces as expected. However, this can lead to biased models (discriminating ones). The public usage of these models can collide with current policies in law practices in several countries; it is helpful to always have a side-by-side comparison of both modalities.

6. ReID and Spatial–Temporal Multi Object ReID Methods

One of the main difficulties of object ReID is to operate in distributed scenarios and account for spatial–temporal constraints for multi-object ReID. Main techniques explore the use of RNN models to construct tracklets to assign IDs to objects, enabling to robustly handle occlusions. In contrast, others rely on 3D CNN to attain temporal dependencies of the object in track.

6.1. Multi Object ReID Datasets with Trajectories

Multi-Object ReID considers attributes, such as shape and category combined with trajectories, and is one of the objectives that go towards the objective to perform multi-object

ReID in a real urban scenario. However, the development of suitable benchmark datasets for attribute recognition remains sparse. Some object ReID datasets contain trajectories collected from various sources, and such heterogeneous collection poses a significant challenge in developing high-quality fine-grained multi-object recognition algorithms.

Among the publicly available datasets for ReID, one example is the NGSIM dataset [78], a publicly available data set with hand-coded Ground Truth (GT) that enables evaluating multi-camera, multi-vehicle tracking algorithms on real data, quantitatively. This dataset includes multiple views of a dense traffic scene with stop-and-go driving patterns, numerous partial and complete occlusions, and several intersections.

Another example is the KITTI Vision Benchmark Suite [79], which is composed of several datasets for a wide range of tasks, such as stereo, optical flow, visual odometry, 3D object detection, and 3D tracking, complemented with accurate ground truth provided by Velodyne laser 3D scanner and real GPS localization system, Figure 9. The datasets were captured by driving around the mid-size city of Karlsruhe, in Germany, in rural areas, and on highways, and on average, there are up to 15 cars and 30 pedestrians per image. A detailed evaluation metric and evaluation site are also provided.

Figure 9. Example of the multi-object tracking and segmentation system (MOTS).

A summary of the public available databases for multi-object car ReID with trajectories is presented on Table 6.

Table 6. Global overview of the public available databases for multi-object car ReID with trajectories.

Dataset	# Identities	# Cameras	# Images	Label Method	Size	Tracking Sequences
NGSIM						
KITTI	-	-	-	R-CNN	1392 × 512	Yes
UA-DETRAC	825	24	1.21 M	Manual	960 × 540	Yes
VehicleID	26,267	-	221 K	Manual	Vary	Yes
VeRi-776	776	18	50 k	Manual	Vary	Yes
CompCar	1687	-	18 k	Manual	Vary	Yes
PKU-Vehicle	-	-	18 M	Manual	Cropped	Yes
MOT20-03	735	-	356 k	R-CNN	1173 × 880	Yes
MOT16	-	-	476 k	R-CNN	1920 × 1080	Yes
TRANCOS	46,796	-	58 M	HOG	Vary	Yes
WebCamT	-	212	60 k	-	-	No

R-CNN—region proposals with CNN, HOG—histogram oriented gradients.

In addition, the PASCAL VOC project [80], provides standardized image datasets for object class recognition, with annotations that enable evaluation and comparison of different methods. Another useful dataset is ImageNet [10]. This image database is organized according to the WordNet hierarchy (currently only the nouns), where each node is represented by hundreds and thousands of images. Currently, each node contains an average of over 500 images. The dataset encompasses a total of 21,841 non-empty sets, forming a total number of images of around 14 million, with several images with bounding box annotations of 1 million. In addition, it contains 1.2 million images with pre-computed SIFT features [81]. Pre-trained networks on these two datasets are commonly reused as backbones for feature extraction to perform several tasks, such as object ReID.

6.2. Spatial–Temporal Constrained and Multi Object ReID Methods

In this section, deep learning-based vehicle ReID methods are grouped into four main categories, as represented in Figure 10; which includes methods for feature learning, sequence learning, deep learning metrics, and tracking, with these categories encompassing several methods. The main aspects of each category method and their experimental results are discussed in the following.

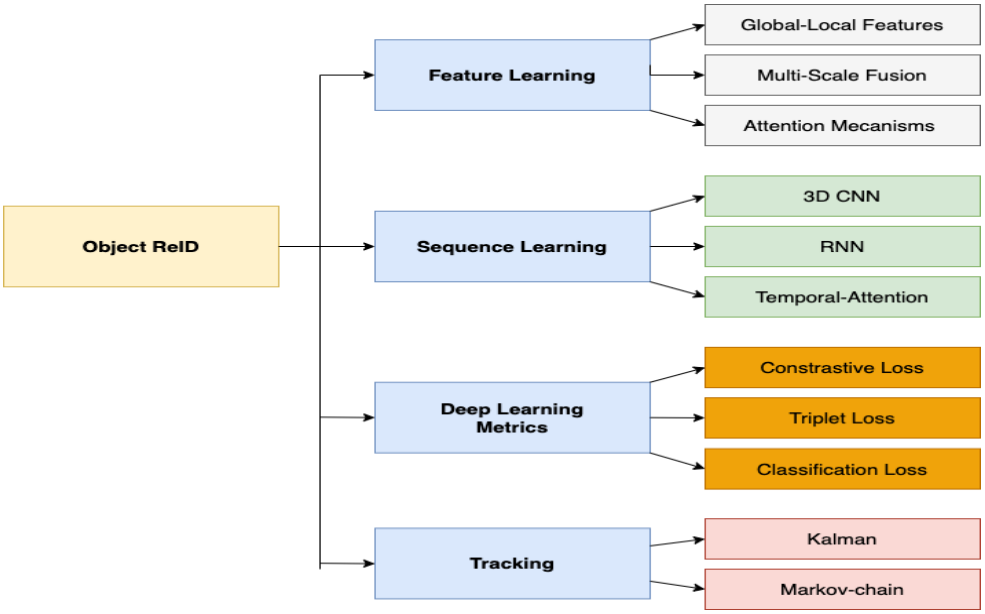

Figure 10. Deep learning-based vehicle re-identification methods.

