mathematics

MDPI

Article

ReID-DeePNet: A Hybrid Deep Learning System for Person Re-Identification

Hussam J. Mohammed [1,*], Shumoos Al-Fahdawi [2], Alaa S. Al-Waisy [3], Dilovan Asaad Zebari [4], Dheyaa Ahmed Ibrahim [3], Mazin Abed Mohammed [5], Seifedine Kadry [6,7,8] and Jungeun Kim [9,*]

[1] Computer Center, University of Anbar, Ramadi 31001, Iraq
[2] Computer Science Department, Al-Ma'aref University College, Ramadi 31001, Iraq
[3] Computer Engineering Technology Department, Information Technology Collage, Imam Ja'afar Al-Sadiq University, Baghdad 10072, Iraq
[4] Department of Computer Science, College of Science, Nawroz University, Duhok 42001, Iraq
[5] College of Computer Science and Information Technology, University of Anbar, Ramadi 31001, Iraq
[6] Department of Applied Data Science, Noroff University College, 4612 Kristiansand, Norway
[7] Department of Electrical and Computer Engineering, Lebanese American University, Byblos P.O. Box 13-5053, Lebanon
[8] Artificial Intelligence Research Center (AIRC), College of Engineering and Information Technology, Ajman University, Ajman P.O. Box 346, United Arab Emirates
[9] Department of Software, Kongju National University, Cheonan 31080, Korea
* Correspondence: hussamjasim@uoanbar.edu.iq (H.J.M.); jekim@kongju.ac.kr (J.K.)

Abstract: Person re-identification has become an essential application within computer vision due to its ability to match the same person over non-overlapping cameras. However, it is a challenging task because of the broad view of cameras with a large number of pedestrians appearing with various poses. As a result, various approaches of supervised model learning have been utilized to locate and identify a person based on the given input. Nevertheless, several of these approaches perform worse than expected in retrieving the right person in real-time over multiple CCTVs/camera views. This is due to inaccurate segmentation of the person, leading to incorrect classification. This paper proposes an efficient and real-time person re-identification system, named ReID-DeePNet system. It is based on fusing the matching scores generated by two different deep learning models, convolutional neural network and deep belief network, to extract discriminative feature representations from the pedestrian image. Initially, a segmentation procedure was developed based on merging the advantages of the Mask R-CNN and GrabCut algorithm to tackle the adverse effects caused by background clutter. Afterward, the two different deep learning models extracted discriminative feature representations from the pedestrian segmented image, and their matching scores were fused to make the final decision. Several extensive experiments were conducted, using three large-scale and challenging person re-identification datasets: Market-1501, CUHK03, and P-DESTRE. The ReID-DeePNet system achieved new state-of-the-art Rank-1 and mAP values on these three challenging ReID datasets.

Keywords: person re-identification; deep learning; deep belief network; Mask R-CNN; GrabCut algorithm; Market-1501 dataset

MSC: 68T10

check for updates

Citation: Mohammed, H.J.; Al-Fahdawi, S.; Al-Waisy, A.S.; Zebari, D.A.; Ibrahim, D.A.; Mohammed, M.A.; Kadry, S.; Kim, J. ReID-DeePNet: A Hybrid Deep Learning System for Person Re-Identification. *Mathematics* **2022**, *10*, 3530. https://doi.org/10.3390/math10193530

Academic Editors: Abeer Alsadoon and Luis Coelho

Received: 16 August 2022
Accepted: 25 September 2022
Published: 28 September 2022

1. Introduction

Surveillance systems have been used immensely in various public and private areas such as airports, universities, schools, streets, houses, etc. These surveillance systems provide massive data, including images and videos, which are helpful in the investigation of criminal activities [1]. However, processing and analyzing these images and videos to track and monitor a person over non-overlapping cameras is a time-consuming and challenging task [2]. Several factors can significantly affect the performance of the person ReID system in practical applications, such as pose variations, illumination changes, occlusions, different camera settings, and background clutter [3]. All these factors cause the appearance of the same person to look extremely different. This leads to a heavy burden on investigators to identify the wanted person from many surveillance videos in a short time. Therefore, the person Re-ID task is still an unsolved problem that is worth further research. However, the ethical and privacy implications of surveillance biometric-based systems (e.g., person ReID) have become significantly critical, and have attracted increasingly critical attention [4]. Some direct questions have been raised concerning whether biometric systems offer society significant advantages over traditional methods of personal identification (e.g., passwords, ID cards, etc.), or whether it constitutes a threat to people's privacy. For instance, CCTVs are used in car parks and cities, and X-ray machines at airports to detect and prevent potential crimes against people or property. However, the dilemma facing the surveillance systems is data collection, and ensuring that collected data will not be used for purposes that are unethical or impinge upon human rights. Thus, the collected information should be protected by ethics and laws, except in specific circumstances (e.g., in a court of law) [5].

In general, the person ReID system is mainly based on three critical steps: automatic pedestrian detection, features extraction, and classification step. Most of the previously published works directly learn the feature representations from the whole pedestrian image, that contains background clutter. Quite recently, several person ReID deep learning-based systems have suggested learning effective feature representations from the detected pedestrian body to reduce the background clutter and improve the robustness of the person ReID system [6–8]. This motivates us to develop an automated image segmentation algorithm to eliminate background noise interference issues and enhance the discriminability of the extracted feature representations, even for an incomplete person, which may contain information that is discriminatory and deserves attention.

In the features extraction step, person ReID systems can be divided into either handcrafted-based systems or deep learning-based systems. Handcrafted-based systems are designed to extract invariant features (e.g., color and texture) for pedestrian description [9]. For instance, Zheng et al. [10] applied the SIFT descriptor to extract a feature vector of 128 values for pedestrian description and employed the bag of words (BOW) for person ReID. Klaser et al. [11] proposed the integration of the histograms oriented gradient (HOG) and histograms of optical flow (HOF) to introduce a 3D pedestrian descriptor, named HOG3D. Although the handcrafted-based descriptors have further improved the performance of person ReID systems, the massive amount of captured data using multiple cameras has made extracting common feature representations from the same pedestrian very hard due to these descriptors lacking a self-learning process. This promotes the appearance of deep learning-based systems for person ReID.

Recently, deep learning methods, for example, convolutional neural networks (CNN), have played an essential role in addressing person re-identification problems due to their ability to jointly handle occlusions, geometric transforms, illumination changes, and background clutter in a unified framework [12]. CNNs can extract discriminative and robust feature representations for either the whole or part body of a pedestrian's image. However, a tremendous amount of data is required to train ReID deep learning-based systems and achieve a satisfactory performance. Some examples of ReID deep learning-based systems can be found in [3,13,14]. Another critical step of person ReID is learning a robust distance or similarity function to address the problems of the matching pedestrian (e.g., heterogeneous face recognition). In this regard, metric learning methods have been developed to solve matching person problems, such as cross-view quadratic discriminant analysis (XQDA) [15], distance metric [16], etc.

In this paper, an automatic hybrid deep learning system is developed and named ReID-DeePNet system, to identify a person in real life using multiple CCTVs/street camera views. Initially, an efficient and reliable image segmentation procedure was developed based on integrating the advantages of the Mask R-CNN and GrabCut algorithm to tackle the adverse effects caused by background clutter. Then, the matching scores from two distinctive deep learning models based on CNN and deep belief network (DBN) models were obtained to establish the person's identity across multiple cameras.

1. An automated and fast image segmentation algorithm was proposed to eliminate background noise interference issues and enhance the feature representations' discriminability in the subsequent steps of the proposed ReID-DeePNet system. Herein, the MASK region-based CNN (Mask R-CNN) algorithm was applied to automatically extract the pixel-wise mask for foreground objects "pedestrians" out of the complex background. However, Mask R-CNN could not ideally detect the object of interest, such as the dynamic pedestrian body, and some parts of the background still appeared in the final segmented image. This could negatively affect the accuracy of the proposed system. Thus, the GrabCut algorithm was applied using the mask generated by Mask R-CNN as the initial seed to reduce the effects of background noise interference and enhance the person's body segmentation accuracy;
2. An effective and real-time person ReID system was developed based on integrating the matching scores generated from two different deep learning approaches, such as CNN and DBN, to extract discriminative feature representations from the pedestrian image. To the best of our knowledge, this was the first attempt to investigate the possibility of training a CNN and DBN from scratch, to address the person ReID problem in a unified system;
3. A parallel architecture for integrating the matching scores generated from the CNN and DBN model was considered that could give end users a high degree of flexibility in establishing a person's identity using the result obtained from one or both adopted models, based on the desired security level and the user's satisfaction. The performance of the proposed system was assessed using different fusion rules at the score level (e.g., sum rule (SR), weighted sum rule (WSR), product rule (PR), max rule, and min rule) and rank level (e.g., highest rank (HR), Borda count (BC), and logistic regression (LR));
4. The accuracy of the proposed ReID-DeePNet system was assessed by carrying out several comprehensive experiments on three large-scale and challenging ReID datasets, including the Market-1501, CUHK03, and P-DESTRE datasets. A new advanced Rank-1 identification rate and mAP were achieved using the ReID-DeePNet system on all the employed datasets.

The rest of this article is organized as follows: A review of the previous works is presented in Section 2, and the proposed framework of ReID-DeePNet in Section 3. The employed ReID datasets and the empirical results are discussed and explained in Section 4. Finally, conclusions and future work guidelines are outlined in the last section.

2. Related Work

Recently, many researchers have focused on addressing the person re-identification problem by developing ideal solutions that can help in recognizing the person's identity across multiple cameras. Many researchers have employed deep learning approaches to address the person ReID task by combining the feature extraction and classification stages in a unified system. For instance, Weilin et al. [17] developed a hybrid framework combining multilevel feature extraction and a multi-loss learning approach to obtain a high description of the pedestrian. The multilevel feature extraction process was achieved using a feature aggregation network (FAN) to extract multilevel attributes from different layers. The multi-loss learning process included two actions: verification and recognition, where the verification aimed to verify that the two images belonged to the same identity, and where the recognition aimed to specify the identity within each image. This was accomplished using recurrent comparative network (RCN) and global average pooling (GAP) algorithms. Their experiments were conducted using four datasets, including CUHK03, CUHK01, Market1501, and DukeMTMC-reID. The best Rank-1 rate of 84.7% and mAP of 65.8% were obtained on Market1501dataset.