6.2.1. Deep Learning Metrics for Vehicle ReID

Similar to person ReID, triplet loss are a excellent tool for vehicle ReID tasks. A variation of the triplet loss commonly used in deep learning methods, defined as a group-sensitive-triplet embedding (GS-TRE), was proposed in [82] to recognize and retrieve vehicles, where the intraclass variance is elegantly modeled by incorporating an intermediate representation "group" between samples and each vehicle in the triplet network learning. The main objective is to address the car ReID problematic, namely the fact that common deep metric learning with a triplet network common configuration ignores the impact of intra-class variance-incorporated embedding on the performance of vehicle ReID, where robust fine-grained features for large-scale vehicle ReID have not been fully studied. In addition, a clustering strategy to derive group labels, and in particular, an online clustering method, is employed, and a mean-valued triplet loss [83] is also proposed to enhance the learning of discriminative features. The performance of the proposed group-sensitive-triplet embedding (GS-TRE) was evaluated on the VeRi-776 and VehicleID datasets, with the model obtaining a mAP of 0.743 on the VehicleID dataset, with the modified triplet loss well suitable to help in the ReID task.

An improvement of triplet loss is presented by [84], focusing on two aspects: first, a stronger constraint, namely classification-oriented loss augmented with the original triplet loss; second, a new triplet sampling method based on pairwise images is proposed in combination with a classification-oriented loss to implicitly impose a constraint for the embedded features of the images of the same vehicle to be similar, and also by ensuring negative samples in one triplet act as positive samples in another triplet. The system architecture consists of three parts: a shared deep CNN to learn a mapping from raw images to Euclidean space, with the distance reflecting the relevance between the images, a triplet stream for calculating the distances and providing the constraint of the triplet loss, and a classification stream for ID level supervision provided by the classification-oriented loss, with the image triplets being generated by the proposed triplet sampling method. The

deep CNN for feature extraction was fine-tuned from VGG CNN using the pre-trained weights from the ILSVRC-2012 dataset [10]. Stochastic gradient descent was employed during the training process. Results were gathered based on the VeRi dataset using the CMC and mAP, with model obtaining a mAP of 0.5740 on this dataset. While the results show a lower level of performance, making evident that simple triplet usage is not sufficient for robust ReID systems.

Contrast loss is a useful tool for pair-wise marching, enabling the model to focus on discriminate features. A novel deep learning-based approach to progressive vehicle ReID, called PROVID, was proposed by [85]. The approach starts by addressing the ReID as two distinct search processes: coarse-to-fine search that operates in the feature space and near-to-distant search to address real-world scenarios. The first searching process employs the appearance attributes of the vehicle for coarse filtering, while simultaneously exploring Siamese Neural Network architecture for license plate verification for vehicle identification. The near-to-distant search process enables to retrieve vehicles´ identity by searching from near to faraway cameras and from close to a distant time. To account for the spatial–temporal domain, an assumption is that two images have a higher probability of being the same object, if they have a small space or time distance among frames, and a lower probability of being the same vehicle if they have large space or time distance. With this in mind, for each query image i and test image j, a spatial–temporal similarity $ST(i, j)$ is defined, and with a re-ranking strategy in combination to model the spatial–temporal information with the appearance and plate features. The model was evaluated using the VeRi-776 dataset, showing that the proposed method achieved a 0.277 mAP. The results are much lower when compared with related works. In addition, the use of discriminating features, such as the vehicle identification plate, may cause problems in the usage on public surveillance systems and comply with legislation in place.

A unified multi-object tracking (MOT) framework is presented in [86], enabling exploring the full potential of the long-term and short-term cues for handling complex cases in MOT scenes. For better association, a switcher-aware classification (SAC) is proposed, exploring the potential of the identity-switch causer (switcher). Specifically, the proposed method incorporates a single object tracking (SOT) subnet to capture short-term cues, a ReID subnet to extract long-term cues and a switcher-aware classifier to make matching object decisions, using extracted features from the main target and the switcher. The main objective of short-term cues is to help find false negatives, while long-term cues avoid critical mistakes when occlusion occurs, with the SAC learning used to combine multiple cues in an effective way to improve robustness. The SOT subnet and the ReID subnet are trained independently. For the SOT subnet, image pairs of targets are generated according to the GT of the videos, and the pairs are extended to include part of the background according to the training schema of Siamese-RPN. On the other hand, for the ReID subnet, each target is regarded as one class, with the network trained to predict the class of the input target. Extensive model evaluations were performed using the challenging MOT16 benchmarks [87], achieving a CLEAR MOT of 71.2%, proving the effectiveness of the switcher to robustly assign long-term occluded objects to corresponding tracklets.

Performance evaluation of vehicle ReID using deep learning metrics is resumed in Table 7.

Since it is similar to person ReID, several authors explore the Siamese and triplet loss for the ReID task. Because the scheme is similar in what concerns the job itself, the leaned features are very distinct, hampering the use of Siamese and triplet loss, both for person and vehicle ReID in a single unified framework.

Table 7. Performance evaluation of the reviewed vehicle ReID using deep leaning metric methods.

Cat	Ref.	Main Technique(s)	# Data Success	Pros/Cons
Triplet	[82]	CNN features, group-sensitive-triplet emb.	VehicleID mAP 0.743	Reproducible, ranking problems
	[84]	VGG features, triplet sampling method	VeRi-776 mAP 0.574	reproducible, robust
Contrast loss	[85]	Siamese Neural Net, spatial–temporal similarity	VeRi-776 mAP 27.77	Simple, not robust
	[86]	Siamese-RPN, switcher -aware classification (SAC)	MOT16 CLEAR 0.712	Complex, trajectory ID handled

mAP—mean average precision (higher, the better; range: $[0.0, 100.0]$).

6.2.2. Sequence Models for Vehicle ReID

To capture temporal features, RNN are commonly employed.

In [88], two end-to-end deep architectures, defined as the spatially concatenated CNN and CNN LSTM bi-directional loop, were proposed to address the problematic of vehicle viewpoint uncertainty. The models exploit the great advantages of CNN and LSTM to learn transformations across different viewpoints of vehicles, enabling to attain multi-view vehicle representation containing all viewpoints; information that can be inferred from the only one input view, and then used for learning to measure distance. The evaluation of the model was performed using a new proposed toy car ReID dataset with images from multiple viewpoints of 200 vehicles and the public multi-view car, VehicleID, and VeRi datasets. Conducted experiments showed that the proposed model could achieve a mAP of 18.13 on the VeRi dataset.