Yutian et al. [18] proposed a Bayesian query expansion (BQE) algorithm to produce a new query from the initial ranking list. They suggested dividing the dataset into three mutually exclusive sets of data: a training set, gallery set, and testing set. Once the algorithm was trained on the training data, the algorithm calculated the probability of images within the initial gallery, producing actual matches. The images of actual matches were used to predict a single vector used to generate a new list as a query expansion process. Extensive experiments on four different datasets were carried out, including Market-1501, DukeMTMC-reID, CUHK03, and MARS. The highest Rank-1 rate of 85.24% and mAP of 69.79% were achieved on the Market-1501 dataset.

Yichao et al. [19] proposed a feature attention block to generate part-level representations for pedestrians. Their method provided a weight for each part of the pedestrian's body by finding various horizontal features. A deep CNN model was utilized in the method to learn discriminative feature representations to compute the distance between pairs of query images in the gallery set and generate a ranking list for each query person. The authors evaluated their method using three datasets, Market-1501, DukeMTMC-ReID, and CUHK03. The experiments showed that the best results were obtained on the Market-1501 dataset by achieving a Rank-1 rate of 93.5% and mAP of 81.8%. Li et al. [20] proposed two branches of CNN network architecture for person feature extraction purposes. These branches considered the global and local features based on loss functions that are commonly used in person re-ID. The highest Rank-1 rate of 93.8% and mAP of 84.6% were achieved on the Market1501 dataset. In the context of multi-modules methods, Xin et al. [21] developed a semi-supervised feature representation approach to obtain discriminative feature representations from pedestrian images across disjoined cameras. They suggested using various CNN models to generate different feature representations from a single labeled image within a dataset. A finely-tuned process was applied to each CNN's feature representation to simultaneously decrease the identification loss and verification loss. Afterward, a multi-view clustering process was utilized to classify the CNN's features into similar groups and dissimilar to different groups, thereby integrating the features into the proper representations. The multi-view clustering process also estimated pseudo labels for unlabeled images to produce a label for each image within the dataset. Two benchmark datasets, including Market1501 and the DukeMCMT-reID dataset, were employed in the conducted experiments. The best performance on the Market1501 dataset was obtained by achieving a Rank-1 rate of 75.2% and mAP of 52.6%.

Isobe et al. [22] investigated the ability to learn discriminative feature representations within the person image. They proposed a framework to reduce the noise with unlabelled images, transfer the knowledge that could be learned from the source to the target image, and add extra training constraints. Therefore, the cluster-wise contrastive learning algorithm (CCL) was utilized with progressive domain adaptation (PDA) followed by Fourier augmentation (FA). Their experiments were performed on various datasets, including Market-1501, Duke, and MSMT. Their results outperformed current state- of-the-art works by achieving a mAP of 8.1%, 9.9%, 11.4%, and 11.1%, on the Market-to-Duke, Duke-to-Market, Market-to-MSMT, and Duke-to-MSMT tasks, respectively. In terms of using low-resolution images, Xia et al. [23] developed a semi-supervised method based on the mixed-space super-resolution model (MSSR) to enhance a person's resolution. Then, a part-based graph convolutional network (PGCN) was performed to obtain discriminative feature representations from the pedestrian images. Their experiments were carried out on the Market1501, CUHK03, and MSMT17 datasets to evaluate the performance of the proposed methods. The results showed that they were able to identify the person with good accuracy compared with many semi-supervised methods, by achieving the highest Rank-1 rate of 73.2% and mAP of 49.8 on the Market-1501 dataset.

Recently, Wu et al. [24] suggested learning more distinctive features for person ReID by jointly optimizing the appearance feature and the information of the ranking context. The authors proposed a hybrid ranking framework composed of two streams for addressing the person ReID problems. In the first stream, the external ranking information was obtained by generating the ranking list for each probe image to learn visible changes among the top ranks of the gallery set. On the other hand, the internal ranking information was obtained using the fine-grained feature in the second stream. The performance of the proposed hybrid ranking framework was assessed using four ReID datasets, including the Market-1501, DukeMTMC-ReID, CUHK03 and MSMT17 datasets. The best performance was achieved on Market-1501 with 94.7 and 86.8, of Rank-1 and mAP respectively. Tang et al. [25] developed a novel harmonious multi-branch network (HMBN) with various stripes on different branches to learn more discriminative feature representations for person ReID. The authors replaced the uniform partition procedure with a horizontal overlapped partition to avoid losing important information within the local regions. The performance of the HMBN was assessed on three different ReID datasets, including DukeMTMC-ReID, CUHK03, and Market-1501. The highest Rank-1 rate of 95.58% and mAP of 94.21% were achieved on the Market1501 dataset.

Gu et al. [26] considered extracting clothes' irrelevant feature representation from the original RGB images of pedestrians. The performance of the developed clothes-based adversarial loss (CAL) was tested on a private ReID dataset, named CCVID dataset. The CAL achieved a Rank-1 rate of 82.6% and mAP of 81.3%. Yang et al. [27] addressed the problem of twin noise labels (TNL) in visible infrared person re-identification (VI-ReID), which refers to noisy annotation and correspondence. The authors developed a new approach for reliable VI-ReID, named DuAlly robust training (DART). DART is mainly based on computing the clean confidence of noisy annotations and rectifying the noisy correspondence with the estimated confidence. The performance of DART has outperformed five state-of-the-art methods using two ReID datasets: SYSU-MM01 and RegDB datasets.

Throughout this review, one can see that several studies in the literature were developed to tackle the person re-identification problem using different deep learning architectures. In general, the models trained in a supervised manner showed better performance than semi-supervised and unsupervised approaches, due to the labeled images playing a substantial role in improving the learning ability of the developed models in recognizing a person's identity across multiple cameras. Although most of the developed approaches have significantly reduced the effects of lighting and pose changes and achieved good performance, the discriminative power of the extracted feature representations may still be affected by background clutter. Thus, the generalization abilities of currently developed approaches are still far from being at an acceptable level in handling real person ReID issues. In this study, several advanced deep learning approaches are integrated to develop a competitive person ReID system.

3. The Proposed ReID-DeePNet System

This section describes the proposed ReID-DeePNet system for person ReID. As depicted in Figure 1, the overall structure of the proposed ReID-DeePNet system is composed of two modules: The background suppression module and the person Re-ID module. In the background suppression module, the issues of background noise interference are solved to enhance the discriminability of feature representations in the subsequent steps of the proposed system. Mask R-CNN is employed to automatically extract the pixel-wise mask for foreground objects "pedestrians" out of the complex background. Nevertheless, the Mask R-CNN algorithm cannot perfectly distinguish between the foreground and background in the input image during the segmentation process. Thus, the output of the Mask R-CNN algorithm is further enhanced using the GrabCut method to reduce the effects of the background noise interference and enhance the person's body segmentation accuracy. This is followed by identifying a pedestrian's identity by integrating the advantages of two distinctive deep learning models (CNN and DBN) to address the person ReID problem.

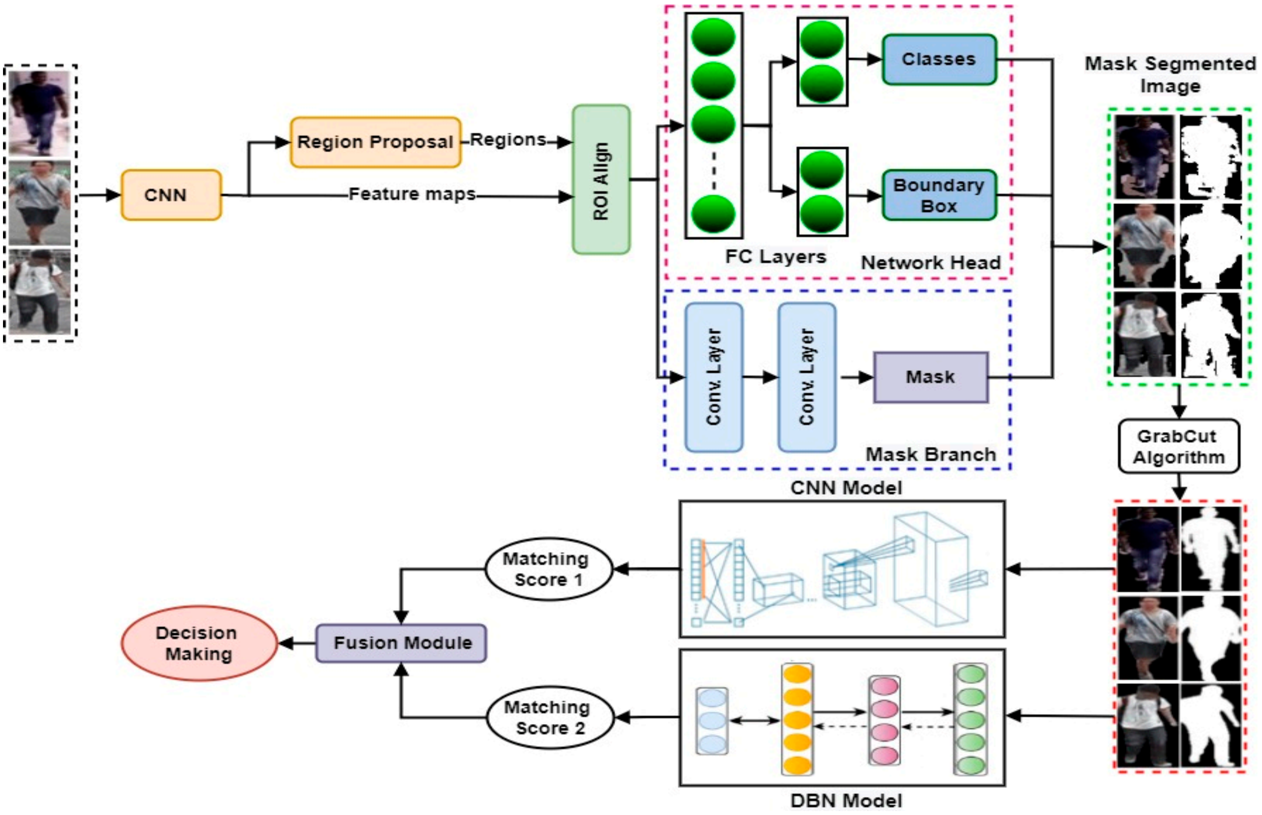

Figure 1. The overall structure of the proposed ReID-DeePNet system.