In [89], a deep spatial–temporal neural network is proposed to solve the task of sequentially counting vehicles from low-quality videos acquired by city cameras (citycams). Citycam videos are characterized by low resolution, low frame rate, high occlusion, and broad perspective, making most existing methods lose efficacy. To overcome the limitations of the current methods and incorporate the temporal information of traffic video, a novel FCN-rLSTM network was proposed to jointly estimate vehicle density and vehicle count by connecting Fully Convolutional Network (FCN) with LSTM in a residual learning approach. The design enables leveraging the strengths of FCN for pixel-level prediction and the strengths of LSTM to learn complex temporal dynamics. The residual learning connection reformulates the vehicle count regression as a learning residual function concerning the sum of densities in each frame, leading to a significant reduction in network training. To preserve feature map resolution, a hyper-atrous combination was proposed to integrate atrous convolution on the FCN, and combine feature maps of different convolution layers. FCN-rLSTM enables refined feature representation and a new end-to-end trainable mapping from pixels to vehicle count. The proposed method was extensively evaluated on different counting tasks using three datasets, with experimental results demonstrating their effectiveness and robustness. In particular, the proposed FCN-rLSTM reduced the Mean Absolute Error (MAE) from 5.31 to 4.21 on the TRANCOS dataset, showing the abilities of LSTM in combination with FCN to track objects in low-resolution videos.

A practical vehicle tracking framework and trajectory-based weighted ranking method, which significantly improves the performance of cars ReID, were proposed in [90]. The proposed approach makes use of a ResNet50 [14] as the backbone for feature extraction, trained using the set of AI City Challenge [91] and VisDrone2018 [92] datasets, and only considering only the vehicle category. In the inference phase, the image is resized into 1440×800, to capture small vehicles in the video. By using the unified multi-object tracking framework proposed by [86], long-term and short-term cues are fully used as a detector with high recall, considering only boxes with higher confidence as input for the multiple target tracking algorithm. The similarity between the two features space is calculated

through cosine similarity with the overall loss function containing the cluster loss, trajectory consistency loss, and classification loss. During the inference phase, a re-ranking with spatial–temporal cue is used, with all trajectories encoded in a feature vector of 2048 dimensions. A density clustering DBSCAN is used to gather similar vehicles with the different ID that belong to the same class. Finally, a ranking with weighted features and trajectory information is set in place to identify individual trajectories from the class group. Conducted experiments using the AI city challenge dataset achieved a mAP of 0.730, showing competitive tracking results in real urban scenarios.

In [93] is proposed an extension to the prevalent task of multi-object tracking and segmentation (MOTS). It explores widely annotated dense pixel-level annotations of two existing tracking datasets using a semi-automatic annotation procedure, containing the masks for 977 distinct objects (cars and pedestrians) in 10,870 video frames. To tackle detection, tracking, and segmentation, i.e., the MOTS task, a neural network is employed jointly with a baseline method built upon the famous Mask R-CNN [39] architecture, which extends the faster R-CNN [94] detector with a mask head.

The TrackR-CNN model provides mask-based detection and association features. Both are used as input to a tracking algorithm that decides which detection to select, and how to integrate temporal context information. The temporal context of the input video is explored by the integration of 3D convolutions (to account for time), into Mask R-CNN on top of a ResNet-101 [14] backbone. The 3D convolutions layers are used to extract the backbone features and to obtain a temporal augmentation context. These new augmented features are then used by the Region Proposal Network (RPN) for the ReID task. For evaluation purposes, the proposed method was set as a baseline that jointly addresses detection, tracking, and segmentation with a single CNN. Conducted experiments demonstrated the relevance of the constructed datasets, enabling them to achieve considerable improvements in performance when trained on MOTS annotations. The datasets, proposed metrics, and baselines, such as MOTSA and sMOTSA, were considered with the baseline model achieving a sMOTSA of 52.7% and MOTSA of 66.9% on the KITTI MOTS dataset, enabling to account for temporal multi-object ReID.

Table 8 summarizes the discussed works and establish comparisons among the performance and used datasets.

Table 8. Performance evaluation of the reviewed vehicle ReID using sequence learning methods.

Cat	Ref.	Main Technique(s)	# Data Success	Pros/Cons
LSTM	[88]	Spatially Concatenated CNN, CNN-LSTM bi-directional loop	VeRi-776 mAP 18.13	Simple, applicable
	[89]	CNN features, gFCN- rLSTM network + Atrous	TRANCOS MAE 4.21	Reproducible, ranking problems
	[90]	ResNet50, LSTM + clustering DBSCAN	AI City MAE 0.730	Complex, trajectory problems
3D	[93]	Mask R-CNN, 3D convolutional layers	KITTI MOTS 0.669	Robust, Short term ID handled

mAP—mean average precision (higher, the better; range: $[0.0, 100.0]$).

6.2.3. Feature Learning for Vehicle ReID

Feature learning are sometimes implicit when using pre-trained backbones, such as ResNets. However, many works explore the use of specialized schemes to explore the potentialities of global and local features, often by using fusion schemes.

In [95], the authors proposed an approach on vehicle ReID without any knowledge about localization or movements of the cars. This method obtains real-time traffic information based on linear regression with SVM, according to feature vectors, which consist of a color histogram and oriented gradients. First, the vehicles are detected in the video by an object classifier model that creates 3D bounding boxes around the cars. Only the side

and front (or back) faces vehicle images are extracted. The extracted image is then fitted into a grid and color histograms to be found in another vehicle image set by simulating a different camera view to be used for the first-round regression. Vehicles with positive first-round regression results are then tested on the second-round regression, where the average Histogram of Oriented Gradients (HOG) vector is used. Cars with both regression results positive are added to another set and are considered as highly potential positive ReID candidates. Experiments were performed on pre-selected semi-automatically 1232 image pairs likely to be matching vehicles, and using a web interface and crowd-sourced people's opinion of vehicles that "are likely to be the same vehicle". The findings showed that 60% of matches could be retrieved (TPR), with only about 10% of False Positive (FP) being included. The proposed method lacks robustness and relies on great percentages on non-robust features that are not optimal to be used in a variety of urban scenarios.

A two-branch CNN scheme was presented in [96] to learn deep features and the distance metric simultaneously. The proposed model uses the late fusion scheme to combine attributes and color features (FACT). It is the late fusion scheme that starts by ranking scores of all test images with the semantic feature learned by GoogLeNet [97] separately. Conducted experiments were performed against other methods, with the rank scores being calculated by the Euclidean distance. The model evaluation was performed on the VeRi dataset, with the proposed model achieving a mAP of 19.92.

In [98], the authors proposed a new spatially constrained similarity measure (SCSM) to handle object rotation, scaling, viewpoint change, and appearance deformation for object ReID in combination with a robust re-ranking method with the k-nearest neighbors of a given query for automatically refining the initial search results. The retrieval system is implemented with SIFT descriptors [81] and fast approximate k-means clustering [99] creating a bag of words (BoW) classification scheme. Extensive performance evaluations on INRIA dataset achieves a mAP of 0.762.