3.1. The Background Suppression Module

It was proved that using the pixel-wise mask of pedestrians can significantly reduce background clutter and improve the robustness of person ReID models under various background conditions. Furthermore, the generated pixel-wise masks have body shape information which is considered an important gait feature and useful for identifying a person, thus, boosting the identification accuracy [28,29]. The Mask R-CNN is an extension to the Fast R-CNN model that works on detecting objects in the input image and produces a binary mask for each object. Mask R-CNN aims to detect various objects in an image or video to produce the bounding box of an object along with its class label and binary mask. In other words, the Mask R-CNN consists of two stages including proposal generation of the potential objects within the input image, and prediction of the object's class then refinement of the surrounding box to generate the binary mask of the presented object within the first stage [30].

In addition, these two stages are connected using a backbone model to predict the person's class. In this study, the backbone component, based on a pre-trained ResNet101 model, is employed to extract more discriminative feature maps of the person from the input image. Then, the generated feature maps are transferred on to the feature pyramid network (FPN) to efficiently extract useful feature representations of different scales in the input image. The FPN uses the feature maps and semantic information to localize the region of object. The FPN also takes the benefits of the inherent and multiple scale nature of CNNs to gain a better detection of the person's object and to perform the semantic segmentation with various scales. This is achieved using the sliding window that applies on the generated feature maps to generate regions of persons within the image in the form of a bounding box. However, these proposals of bonding boxes come with different sizes, causing various issues in generating the segmented person within its mask. Therefore, the ROIAlign is employed to produce fixed feature maps with a unique form. Afterward, these fixed maps are passed into two fully connected layers within the network head component to produce the class of person and the boundary box of that person. In addition, the features maps will also pass into multi-convolutional layers within a mask component to obtain the mask of the segmented person. As a result, the output from the Mask R-CNN is represented by three components including the class of person, the boundary box, and the binary mask. However, the Mask R-CNN cannot perfectly separate the foreground objects from the complicated background during the segmentation process. In this study, the accuracy of the Mask R-CNN algorithm has been further improved using the predicted mask from Mask R-CNN as an initial seed to the GrabCut algorithm to reduce the effects of the background noise interference and enhance the person's body segmentation accuracy.

The GrabCut algorithm is an effective segmentation method used to remove undesired and heterogeneous edges of background from the segmented image of the person, and retain the foreground which is represented as a person's body [31]. Herein, the GrabCut algorithm utilizes the graph cuts method by drawing a boundary box around the foreground object of a person within the input image produced from the Mask R-CNN. Then, the Gaussian Mixture Model (GMM) is applied for estimating the color distribution of the foreground and background. The GMM then learns and predicts class labels for the unknown pixels based on the data from the input image, where each pixel is classified either as a foreground or background depending on its color statistics [32]. The GrabCut algorithm represents the input image as a graph by considering its pixels as vertices and the feature connection between these pixels as the edges (see Figure 2). The GrabCut algorithm loops on all the pixels within an image and breaks the weak connections between them, and then assigns each pixel to either the foreground or background. The implementation of the GrabCut algorithm on the top of the Mask R-CNN has significantly reduced the effects of the image's background and enhanced the segmentation accuracy and contour extraction of the person's body.

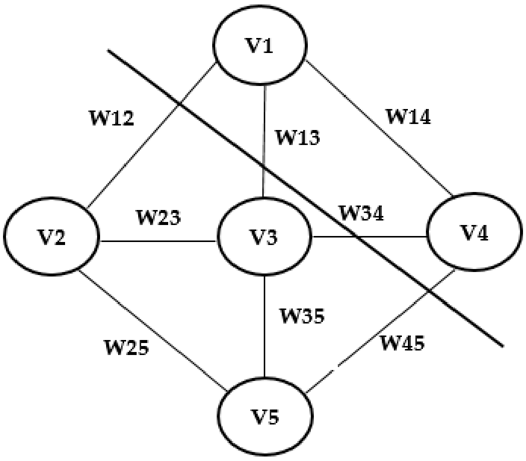

Figure 2. The graphical representation of the foreground and background objects using the GrabCut algorithm.

3.2. The Person Re-ID Module

The output of proposed image segmentation procedure in the background suppression module is the targeted person who should be classified in order to be tracked, based on his appearance. In this study, the person's identity is recognized using two powerful deep learning models (CNN and DBN) trained from scratch to address the person ReID problem. To the best of our knowledge, this is the first study that explores the possible use of CNN and DBN models in a unified person ReID system to extract distinctive local feature representations from pedestrian images. In the next sub-sections, the main architecture and the training methodology of the adopted deep learning models (CNN and DBN) are explained in detail.

3.2.1. CNN for Person ReID

As shown in Figure 3, the main structure of the employed CNN model comprises a combination of four locally-connected convolutional layers, each one followed by (2 × 2) sub-sampling max-pooling layer. Each convolutional layer has an assigned number of trainable filters to learn high-level feature representations from the pedestrian image. Herein, the number of trainable filters are set as 6, 20, 64, and 128, for the employed convolutional layer. In this work, two fully connected layers are employed on top of the proposed CNN model for the multi-class classification tasks. The output of the last fully connected layer is fed into the Softmax classifier, which computes the probability distribution over all of the class labels in the dataset being used to produce the predicted class label. Finally, a suitable loss function based on a cross-entropy is employed to measure the correspondence between the predicted and the target labels and compute the cost value for the proposed CNN model.

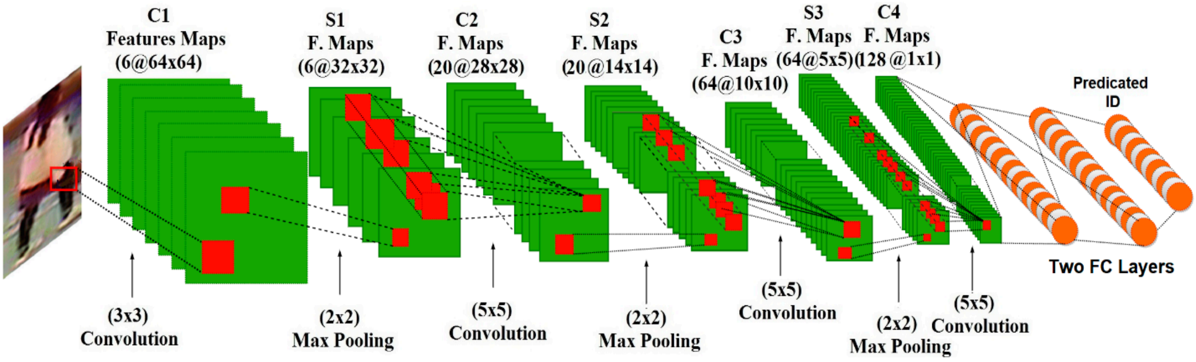

Figure 3. The architecture of the developed CNN model for addressing person ReID problems.

Following the same training procedure described [33], we start training a specific CNN model with a particular structure by splitting the training set into four sets. The CNN model is trained using the first three sets, and the last set is used as a validation set to assess the CNN model's generalization capacity during the learning process. The last trained CNN model with minimum validation error on the validation set is stored to report the real performance using the testing set. To prevent the overfitting problem, an early stopping procedure is applied by stopping the training process when the value of validation error on the validation set starts to increase again, for few times. Furthermore, some of the most widely used data augmentation techniques are implemented to reduce overfitting and enhance the generalization capability of the last trained CNN model during the learning process. In this work, five image regions are randomly cropped from each image in the training set along with their horizontally flipped versions. The main steps of the implemented training procedure can be defined as follows:

1. Divide the dataset into three sets (e.g., training set, validation set, and testing set);
2. Select a particular CNN structure and initialize the value of the hyper-parameters (e.g., number of epochs, learning rate, etc.);
3. Train the selected CNN model with the training set;
4. Assess the performance of the selected CNN model using the validation set during the learning process;
5. Repeat steps 3 through 4 using 300 epochs;
6. Save the weights of the best trained CNN model with less validation error on the validation set;
7. Report the actual performance of the saved CNN model using the testing set.

3.2.2. DBN for Person ReID

On the other hand, the DBN model consists of a single visible layer and multiple hidden layers connected with each other in a strong relationship to learn high-level feature representations from input data [34]. These layers are also utilized to learn the statistical associations between units of a previous layer where each unit within the proceeding layer is connected to all units of the earlier layer, as illustrated in Figure 4. The DBN is a stack of multiple layers of restricted Boltzmann machines (RBMs). RBM is a generative stochastic neural network consisting of two fully-connected layers using symmetric undirected edges with no links between nodes of the same layer. As shown in Figure 4, the employed DBN model for a person ReID is composed of stacking three RBMs as hidden layers. The first two hidden layers are trained one at a time as feature descriptors in a bottom-up fashion utilizing an unsupervised greedy layer wised (GLW) algorithm. Herein, the CD learning algorithm is employed. The applied DBN model's final hidden layer is trained as a discriminative model in combination with a Softmax classifier to perform the classification task. Herein, the suggested training procedure to train the employed DBN model can be outlined into three steps as follows:

1. As per the training procedure described in [35,36], the first two RBMs are trained one at a time using an unsupervised learning algorithm based on the CD learning algorithm. After the training process of the first RBM is finished, its activation outputs can be seen as features learned from the input image. Then, these feature representations are used as input data to train the next RBM in the stack. This unsupervised learning process enables us to train the network with massive amounts of unlabeled data to advance the generalization capability of the proposed DBN model. After finishing the training process of the first two RBMs, they can be seen as feature descriptors that can extract the most discriminative features from the raw images automatically;
2. The training and validation sets, together with their associated class labels, are used to train the last hidden layer in the proposed DBN as a non-linear classifier, which is used to monitor the learning process;
3. To improve classification accuracy, the weights of the whole network are fine-tuned in a top-down fashion using the back-propagation algorithm.

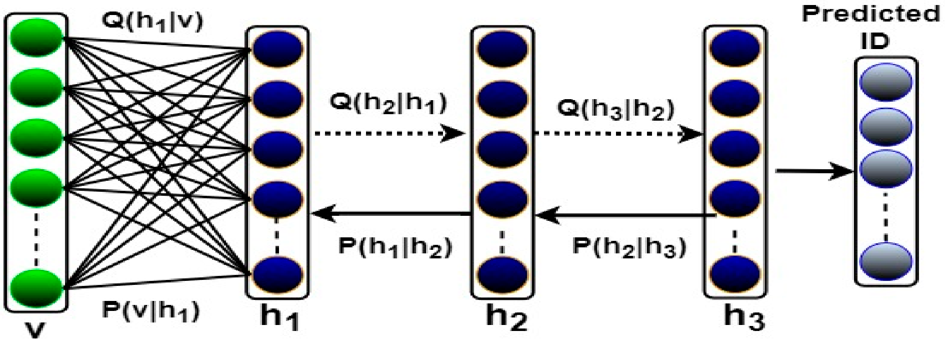

Figure 4. The main architecture of the developed DBN model for addressing person ReID problems.

4. Experimental Results

This section provides a description of the three large-scale and challenging ReID datasets that were employed to evaluate the effectiveness of the proposed ReID-DeePNet system. Then, implementation details of the proposed approaches in the background suppression and person Re-ID module are introduced. Next, the hyper-parameters analysis and visualization of the employed deep learning models are also presented to verify their effectiveness. Finally, we compare the performance of the proposed ReID-DeePNet system with advanced systems on these three datasets. The code of the ReID-DeePNet system was coded in Python programming language and all the experiments were conducted on the Google Colab server platform with 69K GPU graphics card, and 16 GB of RAM on the Windows 10 operating system, Intel(R) Core (TM) i7-4510U GHz CPU.

4.1. Datasets Description

The robustness of the proposed ReID-DeePNet system was tested on three large-scale and challenging Re-ID datasets, including the Market-1501 [10], CUHK03 [37], and P-DESTRE datasets [38]. These three employed Re-ID datasets reflected the main issues that influence person Re-ID in a real-world application, such as perspectives, changes of illumination, occasions, poses of pedestrians, etc. All the conducted experiments followed the standard evaluation protocol and data split setting of these three datasets. The performance evaluation metrics, such as the Rank-1 identification rate and mean average precision (mAP) were computed. Table 1 shows the statistics of the adopted three Re-ID datasets, and some samples from these datasets are shown in Figure 5.

Table 1. The statistics of the adopted three Re-ID datasets.

Datasets	No. Images	No. Identities
Market-1501 [10]	32,688	1501
CUHK03 [37]	13,164	1360
P-DESTRE dataset [38]	—	269

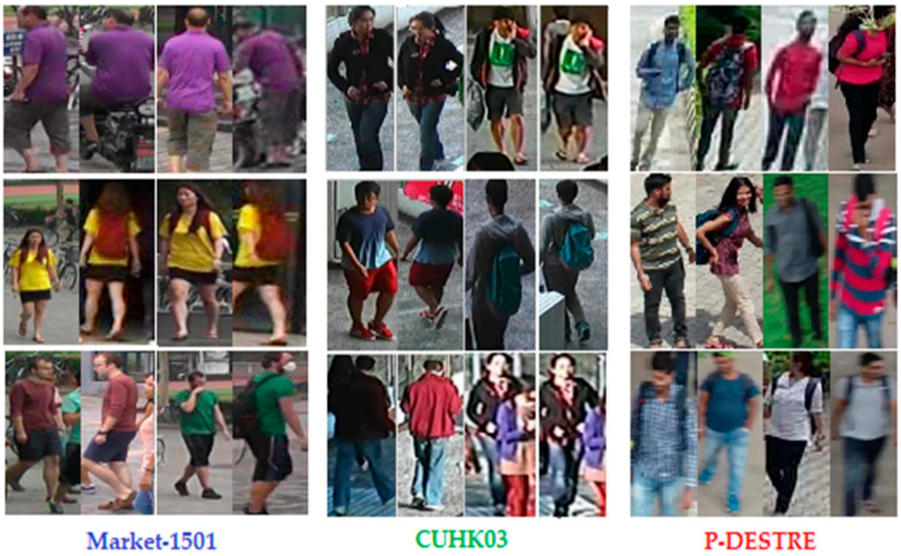

Figure 5. Sample images from the Market-1501, CUHK03, and P-DESTRE datasets.

- **Market-1501 dataset** [10] is a public benchmark dataset containing 1501 identities that were collected by six cameras from different viewpoints. The total number of pedestrian images was 32,688, with approximately 3.6 images on average for each identity from different viewpoints. In addition, all images were in .jpg format. A deformable part models (DPM) pedestrian detector was used to extract and detect the pedestrian within the collected images. Following the standard evaluation protocol, the Market-1501 dataset was divided into two sets, with 750 for training set (e.g., 17.2 images per identity) and 751 for testing set. Thus, all the 12,936 images were used to train the proposed ReID-DeePNet system;
- **CUHK03 dataset** [37] is also a public dataset composed of 1360 identities with 13,164 images in .jpg format. Six surveillance cameras were utilized to capture these images and each two disjoined cameras produced 4.8 images on average for each identity. The captured images within the CUHK03 dataset contained various variations, such as illumination, direction of pedestrians, different cameras settings, etc. Following the training and testing splits described in [37], the dataset was divided into two sets: the training set had 767 IDs, while the testing set contained the remaining 700 identities;

- **P-DESTRE dataset** [38] contained a total of 75 videos and individual tracks sequences with a resolution of (3840 × 2160) pixels. The cameras used to capture these videos were attached to several UAVs. The dataset included videos acquired at altitudes of 5.5 and 6.7 m over many days in crowded outdoor settings. Although most of the bounding boxes included humans with an acceptable resolution, this was not always the case when people were caught from a distance (distances exceeding 40 m), which resulted in low resolution and blur in some situations. Some of the frames had motion blur problems because of the UAVs' fast movements and low altitude. The proposed ReID-DeePNet system's effectiveness on this dataset was evaluated by computing the mean and standard deviation of the results across all five splits, which had five predefined splits of test data and training data. A 10-fold cross validation strategy was employed for the P-DESTRE set, with the data in each split being randomly split into 60% for the training set (45 videos), 20% for the validation set (15 videos), and 20% for the testing set (15 videos).

4.2. The Background Suppression Module Evaluation

The experimental hypothesis of image segmentation was to locate and segment a person in the image among various objects, such as vehicles, trees, animals, etc. In addition, the person could appear in a small part of the image, such as their upper/lower body. Therefore, an automated segmentation step was desperately required to locate the person within the image and apply accurate classification. Herein, the Mask R-CNN was applied as an effective and reliable approach to detect a person's body across multiple cameras. Although the Mask R-CNN has shown encouraging results, it still had some parts of the image's background appearing in the final segmented image that could significantly degrade the accuracy of the developed system. Therefore, the effects of the background noise interference were eliminated by employing the Mask R-CNN as an initial seed to the GrabCut algorithm as a post-process step of the person segmentation procedure. Several experiments were conducted based on different network backbones within the Mask R-CNN, such as ResNet34, ResNet5o, ResNet101, and VGG19. These experiments were repeated with the same network backbones based on merging the advantages of the Mask R-CNN and GrabCut methods to prove the effectiveness of the proposed background suppression module, as illustrated in Table 2. In these experiments, two common evaluation metrics were calculated, including cumulative match characteristic (CMC) which denoted as Rank-1 accuracy, and mAP. All the experiments were carried out using the pre-trained ResNet50 model in the classification stage.

Table 2. Performance comparison of person Re-ID accuracy (%) using a different network backbone for Mask R-CNN.

Methods	**Network Backbone**	**Market-1501**		**CUHK03**		**P-DESTRE**	
		Rank-1	**mAP**	**Rank-1**	**mAP**	**Rank-1**	**mAP**
Mask R-CNN	ResNet34	69.61	59.22	64.22	63.31	43.14 ± 12	35.43 ± 13.2
	ResNet50	56.64	45.23	66.24	55.34	45.67 ± 9.7	34.56 ± 11.8
	ResNet101	84.24	76.34	69.56	65.87	80.89 ±7.1	71.03 ± 8.9
	VGG19	71.64	65.54	70.74	72.24	69.93 ± 8.8	65.95 ± 9.5
Mask R-CNN + GrabCut	ResNet34	76.21	65.25	68.34	65.11	52.64 ± 11.1	50.35 ± 9.8
	ResNet50	76.61	**81.02**	67.22	63.31	59.90 ± 9.7	45.65 ± 10.7
	ResNet101	**84.51**	80.98	**79.11**	**77.89**	**85.67 ± 6.3**	**74.89 ± 6.3**
	VGG19	73.21	74.18	69.99	68.49	76.81 ± 10.3	67.32 ± 8.1

From Table 2, one can see that using only the Mask R-CNN method with ResNet34 and ResNet50 as network backbones demonstrated a varying accuracy among the other network backbones. In terms of the Rank-1 rate, the Mask R-CNN method provided a higher accuracy using ResNet50 on the CUHK03 and P-DESTRE datasets, compared with the ResNet34 model, by achieving a Rank-1 rate of 66.24% and 45.67%, respectively. In contrast, the ResNet34 presented a better accuracy on the Market-1501 dataset by achieving a Rank-1 rate of 69.61%. Moreover, it was obvious that using ResNet34 model obtained higher mAP values on Market-1501 and CUHK03 datasets by achieving a mAP of 59.22% and 63.31, respectively. However, using Mask R-CNN along with the ResNet101 model, better segmentation accuracy was obtained, compared with the other three models across all the employed datasets. However, a slightly higher mAP value was obtained using the VGG19 model as a network backbone on the CUHK03 dataset. In general, the best segmentation accuracy was obtained using the ResNet101 model as a network backbone by producing a Rank-1 rate of 84.24%, 72.56%, and 80.89% on the Market-1501, CUHK03 and P-DESTRE datasets, respectively.