In [100], the authors presented a forecasting mechanism to forecast pedestrian destinations in a large area with a limited number of observations. To address the challenges posed by a limited number of observations (e.g., sparse cameras), and change in pedestrian appearance cues across different cameras, a new descriptor is defined as social affinity maps (SAMs) to link broken, or unobserved trajectories of individuals in the crowd. To continuously track the pedestrians, a Markov-chain model is used to connect every intermediate track x_t^i in trajectory T, to subsequent track x_t^{i+1} with a given probability encoded as priors over Origin and Destination (OD) preferences. In addition, the proposed work also introduces a dataset of 42 million trajectories collected in train stations. The conducted experiments were performed using SAM features, and results showed that the performance of OD forecasting with a different number of in-between cameras increased, and more accurate trajectories were predicted, obtaining an overall OD error rate of 0.672. The SAM enables the extraction of relevant features to maintain continuous track in occluded pedestrians over time.

The fine-grained recognition of vehicles, mainly in traffic surveillance applications, is addressed in [101]. The approach is based on recent advancements in fine-grained recognition: automatic part discovery and bilinear pooling. In contrast to other methods that focus on fine-grained recognition of vehicles, viewpoints are not limited only to a frontal/rear viewpoint, but it allows the vehicles to be seen from any viewpoint. The approach is based on 3D bounding boxes built around the vehicles that are automatically constructed from traffic surveillance data. An CNN based on ResNet50 is used for the estimation of the directions towards the identified vanishing points by feeding the vehicle image into a ResNet50 with three separate outputs regarding the probabilities for directions of vanishing points in quantized angle space. A new annotated dataset BoxCars116k is proposed, focusing on images gathered from surveillance cameras. Several experiments were conducted, with the proposed method significantly improving the CNN classification accuracy, achieving a 12% increase (80.8%) on bounding box determination.

A summary regarding the methods for vehicle ReID is presented in Table 9.

Table 9. Performance evaluation of the reviewed vehicle ReID using feature learning methods.

Cat	Ref.	Main Technique(s)	# Data Success	Pros/Cons
Fusion	[95]	SVM, HOG	Own -	Not replicable,
	[96]	GoogLeNet, Feature fusion	VeRi-776 mAP 19.92	Simple, Baseline
	[98]	SIFT + BOW, re-ranking	INRIA mAP 0.762	OLD fashion, Not SOTA
	[100]	Social Affinity Maps (SAM), Markov-chain model	Own –	Complex, Only indoors
	[101]	3D box prediction, ResNet	BoxCars116k ACC 0.808	Not useful, simple

mAP—mean average precision (higher, the better; range: $[0.0, 100.0]$).

6.2.4. Tracking for Vehicle ReID

Continuously tracking of objects of interest is a common requirement for any ReID system that has been addressed using different strategies ranging form Kalman filtering based, Optical flow, and many combinations of those with other mechanism.

In [102], the authors proposed an object tracking and 3D reconstruction method to perform 3D object motion estimation. Object tracking and 3D reconstruction are often performed together, with tracking used as input for the 3D reconstruction. To improve tracking performance, a novel method is proposed to close this gap, by first tracking and reconstructing to track. The proposed multi-object tracking, segmentation, and dynamic object fusion (MOTSFusion) approach exploits the 3D motion information extracted from dynamic object reconstructions to track objects through long periods with complete occlusion and recover missing detections. The method first builds up short tracklets using 2D optical flow and then fuses them into dynamic 3D object reconstructions. The precise 3D object motion of these reconstructions is used to merge tracklets through occlusion into long-term tracks and to locate objects in the absence of detection. Conducted experiments were performed on the KITTI platform [93,103], with the reconstruction-based tracking reducing the number of ID switches of the initial tracklets by more than 50%. CLEARMOT [104] was adopted as evaluation metric for bounding box tracking to rank it in terms of MOTA [103], which incorporates FP, False Negative (FN), ID switches (IDS) and sMOTSA to account the segmentation Intersection over Union (IoU) accuracy, with the method achieving a MOTA of 84.83%. The method enables to robustly incorporate long-term occluded objects in an optimized manner.

A recent and widely used tracking-by-detection algorithm is the DeepSort [105]. Simple online and real-time tracking (SORT) enables tracking multiple objects for more extended periods. The method relies on object detector backbones, such as Yolo [106], where an appearance integration is integrated into the Kalman filter to effectively track uniquely similar objects in strong occluded environments, while reducing the number of tracking or ID switch. When an object in track by the Kalman filter, the update of the filter is performed considering the dynamics, the association mechanism based on the Mahalanobis distance, with an extra feature concerning the object appearance, for each object bounding box appearance, a gallery of associated appearance descriptors is kept for each track. Cosine similarity on new observation is used to compute the distance of the current object in track and the new observation, enabling to correctly associate the observation even for an object with high dynamics that cannot be handled solely by the Kalman filter association metrics. The experimental evaluation shows that the inclusion of deep features into the Kalman filter reduces the number of identity switches by 45%.

Table 10 summarizes the discussed works and establish comparisons among the performance and used datasets.

Table 10. Performance evaluation of the reviewed vehicle ReID Tracking methods.

Cat	Ref.	Main Technique(s)	# Data Success	Pros/Cons
	[102]	3D reconstruction, 2D optical flow	KITTI MOTA 0.848	Robust, Short term ID handled
	[105]	CNN features, Kalman + Association	- -	Robust, Short term ID handled

mAP—mean average precision (higher, the better; range: $[0.0, 100.0]$).

6.2.5. Summary of Vehicle ReID Methods

Several research works concerning object ReID with spatial–temporal constraints can be identified in the literature. However, the ReID in non-overlapping cameras with tracking is commonly accepted to be difficult task, and the range of works that address this problem is not vast. However, there is a clear trend in the use of deep learning features or 3D CNN commonly used in action recognition, explored in the ReID task to capture spatial–temporal invariant features to improve ReID generalization performance in unseen objects over time; however, ReID with spatial–temporal constraints is a difficult task to accomplish soon, mainly in urban scenarios due to the infinite number of partial occlusions, uneven, and dynamic illumination conditions.

7. Methods for Image Enhancement

Severe weather conditions, such as rain and snow, adversely may affect the visual quality of the images acquired under such conditions; thus, rendering them useless for further usage and sharing. In addition, such degraded images usually drastically affect the performance of vision systems. Mainly, it is essential to address the problem of single image de-raining. However, the inherent ill-posed nature of the situation presents several challenges.