Although higher results were obtained on the CUHK0 dataset using the VGG19 model by achieving a Rank-1 rate of 70.74% and mAP of 72.24%, inferior results were obtained on the other two datasets compared with the ResNet101 model. On the other hand, one can see that the overall results in terms of Rank-1 rate, and mAP were further improved by merging the advantages of the Mask R-CNN and GrabCut algorithm. However, a slightly lower Rank-1 rate of 69.99% and mAP of 68.49% were obtained using the VGG19 model as a network backbone on the CUHK03 dataset. Generally, the highest Rank-1 rates of 84.51%, 79.11%, and 85.67% were acquired using the ResNet101 model as a network backbone on the Market-1501, CUHK03 and P-DESTRE datasets, respectively. However, a slightly higher mAP value of 81.02% was acquired using the ResNet50 model on the Market-1501 dataset compared with inferior results on the other two datasets by achieving a mAP of 63.31% and 45.65% on the CUHK03 and P-DESTRE datasets, respectively. Figure 6 shows some results of applying the proposed person's body segmentation procedure using Mask R-CNN (e.g., using the ResNet101 model as a network backbone) and the GrabCut algorithm on the Market-1501dataset. Furthermore, some examples of the created attention masks using the proposed background suppression module are shown in Figure 7. The proposed background suppression module could effectively focus on several unique parts of the human body and eliminated background noise interference to significantly improve the accuracy of the subsequence steps of the proposed system.

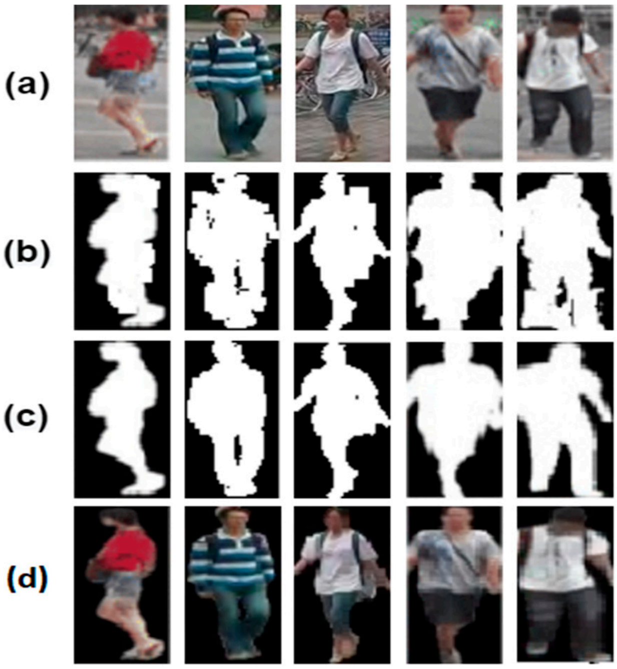

Figure 6. Some examples of applying the proposed person's body segmentation procedure: (**a**) original images, (**b**) body mask generated from the Mask R-CNN, (**c**) body mask generated from the enhanced Mask R-CNN using the GrabCut algorithm, and (**d**) the final detected person's body.

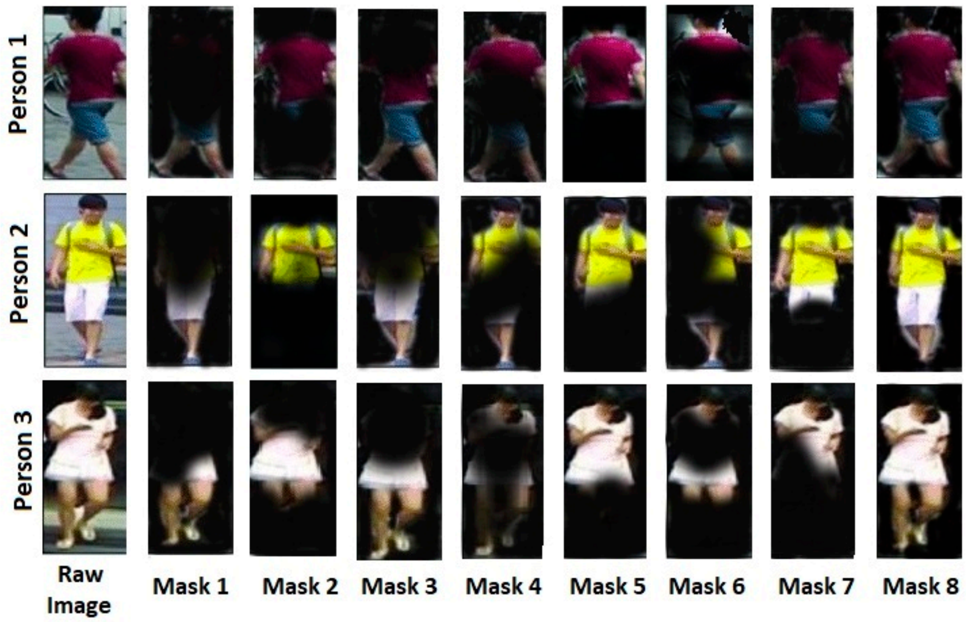

Figure 7. Some examples of the created attention masks using the proposed background suppression module.

4.3. Person Re-ID Module Evaluation

This section seeks to validate the automated approach of classification in finding the person of interest based on the given query image. In this study, we examine two powerful deep learning models, including the CNN model and the DBN model trained from scratch, on top of the output of the proposed segmentation procedure. All experiments were carried out on the three ReID datasets described above to finely-tune all the hyper-parameters of each model.

4.3.1. The Evaluation of the CNN Model

In this section, a set of comprehensive experiments conducted to find the optimal CNN model for the person Re-ID system, are presented. In these experiments, the effects of some hyper-parameters and a set of CNN architectures were assessed to find the optimal CNN model with optimal values of hyper-parameters to address the person Re-ID problem. Initially, the influence of the learning rate values was assessed using the AdaGrad optimization method. Using the suggested training methodology for the CNN model, an initial value of learning rate was set as 0.001. However, it was noticed that the CNN model took a long time to converge during the learning process due to the value of the learning rate being too small, and it was continuously reduced after each epoch using the AdaGrad optimization method. Thus, an initial value of the learning rate of 0.01 was set for all the remaining experiments. At the same time, the first number of epochs was set as 100, and using the same training methodology, the performance of larger numbers was also tested, including 200, 300, and 400 epochs. It was observed that if the CNN model was trained with a larger number of epochs than 100 epochs, its performance improved on the validation set. However, the CNN model started overfiting the training data and its performance on the validation set started to decline when it trained 400 epochs. As a result, the number of epochs was set as 300 epochs for all remaining experiments as the last trained CNN model had a good generalization ability without overfitting the training data. Table 3 shows the values of the employed hyper-parameters for the best obtained CNN model.

Table 3. The values of the hyper-parameters for the best obtained CNN model for addressing person ReID problems.

Hyper-Parameters	Values
No. of Conv. Layers	4
No. of Max-Pooling Layers	3
Optimization Method	Adagrad
Activation Function	ReLU
Momentum	0.90
Weight-decay	0.0002
Dropout	0.5
Batch Size	64
Learning Rate	0.01
Total No. of Epochs	300

In addition, image size plays an important role in the training speed and accuracy of the CNN model. Herein, the image size was set as 64 × 64 pixels, as the quality of the image becomes very poor for a lower image size, while a larger image size can require higher memory requirements and higher computational costs. A zero-padding of 1 pixel was applied only to the input layer of the proposed CNN to avoid a rapid decline in the amount of input data. On the other hand, to prevent the proposed CNN model from overfitting the training set, a dropout method was employed by ignoring the individual nodes within each training iteration. The dropout probability within each iteration was set to 0.5 to reduce the complexity of nodes co-adaptation by avoiding interdependency emerging between the nodes. The ReLU was employed as an activation function on the top of the convolutional and fully connected layers. The aim of the ReLU activation function was to increase the non-linearity of the CNN model. Based on knowledge from previous works the values of the weight decay, momentum, and batch size, were set to 0.0002, 0.9, and 64, respectively. Table 3 illustrates hyper-parameters that were employed in the best CNN model.

As shown in Table 4, several comprehensive experiments were conducted using various network architectures on Market-1501, CUHK03, and P-DESTRE datasets to obtain the best CNN architecture for personal Re-ID purposes. Initially, the CNN model used three layers with a different number of filters of each layer, such as 6, 20, and 32. The proposed model presented poor results across all the employed datasets for the Rank-1 rate and mAP accuracy. Afterward, the filter configuration of the third layer was duplicated to become 64 filters instead of 32. It was observed that the Rank-1 rate and mAP were enhanced by roughly 15%, compared with the previous setting. As a result, it was obvious that the number of filters in each convolutional layer had a strong impact on the accuracy of the CNN model. Thus, the number of filters within the second layer was also increased to become 32 filters instead of 20. One can see that the overall performance of the CNN model improved on the Market-1501 and P-DESTRE datasets. However, slightly lower values of Rank-1 rate and mAP on the CUHK03 dataset were obtained, by achieving 72.91% and 68.98%, respectively. Furthermore, it was also noticed that the accuracy of the CNN model was enhanced as we added more layers and increased the number of filters within the convolutional layer. From Table 4, the overall results in terms of Rank-1 rate and mAP were significantly improved for all adopted ReID datasets by adding a new convolutional layer on the top of the CNN model. As shown in Figure 3, we chose the last CNN architecture (6, 20, 64, and 128) in Table 4 as the adopted CNN architecture for recognizing a person's identity due to it providing the highest Rank-1 rate and mAP values for all the three datasets. As shown in Figure 8, the performance of the best CNN model for person Re-ID tasking on three different datasets is expressed via the CMC curves.

Table 4. Results for several CNN models utilizing images with a 64 × 64 pixel size from three ReID datasets. Each CNN model has 3 or 4 layers and shows the number of filters in each layer.

Network Architecture	Market-1501		CUHK03		P-DESTRE	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
[6,20,32]	56.64	45.23	66.24	55.34	57.32 ± 11.1	51.67 ± 10
[6,20,64]	76.58	72.34	75.18	73.45	78.41 ± 9.5	74.98 ± 8.2
[6,32,64]	82.81	80.45	72.91	68.98	83.91 ± 7.8	79.87 ± 6.4
[6,20,32,64]	87.21	88.34	87.87	84.23	89.15 ± 6.8	81.76 ± 6.2
[6,20,64,128]	**98.65**	**94.78**	**96.08**	**94.89**	**93.94 ± 5.5**	**92.95 ± 4.5**

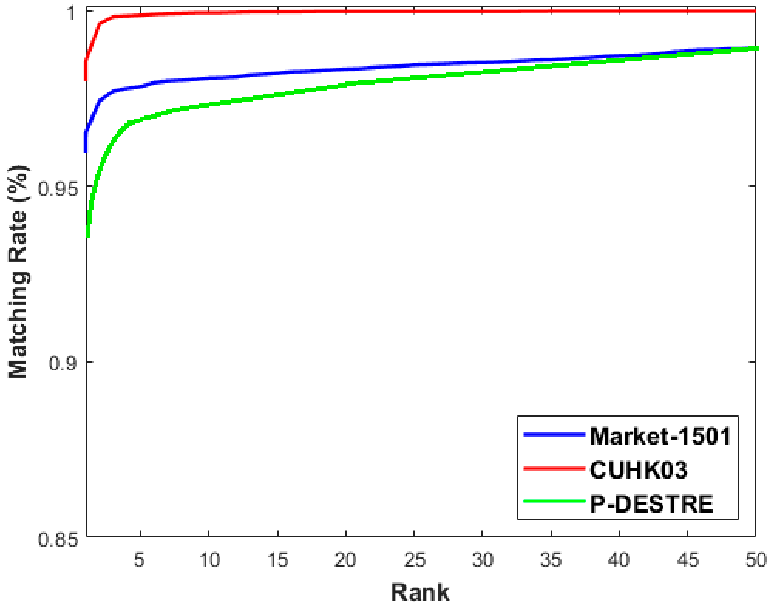

Figure 8. CMC curves for the best trained CNN model over three different ReID datasets.

4.3.2. The Evaluation of the DBN Model

The number of DBN architectures and hyper-parameters that need to be verified, such as the number of RBMs and the number of units per RBM, the number of epochs, the learning rate, etc., make the process of training a DBN model from scratch a challenging and difficult task. In training process of a DBN model, the value of a particular parameter could be affected by the values set for other hyper-parameters, which could be affected by the values set for other hyper-parameters. Additionally, the hyper-parameter values set in one RBM may depend on the values set in other RBMs in the stack. Consequently, the fine-tuning process of the hyper-parameter in a DBN model is quite expensive. Herein, a coarse search for all possible values was employed to carry out the fine-tuning procedure to identify the optimal hyper-parameter values. Using the training methodology described before, the DBN model was trained from scratch in a greedy manner using different numbers of hidden units per each RBM. After the training process of a specific RBM was finished, its weights matrix was preserved, and its activations were utilized as input to train the following RBM in the stack.

In this work, an initial DBN model composed of three hidden layers with a different number of hidden units (e.g., 1024-1024-1024) was greedily trained in a bottom-up fashion to assess the different values of the hyper-parameters. The first two hidden layers (RBMs) were trained separately in an unsupervised manner utilizing the CD learning algorithm using one-step Gibbs sampling (CD-1). The first two hidden layers were trained for 200 epochs, a weight decay of 0.0005, a momentum of 0.91, and mini-batch size of 64. The value of the learning rate was set to 0.001 for each RBM model, but it was noticed that the RBMs models needed a long time to converge because the learning rate value was very small. Thus, the learning rate value was set as 0.01 for all the remaining experiments. Next, the discriminative performance of the last RBM model was assessed by training it in supervised manner as a non-linear classifier. The last RBM model was trained using the same hyper-parameter values as the first two RBM models, with the exception that it was trained for 300 epochs. Finally, to minimize overfitting issues and improve the generalizability of the last trained DBN model, the complete network was trained in a top-down fashion using the back-propagation algorithm supported by the dropout technique. The dropout ratio was set to 0.5. The early stopping procedure was employed to determine the number of epochs during the fine-tuning phase, which was around 500 epochs. The values of the hyper-parameters for the best obtained DBN model are listed in Table 5.

Table 5. The values of the hyper-parameters for the best obtained DBN model for addressing person ReID problems.

Hyper-Parameters	Values
CD Learning Algorithm	1 step of Gibbs sampling
No. of Layers	3 RBMs
No. of Epochs for Each RBMs	200
Momentum	0.91
Weight-decay of	0.0005
Dropout	0.5
Batch Size	64
Learning Rate	0.01
Total No. of Epochs (BW)	500

Using the hyper-parameters shown in Table 5, different experiments were conducted by training a DBN model composed of three layers, but with a different number of hidden units per layer on the top of the segmented images generated from three different datasets. As shown in Table 6, four DBNs models were trained using a different number of hidden units per layer, ranging from 1024 to 3048 units. These models received the input image size of 64 × 64 pixels for all datasets. The first DBN model was composed of three hidden layers with the same number of hidden units (e.g., 1024). This DBN model presented the lowest accuracy among the other models, in terms of Rank-1 rate and mAP on all the employed datasets. Therefore, the number of hidden units within the second hidden layer was increased (to become, e.g., 2024). Notably, better results were obtained compared with the previous one, by achieving Rank-1 rates of over 80% on the Market-1501 and P-DESTRE datasets. Another experiment was also conducted using the DBN model, composed of three hidden layers with the number of hidden units set to 3048, 2024, and 1024, respectively. One can see that the overall performance of the last trained DBN model was significantly improved, by achieving the highest Rank-1 rates of 96.86%, 93.97%, and 91.81%, and mAP of 97.85%, 92.04%, and 87.94%, on the Market-1501, CUHK03 and P-DESTRE datasets, respectively. However, by increasing the number of the hidden units in last layer to 2024 units, it was observed that the values of the Rank-1 rate and mAP were reduced by approximately 10% compared with the third DBN model in Table 6. Therefore, the third DBN model was adopted in all the remaining experiments to identify the person's identity using the proposed ReID-DeePNet system (See Figure 4). As shown in Figure 9, the performance of the best DBN model for person Re-ID task on three different datasets is expressed via the CMC curves.

Table 6. Results acquired for different DBN models using the image size of 64 × 64 pixels from three different ReID datasets.

Network Architecture	Market-1501		CUHK03		P-DESTRE	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
DBN (1024-1024-1024)	65.24	56.45	53.56	55.33	58.98 ± 8.9	56.89 ± 7.6
DBN (1024-2024-1024)	81.56	80.34	71.33	63.55	80.91 ± 6.3	72.78 ± 7.5
DBN (3048-2024-1024)	**96.86**	**97.85**	**93.97**	**92.04**	**91.81 ± 4.5**	**87.94 ± 5.5**
DBN (3048-2024-2024)	84.89	85.68	83.69	80.81	83.91 ± 7.1	80.35 ± 6.7

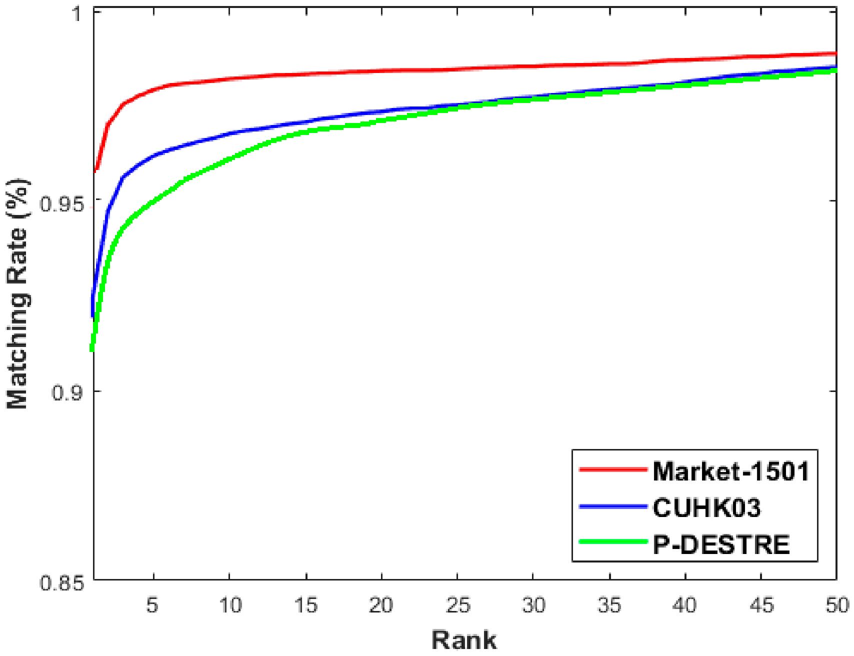

Figure 9. CMC curves for the best trained DBN model over three different ReID datasets.

4.4. The Evaluation of Fusion Rules

Using the proposed ReID-DeePNet system as a personal Re-ID system each time a query image is presented, **N** matching scores were generated from two different deep learning models (CNN and DBN models). These matching scores were either fused directly using one of matching scores rules (e.g., SR, WSR, PR, max, and min rule) or sorted in descending order to generate the ranking list of matching identities, which was fused using one of the ranking rules (e.g., HR, BC, and LR) to make the final decision. As can be seen from Tables 7 and 8, that the best results were obtained using the WSR rule in the matching score level by achieving Rank-1 rates of 99.91%, 98.92%, and 99.69%, and mAP of 99.67%, 98.34%, and 94.79%, on the Market-1501, CUHK03 and P-DESTRE datasets, respectively. In this work, using the WSR rule, a highest weight was given to the CNN model in making the final decision, due to its better performance compared with the performance of the DBN model on all the employed datasets. On the other hand, the BC rule in the ranking level achieved the highest mAP of 98.56% on the CUHK03 dataset. However, as shown in Tables 9 and 10, the HR rule produced the highest results compared with other ranking rules by achieving Rank-1 rates of 99.54%, 98.67%, and 94.85%, and mAP of 98.01%, 97.89%, and 93.95%, on the Market-1501, CUHK03 and P-DESTRE datasets, respectively.