In [107], it is proposed an image de-raining conditional generative adversarial network (ID-CGAN) that account for quantitative, visual, and also discriminative performance into the objective function. The proposal method explores the capabilities of conditional generative adversarial networks (CGAN), in combination with additional constraint to enforce the de-rained image to be indistinguishable from its corresponding GT clean image. A refined loss function and other architectural novelties in the generator–discriminator pair were also introduced, with the loss function aimed towards the reduction of artifacts introduced by GAN, ensuring better visual quality. The generator sub-network is constructed using densely connected networks, whereas the discriminator is designed to leverage global and local information and between real/fake images. Exhaustive experiments were conducted against several State-of-the-Art (SOTA) methods using synthetic datasets derived from the UCID [108] and BSD-500 [109] datasets, and with external noise artifacts added. The experiments were evaluated on synthetic and real images using several evaluation metrics such as peak signal to noise ratio (PSNR), structural similarity index (SSIM) [110], universal quality index (UQI) [111], and visual information fidelity (VIF) [112], with the proposed model achieving an PSNR (DB) of 24.34. Moreover, experimental results evaluated on object detection methods, such as FasterRCNN [94], demonstrated the effectiveness of the proposed method in improving the detection performance on images degraded by rain.

A single-image-based rain removal framework was proposed in [113] by properly formulating the rain removal problem as an image decomposition problem based on the morphological decomposition analysis. The alternative to applying a conventional image decomposition technique, the proposed method first decomposes an image into the low and high-frequency (HF) components by employing a bilateral filter. The HF part is then decomposed into a "rain component" and a "non-rain component" using sparse coding. The model experiments were conducted on synthetic rain images built using an image software, with the model achieving a VIF of 0.60. While the method has some degree of

performance with common rain conditions, it has difficulties to handling more complex rain dynamic scenarios.

In [114], an effective method based on simple patch-based priors for both the background and rain layers is proposed, which is based on the Gaussian mixture model (GMM) to accommodate multiple orientations and scales of the rain streaks. The two GMMs for the background and rain layers, defined as GB and GR, are based on a pre-trained GMM model with 200 mixture components. The method was evaluated using synthetic and real images, and the results compared to SOTA methods, with the proposed method achieving an SSIM of 0.88.

In [115], it is proposed a DNN architecture called DerainNet for removing rain streaks from an image. The architecture is based on a CNN, enabling the direct map of the relationship between rainy and clean image detail layers from the data. For effective enhancement, each image is decomposed into a low-frequency base layer and a high-frequency detail layer. The detail layer corresponds to the input to the CNN for rain removal to be combined at a final stage with the low-frequency component. The CNN model was trained using synthesized images with rain, with the model achieving an SSIM of 0.900, increasing by 2% the performance in comparison to [114] using GMM.

A performance evaluation of the reviewed image de-raining methods is given in Table 11.

Table 11. Performance evaluation of the reviewed image de-raining methods.

Reference	Main Techniques	# Data Success	Pros/Cons
[107]	GANS, conditional GAN	UCID PSNR 24.34	Robust, SOTA
[113]	Bilateral filter, image decomposition	Systeticg VIF 0.60	Simple, Parameter dependent
[114]	GMM, image decomposition	Systetic SSIM 0.880	Simple, Pre-trained dependent
[115]	CNN, HF component layer	Systetic SSIM 0.900	Simple, Robust

PSNR—peak signal to noise ratio (higher, the better; range: $[0.0, -]$), VIF—Information Fidelity (higher the better, range: $[0.0, 1.0]$), SSIM—structural similarity index (higher, the better; range: $[0.0, 1.0]$).

While many other methods can be found, the aforementioned ones highlight the most common approaches to the image enhancement problematic when operating in urban scenarios, where illumination conditions are not constant, due to rain, fog, and illumination, which potentially hamper the performance of the ReID methods.

8. Conclusions

A detailed overview of SOTA methods to date were presented in this paper, including comparisons to identify the main advantages and problems the methods present. In addition, the most commonly used image datasets and their main characteristic were identified.

Image enhancement is a vital component of any computer vision system. It can improve the performance of the initial object detectors and classification, leading to improved ReID systems. Most of the works explore the use of pre-trained DNN, acting as a backbone for feature extraction, with most of them exploring a residual network, enabling to easily reuse the extracted feature maps for ReID model variations. However, the person and vehicle ReID are addressed separately, and fewer research studies have proposed long-term tracking with ReID simultaneously.

The published articles for person ReID concerned DNN using Siamese networks are the most prominent, exhibiting good performance results. Most of the identified works explore novel augmentation and dropout techniques during training, framed with different

triplet loss variations. Most of the research studies have obtained competitive results in the common ReID datasets. However, without proper generalization evaluation in real scenarios where light conditions are more challenging.

In this review, the focus was only on object ReID methods; however, a reliable system comprehends two stages for the task, the detection process and the ReID mechanism by itself. However, a fully end-to-end object ReID requires high precision of the object detection, as well as unlabeled ones, and difficulties on the effective combinations of object detection and ReID in a fully integrated ReID system are directions that require attention in the near future.

From this review, it is possible to conclude that there is a lot of room for improvement regarding the multi-object ReID and long-term tracking that is still not explored in the scientific community, combined with the object detection stage, is still an open problem, and commonly not addressed in the identified ReID works.

Author Contributions: Conceptualization, funding acquisition, and supervision by J.M.R.S.T.; investigation, data collection, formal analysis, and writing original draft preparation by H.S.O.; writing review and editing by H.S.O., J.J.M.M., and J.M.R.S.T. All authors have read and agreed to the published version of the manuscript.