Table 7. Rank-1 identification rate of the developed ReID-DeePNet system on three different ReID datasets using score-level fusion.

Datasets	CNN	DBN	Score Fusion Methods				
			SR	WSR	PR	Max	Min
Market-1501	98.65	96.86	98.76	**99.91**	99.01	98.11	98.23
CUHK03	96.08	93.97	96.89	**98.92**	97.87	97.98	97.85
P-DESTRE	93.94 ± 5.5	91.81 ± 5.8	94.26 ± 5.1	**94.79 ± 4.3**	93.96 ± 6.7	94.14 ± 7.1	93.98 ± 7.4

Table 8. mAP of the developed ReID-DeePNet system on three different ReID datasets using score-level fusion.

Datasets	CNN	DBN	Score Fusion Methods				
			SR	WSR	PR	Max	Min
Market-1501	94.78	97.85	98.19	**99.67**	97.89	98.02	98.12
CUHK03	94.89	92.04	97.34	**98.34**	98.01	96.34	97.01
P-DESTRE	92.95 ± 4.5	87.94 ± 5.5	93.18 ± 7.6	**94.15 ± 4.4**	92.89 ± 6.6	93.16 ± 6.8	93.23 ± 5.7

Table 9. Rank-1 identification rate of the developed ReID-DeePNet system on three different ReID datasets using rank-level fusion.

Datasets	CNN	DBN	Rank Fusion Methods		
			HR	BC	LR
Market-1501	98.65	96.86	**99.54**	99.11	99.59
CUHK03	96.08	93.97	**98.67**	98.34	98.23
P-DESTRE	93.94 ± 5.5	91.81 ± 4.5	**94.85 ± 5.2**	**94.15 ± 6.4**	**93.94 ± 7.1**

Table 10. mAP of the developed ReID-DeePNet system on three different ReID datasets using rank-level fusion.

Datasets	CNN	DBN	Rank Fusion Methods		
			HR	BC	LR
Market-1501	94.78	97.85	98.01	**98.56**	98.23
CUHK03	94.89	92.04	**97.89**	96.78	97.45
P-DESTRE	92.95 ± 4.5	87.94 ± 5.5	**93.95 ± 4.6**	93.53 ± 6.7	93.66 ± 6.6

4.5. Comparison Study and Discussion

The performance of the proposed ReID-DeePNet system was compared against other existing state-of-the-art personal Re-ID systems. For a fair comparison, Rank-1 and mAP values on all three datasets were reported. As shown in Table 11, the proposed ReID-DeePNet system outperformed all state-of-the-art personal Re-ID systems in terms of Rank-1 rates and mAP, using WSR and HR, on all the employed datasets. It is worthwhile noting that the accuracy of the proposed ReID-DeePNet system using WSR on all three ReID datasets, was higher than its performance using HR. Another observation was that the accuracy of the proposed ReID-DeePNet system on the P-DESTRE dataset was lower compared with the other datasets.

In general, the obtained results indicate that learning effective feature representations from the detected pedestrian body can significantly reduce background clutter and improve accuracy of the proposed ReID-DeePNet system. However, despite good preparation of the employed ReID datasets, the accuracy of detecting pedestrians' bodies cannot reach an optimal level, since it depends on image contrast, pose variations, illumination changes, occlusions, different camera settings, the location of the objects within the image, and the effect of the overlapped objects. Thus, pedestrian detection accuracy may be further improved by investigating the possibility of applying many methods, such as single shot multibox detector (SSD) [39], YOLO [40], and fast R-CNN [41] along with the GrabCut algorithm to obtain an accurate segmentation of individuals within the image.

Table 11. Performance comparison with state-of-the-art approaches on three large-scale and challenging ReID datasets.

Methods	Market-1501		CUHK03		P-DESTRE	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
DPA [6]	94.14	90.31	63.04	61.73	—	—
SegHAN [8]	92.3	76.1	88.3	—	—	—
RANGEv2 [24]	94.7	86.8	64.3	67.4	—	—
HMBN (RK)[25]	95.58	94.21	84.16	82.64	—	—
Siamese [28]	83.79	74.33	50.14	50.21	—	—
PAP-S-PS [42]	94.6	85.6	72.5	66.8	—	—
GoogLeNet [43]	81.0	63.4	85.4	—	—	—
HPM [44]	94.2	82.7	63.9	57.5	—	—
EDAAN [45]	95.3	86.8	94.7	83.4	—	—
DSA-reID [46]	95.7	87.6	78.9	75.2	—	—
M^{3}L (IBN-Net50) [47]	75.9	50.2	33.1	32.1	—	—
+NFormer [48]	95.7	93.0	80.6	79.1	—	—
COSAM [38]	—	—	—	—	80.2 ± 12.9	80.6 ± 11.9
GLTR [38]	—	—	—	—	81.0 ± 12.5	79.7 ± 12.0
OSNet [49]	—	—	—	—	82.9 ± 7.7	84.0 ± 7.4
Deep SORT + OSNet [50]	—	—	—	—	77.9 ± 5.1	70.5 ± 4.8
ReID-DeePNet (WSR)	**99.91**	**99.67**	**98.92**	**98.34**	**94.79 ± 4.3**	**94.15 ± 4.4**
ReID-DeePNet (HR)	**99.54**	**98.01**	**98.67**	**97.89**	**94.85 ± 5.2**	**93.95 ± 4.6**

In this study, the processes of learning discriminative feature representations and producing the final matching scores were jointly optimized using two powerful deep learning models, the CNN and the DBN models. These two deep learning models were trained from scratch, using the top of the detected pedestrian body instead of the whole raw pedestrian image. Herein, a parallel architecture was used to combine the matching scores acquired from the adopted models, providing a high degree of flexibility in establishing the person's identity. The results from the proposed ReID-DeePNet system are encouraging, especially given that they were derived from three different ReID datasets made up of more than 1000 IDs and a significant number of pedestrian images, which is relevant to real-world applications. Therefore, we believe that the proposed ReID-DeePNet system can be readily used for real-time application. Nevertheless, it should be pointed out that at the current stage of work, the proposed ReID-DeePNet system has not yet been applied in any real commercial application.

5. Conclusions and Future Work

In this paper, an efficient and real-time ReID-DeePNet system was proposed to match a person across non-overlapping cameras by various viewpoints. This system combined the Mask R-CNN followed by the GrabCut algorithm to obtain an accurate segmentation of individuals within the image. The developed segmentation approach worked in an automated method to obtain the person from among other objects. Afterward, a fusion module based on CNN and DBN was also developed to extract discriminative and robust features, thereby obtaining a correct classification. The effectiveness and robustness of the ReID-DeePNet system was tested on three challenging ReID datasets, namely, Market-1501, CUHK03, and P-DESTRE datasets. It produced higher results than existing state-of-the-art personal Re-ID systems, by achieving Rank-1 rates of 99.91%, 98.92%, and 94.79%, and mAP of 99.67%, 98.34%, and 94.15%, on the Market-1501, CUHK03 and P-DESTRE datasets, respectively. Based on the experimental results, it is obvious that the proposed system has illustrated its ability to segment and identify individuals using various fusion approaches at the score and rank levels.

The focus of future research will be on evaluating the effectiveness of the proposed ReID-DeePNet system using more difficult ReID datasets. We are also working on expanding the current background suppression module by combining person masks and key points to match body parts accurately, eliminate undesired information (e.g., background clutter) and achieve higher accuracy. Another important factor to investigate is the size of deep learning models, since large trained models require more storage space, which makes them difficult to store on small embedded devices. Therefore, models with fewer hyper-parameters and equal or better matching accuracy should be considered.

Author Contributions: Conceptualization, H.J.M., S.A.-F., A.S.A.-W., D.A.Z., D.A.I., M.A.M., S.K. and J.K.; methodology, H.J.M., S.A.-F., A.S.A.-W., D.A.Z., D.A.I., M.A.M., S.K. and J.K.; software, H.J.M. and A.S.A.-W.; formal analysis, H.J.M., S.A.-F., A.S.A.-W., D.A.Z., D.A.I., M.A.M., S.K. and J.K.; investigation, A.S.A.-W.; resources, M.A.M. and S.K.; data curation, H.J.M., A.S.A.-W. and D.A.I.; writing—original draft preparation, H.J.M., S.A.-F., A.S.A.-W., D.A.Z., D.A.I., M.A.M., S.K. and J.K.; writing—review and editing, H.J.M., S.A.-F., A.S.A.-W., D.A.Z., D.A.I., M.A.M., S.K. and J.K.; project administration, A.S.A.-W. and M.A.M.; All authors have read and agreed to the published version of the manuscript.

Funding: This research was partly supported by the Technology Development Program of MSS [No. S3033853] and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1A4A1031509).

Data Availability Statement: Not applicable.