Funding: This article is a result of the project Safe Cities—"Inovação para Construir Cidades Seguras", with reference POCI-01-0247-FEDER-041435, co-funded by the European Regional Development Fund (ERDF), through the Operational Programme for Competitiveness and Internationalization (COMPETE 2020), under the PORTUGAL 2020 Partnership Agreement.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wu, L.; Wang, Y.; Gao, J.; Li, X. Deep adaptive feature embedding with local sample distributions for person re-identification. *Pattern Recognit.* **2018**, *73*, 275–288. [CrossRef]
2. Zhang, W.; Ma, B.; Liu, K.; Huang, R. Video-based pedestrian re-identification by adaptive spatio-temporal appearance model. *IEEE Trans. Image Process.* **2017**, *26*, 2042–2054. [CrossRef]
3. Varior, R.R.; Haloi, M.; Wang, G. Gated Siamese Convolutional Neural Network Architecture for Human Re-Identification. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 791–808.
4. Xiao, T.; Li, H.; Ouyang, W.; Wang, X. Learning deep feature representations with domain guided dropout for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Vegas, NV, USA, 27–30 June 2016; pp. 1249–1258.
5. McLaughlin, N.; Martinez del Rincon, J.; Miller, P. Recurrent convolutional network for video-based person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Vegas, NV, USA, 27–30 June 2016; pp. 1325–1334.
6. Yan, Y.; Ni, B.; Song, Z.; Ma, C.; Yan, Y.; Yang, X. Person re-identification via recurrent feature aggregation. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 701–716.
7. Yi, D.; Lei, Z.; Liao, S.; Li, S.Z. Deep metric learning for person re-identification. In Proceedings of the 2014 22nd International Conference on Pattern Recognition, Washington, DC, USA, 24–28 August 2014; pp. 34–39.
8. Zheng, Z.; Zheng, L.; Yang, Y. A discriminatively learned cnn embedding for person reidentification. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* **2017**, *14*, 1–20. [CrossRef]
9. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
10. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]
11. LeCun, Y.; Boser, B.; Denker, J.; Henderson, D.; Howard, R.; Hubbard, W.; Jackel, L. Handwritten digit recognition with a back-propagation network. *Adv. Neural Inf. Process. Syst.* **1989**, *2*, 396–404.
12. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.

13. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. Inception-v4, inception-resnet and the impact of residual connections on learning. *arXiv* **2016**, arXiv:1602.07261.
14. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
15. Canziani, A.; Paszke, A.; Culurciello, E. An analysis of deep neural network models for practical applications. *arXiv* **2016**, arXiv:1605.07678.
16. Gong, S.; Cristani, M.; Yan, S.; Loy, C.C.; Re-Identification, P. Springer Publishing Company. *Incorporated* **2014**, *1447162951*, 9781447162957.
17. Li, D.; Zhang, Z.; Chen, X.; Ling, H.; Huang, K. A richly annotated dataset for pedestrian attribute recognition. *arXiv* **2016**, arXiv:1603.07054.
18. Gray, D.; Tao, H. Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features. In Proceedings of the European Conference on Computer Vision, Marseille, France, 12–18 October 2008; Springer: Berlin/Heidelberg, Germany, 2008; pp. 262–275.
19. Nguyen, T.B.; Le, T.L.; Nguyen, D.D.; Pham, D.T. A Reliable Image-to-Video Person Re-Identification Based on Feature Fusion. In Proceedings of the Asian Conference on Intelligent Information and Database Systems, Dong Hoi City, Vietnam, 19–21 March 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 433–442.
20. Pham, T.T.T.; Le, T.L.; Vu, H.; Dao, T.K. Fully-automated person re-identification in multi-camera surveillance system with a robust kernel descriptor and effective shadow removal method. *Image Vis. Comput.* **2017**, *59*, 44–62. [CrossRef]
21. Cheng, D.S.; Cristani, M.; Stoppa, M.; Bazzani, L.; Murino, V. Custom pictorial structures for re-identification. *BMVC* **2011**, *1*, 6.
22. Das, A.; Chakraborty, A.; Roy-Chowdhury, A.K. Consistent Re-Identification in a Camera Network. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 330–345.
23. Moon, H.; Phillips, P.J. Computational and performance aspects of PCA-based face-recognition algorithms. *Perception* **2001**, *30*, 303–321. [CrossRef] [PubMed]
24. Nguyen, T.B.; Le, T.L.; Ngoc, N.P. Fusion schemes for image-to-video person re-identification. *J. Inf. Telecommun.* **2019**, *3*, 74–94. [CrossRef]
25. Matsukawa, T.; Okabe, T.; Suzuki, E.; Sato, Y. Hierarchical gaussian descriptor for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Vegas, NV, USA, 27–30 June 2016; pp. 1363–1372.
26. Li, W.; Zhu, X.; Gong, S. Person re-identification by deep joint learning of multi-loss classification. *arXiv* **2017**, arXiv:1705.04724.
27. Argyriou, A.; Evgeniou, T.; Pontil, M. Multi-task feature learning. *Adv. Neural Inf. Process. Syst.* **2007**, *19*, 41–48.
28. Kong, D.; Fujimaki, R.; Liu, J.; Nie, F.; Ding, C. Exclusive Feature Learning on Arbitrary Structures via $l_{1,2}$-norm. *Adv. Neural Inf. Process. Syst.* **2014**, *1*, 1655–1663.
29. Wang, H.; Nie, F.; Huang, H. Multi-view clustering and feature learning via structured sparsity. *Int. Conf. Mach. Learn.* **2013**, *28*, 352–360.
30. Gray, D.; Brennan, S.; Tao, H. Evaluating appearance models for recognition, reacquisition, and tracking. In Proceedings of the IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS), Rio de Janeiro, Brazil, 14 October 2007; Volume 3, pp. 1–7.
31. Zhou, K.; Yang, Y.; Cavallaro, A.; Xiang, T. Omni-scale feature learning for person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–3 November 2019; pp. 3702–3712.
32. Reddi, S.J.; Kale, S.; Kumar, S. On the convergence of adam and beyond. *arXiv* **2019**, arXiv:1904.09237.
33. Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; Yang, Y. Random erasing data augmentation. *arXiv* **2017**, arXiv:1708.04896.
34. Ning, X.; Gong, K.; Li, W.; Zhang, L.; Bai, X.; Tian, S. Feature refinement and filter network for person Re-identification. *IEEE Trans. Circ. Syst. Video Technol.* **2020**, *31*, 3391–3402. [CrossRef]
35. Quan, R.; Dong, X.; Wu, Y.; Zhu, L.; Yang, Y. Auto-ReID: Searching for a part-aware ConvNet for person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–3 November 2019; pp. 3750–3759.
36. Liu, H.; Simonyan, K.; Yang, Y. Darts: Differentiable architecture search. *arXiv* **2018**, arXiv:1806.09055.
37. Yaghoubi, E.; Borza, D.; Alirezazadeh, P.; Kumar, A.; Proença, H. An Implicit Attention Mechanism for Deep Learning Pedestrian Re-identification Frameworks. *arXiv* **2020**, arXiv:2001.11267.
38. Luo, H.; Gu, Y.; Liao, X.; Lai, S.; Jiang, W. Bag of tricks and a strong baseline for deep person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–20 June 2019; pp. 4321–4329.
39. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
40. Liu, X.; Zhao, H.; Tian, M.; Sheng, L.; Shao, J.; Yi, S.; Yan, J.; Wang, X. Hydraplus-net: Attentive deep features for pedestrian analysis. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 350–359.
41. Hou, R.; Chang, H.; Ma, B.; Huang, R.; Shan, S. BiCnet-TKS: Learning Efficient Spatial–Temporal Representation for Video Person Re-Identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 21–24 June 2021; pp. 2014–2023.

42. Ning, X.; Gong, K.; Li, W.; Zhang, L. JWSAA: Joint weak saliency and attention aware for person re-identification. *Neurocomputing* **2021**, *453*, 801–811. [CrossRef]
43. Shen, C.; Jin, Z.; Zhao, Y.; Fu, Z.; Jiang, R.; Chen, Y.; Hua, X.S. Deep siamese network with multi-level similarity perception for person re-identification. In Proceedings of the 25th ACM International Conference on Multimedia, Mountain View, CA, USA, 23–27 October 2017; pp. 1942–1950.
44. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
45. Schroff, F.; Kalenichenko, D.; Philbin, J. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–11 June 2015; pp. 815–823.
46. Li, W.; Zhao, R.; Xiao, T.; Wang, X. Deepreid: Deep filter pairing neural network for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 152–159.
47. Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; Tian, Q. Scalable person re-identification: A benchmark. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–11 June 2015; pp. 1116–1124.
48. Li, W.; Wang, X. Locally aligned feature transforms across views. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 3594–3601.
49. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
50. Lv, J.; Chen, W.; Li, Q.; Yang, C. Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7948–7956.
51. Loy, C.C.; Xiang, T.; Gong, S. Multi-camera activity correlation analysis. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 1988–1995.
52. Ahmed, E.; Jones, M.; Marks, T.K. An improved deep learning architecture for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–11 June 2015; pp. 3908–3916.
53. Hermans, A.; Beyer, L.; Leibe, B. In defense of the triplet loss for person re-identification. *arXiv* **2017**, arXiv:1703.07737.
54. Weinberger, K.Q.; Saul, L.K. Distance metric learning for large margin nearest neighbor classification. *J. Mach. Learn. Res.* **2009**, *10*, 207–244.
55. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv* **2015**, arXiv:1502.03167.
56. Glorot, X.; Bordes, A.; Bengio, Y. Deep sparse rectifier neural networks. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, Ft. Lauderdale, FL, USA, 11–13 April 2011; pp. 315–323.
57. Baldassarre, F.; Morín, D.G.; Rodés-Guirao, L. Deep koalarization: Image colorization using cnns and inception-resnet-v2. *arXiv* **2017**, arXiv:1712.03400.
58. Xu, B.; Wang, N.; Chen, T.; Li, M. Empirical evaluation of rectified activations in convolutional network. *arXiv* **2015**, arXiv:1505.00853.
59. Zheng, L.; Bie, Z.; Sun, Y.; Wang, J.; Su, C.; Wang, S.; Tian, Q. Mars: A Video Benchmark for Large-Scale Person Re-Identification. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 868–884.
60. Cheng, D.; Gong, Y.; Zhou, S.; Wang, J.; Zheng, N. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In Proceedings of the iEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1335–1344.
61. Yuan, X.; He, P.; Zhu, Q.; Li, X. Adversarial examples: Attacks and defenses for deep learning. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 2805–2824. [CrossRef] [PubMed]
62. Hu, Y.; Yi, D.; Liao, S.; Lei, Z.; Li, S.Z. Cross Dataset Person Re-Identification. In Proceedings of the Asian Conference on Computer Vision, Singapore, 1–5 November 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 650–664.
63. Hirzer, M.; Beleznai, C.; Roth, P.M.; Bischof, H. Person Re-Identification by Descriptive and Discriminative Classification. In Proceedings of the 17th Scandinavian Conference on Image Analysis, Ystad, Sweden, 23–25 May 2011; Springer: Berlin/Heidelberg, Germany, 2011; pp. 91–102.
64. Liao, X.; He, L.; Yang, Z.; Zhang, C. Video-Based Person Re-Identification Via 3D Convolutional Networks and Non-Local Attention. In Proceedings of the Asian Conference on Computer Vision, Perth, Australia, 2–6 December 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 620–634.
65. Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. The kinetics human action video dataset. *arXiv* **2017**, arXiv:1705.06950.
66. Li, J.; Zhang, S.; Huang, T. Multi-scale 3d convolution network for video based person re-identification. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 8618–8625.
67. Zhou, Z.; Huang, Y.; Wang, W.; Wang, L.; Tan, T. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 June 2017; pp. 4747–4756.

68. Ge, Y.; Li, Z.; Zhao, H.; Yin, G.; Yi, S.; Wang, X. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. In Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montreal, QC, Canada, 3–8 December 2018; pp. 1222–1233.
69. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
70. Zhong, Z.; Zheng, L.; Zheng, Z.; Li, S.; Yang, Y. Camera style adaptation for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5157–5166.
71. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
72. Zou, Y.; Yang, X.; Yu, Z.; Kumar, B.V.; Kautz, J. Joint disentangling and adaptation for cross-domain person re-identification. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part II 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 87–104.
73. Fan, X.; Jiang, W.; Luo, H.; Fei, M. Spherereid: Deep hypersphere manifold embedding for person re-identification. *J. Vis. Commun. Image Represent.* **2019**, *60*, 51–58. [CrossRef]
74. Sun, Y.; Zheng, L.; Yang, Y.; Tian, Q.; Wang, S. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 480–496.
75. Wen, Y.; Zhang, K.; Li, Z.; Qiao, Y. A Discriminative Feature Learning Approach for Deep Face Recognition. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 499–515.
76. Zhong, Z.; Zheng, L.; Cao, D.; Li, S. Re-ranking person re-identification with k-reciprocal encoding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 June 2017; pp. 1318–1327.
77. Dietlmeier, J.; Antony, J.; McGuinness, K.; O'Connor, N.E. How important are faces for person re-identification? In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 6912–6919.
78. Lu, X.Y.; Skabardonis, A. Freeway traffic shockwave analysis: Exploring the NGSIM trajectory data. In Proceedings of the 86th Annual Meeting of the Transportation Research Board, Washington, DC, USA, 21–25 January 2007.
79. Geiger, A.; Lenz, P.; Stiller, C.; Urtasun, R. Vision meets robotics: The kitti dataset. *Int. J. Robot. Res.* **2013**, *32*, 1231–1237. [CrossRef]
80. Everingham, M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]
81. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
82. Bai, Y.; Lou, Y.; Gao, F.; Wang, S.; Wu, Y.; Duan, L.Y. Group-sensitive triplet embedding for vehicle reidentification. *IEEE Trans. Multimed.* **2018**, *20*, 2385–2399. [CrossRef]
83. Em, Y.; Gag, F.; Lou, Y.; Wang, S.; Huang, T.; Duan, L.Y. Incorporating intra-class variance to fine-grained visual recognition. In Proceedings of the 2017 IEEE International Conference on Multimedia and Expo (ICME), Hong Kong, China, 10–14 July 2017; pp. 1452–1457.
84. Zhang, Y.; Liu, D.; Zha, Z.J. Improving triplet-wise training of convolutional neural network for vehicle re-identification. In Proceedings of the 2017 IEEE International Conference on Multimedia and Expo (ICME), Hong Kong, China, 10–14 July 2017; pp. 1386–1391.
85. Liu, X.; Liu, W.; Mei, T.; Ma, H. Provid: Progressive and multimodal vehicle reidentification for large-scale urban surveillance. *IEEE Trans. Multimed.* **2017**, *20*, 645–658. [CrossRef]
86. Feng, W.; Hu, Z.; Wu, W.; Yan, J.; Ouyang, W. Multi-object tracking with multiple cues and switcher-aware classification. *arXiv* **2019**, arXiv:1901.06129.
87. Milan, A.; Leal-Taixé, L.; Reid, I.; Roth, S.; Schindler, K. MOT16: A benchmark for multi-object tracking. *arXiv* **2016**, arXiv:1603.00831.
88. Zhou, Y.; Liu, L.; Shao, L. Vehicle re-identification by deep hidden multi-view inference. *IEEE Trans. Image Process.* **2018**, *27*, 3275–3287. [CrossRef]
89. Zhang, S.; Wu, G.; Costeira, J.P.; Moura, J.M. Fcn-rlstm: Deep spatio-temporal neural networks for vehicle counting in city cameras. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3667–3676.
90. He, Z.; Lei, Y.; Bai, S.; Wu, W. Multi-Camera vehicle tracking with powerful visual features and spatial-temporal cue. In Proceedings of the CVPR Workshops, Long Beach, CA, USA, 16–20 June 2019; pp. 203–212.
91. Naphade, M.; Anastasiu, D.C.; Sharma, A.; Jagrlamudi, V.; Jeon, H.; Liu, K.; Chang, M.C.; Lyu, S.; Gao, Z. The nvidia ai city challenge. In Proceedings of the 2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), San Francisco, CA, USA, 4–8 August 2017; pp. 1–6.
92. Zhu, P.; Wen, L.; Bian, X.; Ling, H.; Hu, Q. Vision meets drones: A challenge. *arXiv* **2018**, arXiv:1804.07437.
93. Voigtlaender, P.; Krause, M.; Osep, A.; Luiten, J.; Sekar, B.B.G.; Geiger, A.; Leibe, B. MOTS: Multi-object tracking and segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7942–7951.

94. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In Proceedings of the 32nd Conference on Neural Information Processing Systems, Montreal, QC, Canada, 7–10 December 2015; pp. 91–99.
95. Zapletal, D.; Herout, A. Vehicle re-identification for automatic video traffic surveillance. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 25–31.
96. Liu, X.; Liu, W.; Ma, H.; Fu, H. Large-scale vehicle re-identification in urban surveillance videos. In Proceedings of the 2016 IEEE International Conference on Multimedia and Expo (ICME), Seattle, WA, USA, 11–15 July 2016; pp. 1–6.
97. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
98. Shen, X.; Lin, Z.; Brandt, J.; Avidan, S.; Wu, Y. Object retrieval and localization with spatially-constrained similarity measure and k-nn re-ranking. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, 16–21 June 2012; pp. 3013–3020.
99. Muja, M.; Lowe, D.G. Fast approximate nearest neighbors with automatic algorithm configuration. *VISAPP* **2009**, *2*, 2.
100. Alahi, A.; Ramanathan, V.; Fei-Fei, L. Socially-aware large-scale crowd forecasting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 2203–2210.
101. Sochor, J.; Špaňhel, J.; Herout, A. BoxCars: Improving Fine-Grained Recognition of Vehicles Using 3-D Bounding Boxes in Traffic Surveillance. *IEEE Trans. Intell. Transp. Syst.* **2018**, *20*, 97–108. [CrossRef]
102. Luiten, J.; Fischer, T.; Leibe, B. Track to reconstruct and reconstruct to track. *IEEE Robot. Autom. Lett.* **2020**, *5*, 1803–1810. [CrossRef]
103. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for autonomous driving? The kitti vision benchmark suite. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, 16–21 June 2012; pp. 3354–3361.
104. Bernardin, K.; Stiefelhagen, R. Evaluating multiple object tracking performance: The CLEAR MOT metrics. *EURASIP J. Image Video Process.* **2008**, *2008*, 1–10. [CrossRef]
105. Wojke, N.; Bewley, A.; Paulus, D. Simple online and realtime tracking with a deep association metric. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 3645–3649.
106. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
107. Zhang, H.; Sindagi, V.; Patel, V.M. Image de-raining using a conditional generative adversarial network. *IEEE Trans. Circ. Syst. Video Technol.* **2019**, *30*, 3943–3956. [CrossRef]
108. Schaefer, G.; Stich, M. UCID: An uncompressed color image database. In *Storage and Retrieval Methods and Applications for Multimedia 2004*; International Society for Optics and Photonics: Washington, DC, USA, 2003; Volume 5307, pp. 472–480.
109. Arbelaez, P.; Maire, M.; Fowlkes, C.; Malik, J. Contour detection and hierarchical image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *33*, 898–916. [CrossRef]
110. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]
111. Wang, Z.; Bovik, A.C. A universal image quality index. *IEEE Signal Process. Lett.* **2002**, *9*, 81–84. [CrossRef]
112. Sheikh, H.R.; Bovik, A.C. Image information and visual quality. *IEEE Trans. Image Process.* **2006**, *15*, 430–444. [CrossRef]
113. Kang, L.W.; Lin, C.W.; Fu, Y.H. Automatic single-image-based rain streaks removal via image decomposition. *IEEE Trans. Image Process.* **2011**, *21*, 1742–1755. [CrossRef] [PubMed]
114. Li, Y.; Tan, R.T.; Guo, X.; Lu, J.; Brown, M.S. Rain streak removal using layer priors. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2736–2744.
115. Fu, X.; Huang, J.; Ding, X.; Liao, Y.; Paisley, J. Clearing the skies: A deep network architecture for single-image rain removal. *IEEE Trans. Image Process.* **2017**, *26*, 2944–2956. [CrossRef] [PubMed]