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

References

1. Wu, D.; Zheng, S.J.; Zhang, X.P.; Yuan, C.A.; Cheng, F.; Zhao, Y.; Lin, Y.J.; Zhao, Z.Q.; Jiang, Y.L.; Huang, D.S. Deep learning-based methods for person re-identification: A comprehensive review. *Neurocomputing* **2019**, *337*, 354–371. [CrossRef]
2. Wu, D.; Zheng, S.J.; Bao, W.Z.; Zhang, X.P.; Yuan, C.A.; Huang, D.S. A novel deep model with multi-loss and efficient training for person re-identification. *Neurocomputing* **2019**, *324*, 69–75. [CrossRef]
3. Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; Hoi, S.C.H. Deep Learning for Person Re-identification: A Survey and Outlook. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 2872–2893. [CrossRef] [PubMed]
4. Marcel, S.; Nixon, M.S.; Li, S.Z. *Handbook of Biometric Anti-Spoofing-Trusted Biometrics under Spoofing Attacks*; Advances in Computer Vision and Pattern Recognition; Springer: Berlin/Heidelberg, Germany, 2014; pp. 1–279. ISBN 978-1-4471-6523-1. [CrossRef]
5. Tistarelli, M.; Li, S.Z.; Rama, C. *Handbook of Remote Biometrics- for Surveillance and Security*; Springer: London, UK, 2009; pp. 1–281. ISBN 9781849965064. [CrossRef]
6. Yao, Z.; Wu, X.; Xiong, Z.; Ma, Y. A dynamic part-attention model for person re-identification. *Sensors* **2019**, *19*, 2080. [CrossRef] [PubMed]
7. Khatun, A.; Denman, S.; Sridharan, S.; Fookes, C. Pose-driven Attention-guided Image Generation for Person Re-Identification. *arXiv* **2021**, arXiv:2104.13773.
8. Geng, S.; Yu, M.; Yu, Y.; Guo, Y. A segmentation-based human alignment network for person re-identification with frequency weighting re-ranking. *Acad. J. Sci. Res.* **2019**. [CrossRef]
9. Baabou, S.; Bremond, F.; Fradj, A.; Farah, M.A.; Kachouri, A.; Baabou, S.; Bremond, F.; Fradj, A.B.; Farah, M.A.; Kachouri, A. Hand-Crafted System for Person Re-Identification:A Comprehensive Review. In Proceedings of the International Conference on Smart, Monitored and Controlled Cities (SM2C), Sfax, Tunisia, 17–19 February 2017.
10. Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; Tian, Q. Scalable Person Re-identification: A Benchmark University of Texas at San Antonio. In Proceedings of the IEEE International Conference on Computer Vision, Washington, DC, USA, 7–13 December 2015; pp. 1116–1124.
11. Kläser, A.; Marszałek, M.; Schmid, C. A spatio-temporal descriptor based on 3D-gradients. In Proceedings of the British Machine Vision Conference, Leeds, UK, 1–4 September 2008. [CrossRef]
12. Li, D.; Chen, X.; Zhang, Z.; Huang, K. Learning deep context-Aware features over body and latent parts for person re-identification. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition CVPR 2017, Honolulu, HI, USA, 21–26 July 2016; pp. 7398–7407. [CrossRef]
13. Ming, Z.; Zhu, M.; Wang, X.; Zhu, J.; Cheng, J.; Gao, C.; Yang, Y.; Wei, X. Deep learning-based person re-identification methods: A survey and outlook of recent works. *Image Vis. Comput.* **2022**, *119*, 104394. [CrossRef]

14. Lavi, B.; Ullah, I.; Fatan, M.; Rocha, A. Survey on Reliable Deep Learning-Based Person Re-Identification Models: Are We There Yet? *arXiv* **2020**, arXiv:2005.00355.
15. Liao, S.; Hu, Y.; Zhu, X.; Li, S.Z. Person Re-identification by Local Maximal Occurrence Representation and Metric Learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 2197–2206.
16. Li, Z.; Chang, S.; Liang, F.; Huang, T.S.; Cao, L.; Smith, J.R. Learning locally-adaptive decision functions for person verification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, 23–28 June 2013; pp. 3610–3617. [CrossRef]
17. Zhong, W.; Jiang, L.; Zhang, T.; Ji, J.; Xiong, H. Combining multilevel feature extraction and multi-loss learning for person re-identification. *Neurocomputing* **2019**, *334*, 68–78. [CrossRef]
18. Lin, Y.; Zheng, Z.; Zhang, H.; Gao, C.; Yang, Y. Bayesian query expansion for multi-camera person re-identification. *Pattern Recognit. Lett.* **2020**, *130*, 284–292. [CrossRef]
19. Yan, Y.; Ni, B.; Liu, J.; Yang, X. Multi-level attention model for person re-identification. *Pattern Recognit. Lett.* **2019**, *127*, 156–164. [CrossRef]
20. Yang, B.; Shan, Y.; Peng, R.; Li, J.; Chen, S.; Li, L. A feature extraction method for person re-identification based on a two-branch CNN. *Multimed. Tools Appl.* **2022**, *14*, 1–16. [CrossRef]
21. Xin, X.; Wang, J.; Xie, R.; Zhou, S.; Huang, W.; Zheng, N. *Semi-Supervised Person Re-Identification Using Multi-View Clustering*; Elsevier Ltd.: Amsterdam, The Netherlands, 2019; Volume 88, pp. 285–297. ISBN 8602983395146. [CrossRef]
22. Isobe, T.; Li, D.; Tian, L.; Chen, W.; Shan, Y.; Wang, S. Towards Discriminative Representation Learning for Unsupervised Person Re-identification. In Proceedings of the IEEE International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 8506–8516. [CrossRef]
23. Xia, L.; Zhu, J.; Yu, Z. Real-World Person Re-Identification via Super-Resolution and Semi-Supervised Methods. *IEEE Access* **2021**, *9*, 35834–35845. [CrossRef]
24. Wu, G.; Zhu, X.; Gong, S. Learning hybrid ranking representation for person re-identification. *Pattern Recognit.* **2022**, *121*, 108239. [CrossRef]
25. Tang, Z.; Huang, J. Harmonious Multi-branch Network for Person Re-identification with Harder Triplet Loss. *ACM Trans. Multimed. Comput. Commun. Appl.* **2022**, *18*, 98. [CrossRef]
26. Gu, X.; Chang, H.; Ma, B.; Bai, S.; Shan, S.; Chen, X. Clothes-Changing Person Re-identification with RGB Modality Only. *arXiv* **2022**, arXiv:2204.06890.
27. Yang, M.; Huang, Z.; Hu, P.; Li, T.; Lv, J.; Peng, X. Learning with Twin Noisy Labels for Visible-Infrared Person Re-Identification. *Cvpr* **2022**, *1*, 14308–14317.
28. Song, C.; Huang, Y.; Ouyang, W.; Wang, L. Mask-Guided Contrastive Attention Model for Person Re-identification. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1179–1188. [CrossRef]
29. Yang, X.; Tang, Y.; Wang, N.; Song, B.; Gao, X. An End-to-End Noise-Weakened Person Re-Identification and Tracking with Adaptive Partial Information. *IEEE Access* **2019**, *7*, 20984–20995. [CrossRef]
30. Bello, R.W.; Mohamed, A.S.A.; Talib, A.Z. Contour Extraction of Individual Cattle from an Image Using Enhanced Mask R-CNN Instance Segmentation Method. *IEEE Access* **2021**, *9*, 56984–57000. [CrossRef]
31. Abed, S.H.; Al-Waisy, A.S.; Mohammed, H.J.; Al-Fahdawi, S. A modern deep learning framework in robot vision for automated bean leaves diseases detection. *Int. J. Intell. Robot. Appl.* **2021**, *5*, 235–251. [CrossRef]
32. Hegadi, R.S. Improved GrabCut Technique for Segmentation of Color Images. *Int. J. Comput. Appl.* **2014**, *975*, 975–8887.
33. Al-Waisy, A.S.; Qahwaji, R.; Ipson, S.; Al-Fahdawi, S.; Nagem, T.A.M. A multi-biometric iris recognition system based on a deep learning approach. *Pattern Anal. Appl.* **2018**, *21*, 783–802. [CrossRef]
34. Wang, Z.; Zeng, Y.; Liu, Y.; Li, D. Deep Belief Network Integrating Improved Kernel-Based Extreme Learning Machine for Network Intrusion Detection. *IEEE Access* **2021**, *9*, 16062–16091. [CrossRef]
35. Hinton, G.E.; Salakhutdinov, R.R. Reducing the Dimensionality of Data with Neural Networks. *Science* **2006**, *313*, 504–507. [CrossRef] [PubMed]
36. Al-Waisy, A.S.; Qahwaji, R.; Ipson, S.; Al-Fahdawi, S. A multimodal deep learning framework using local feature representations for face recognition. *Mach. Vis. Appl.* **2017**, *29*, 35–54. [CrossRef]
37. Li, W.; Zhao, R.; Xiao, T.; Wang, X.; Li, W.; Zhao, R.; Xiao, T.; Wang, X. DeepReID: Deep Filter Pairing Neural Network for Person Re-identification. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 152–159. [CrossRef]
38. Kumar, S.V.A.; Yaghoubi, E.; Das, A.; Harish, B.S.; Proenca, H. The P-DESTRE: A Fully Annotated Dataset for Pedestrian Detection, Tracking, and Short/Long-Term Re-Identification from Aerial Devices. *IEEE Trans. Inf. Forensics Secur.* **2021**, *16*, 1696–1708. [CrossRef]
39. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In *Computer Vision—ECCV 2016*; Springer: Cham, Switzerland, 2016; Volume 9905, pp. 21–37. [CrossRef]
40. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.

41. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
42. Huang, H.; Yang, W.; Chen, X.; Zhao, X.; Huang, K.; Lin, J.; Huang, G.; Du, D. EANet: Enhancing Alignment for Cross-Domain Person Re-identification. *arXiv* **2018**, arXiv:1812.11369.
43. Zhao, L.; Li, X.; Zhuang, Y.; Wang, J. Deeply-Learned Part-Aligned Representations for Person Re-identification. In Proceedings of the 2017 IEEE International Conference on Computer Vision ICCV, Venice, Italy, 22–29 October 2017; pp. 3239–3248. [CrossRef]
44. Fu, Y.; Wei, Y.; Zhou, Y.; Shi, H.; Huang, G.; Wang, X.; Yao, Z.; Huang, T. Horizontal pyramid matching for person re-identification. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 8295–8302. [CrossRef]
45. Khatun, A.; Denman, S.; Sridharan, S.; Fookes, C. End-To-End Domain Adaptive Attention Network for Cross-Domain Person Re-Identification. *IEEE Trans. Inf. Forensics Secur.* **2021**, *16*, 3803–3813. [CrossRef]
46. Zhang, Z.; Lan, C.; Zeng, W.; Chen, Z. Densely semantically aligned person re-identification. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 667–676. [CrossRef]
47. Zhao, Y.; Zhong, Z.; Yang, F.; Luo, Z.; Lin, Y.; Li, S.; Sebe, N. Learning to Generalize Unseen Domains via Memory-based Multi-Source Meta-Learning for Person Re-Identification. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 6273–6282. [CrossRef]
48. Wang, H.; Shen, J.; Liu, Y.; Gao, Y.; Gavves, E. NFormer: Robust Person Re-identification with Neighbor Transformer. *arXiv* **2022**, arXiv:2204.09331.
49. Moritz, L.; Specker, A.; Schumann, A. A study of person re-identification design characteristics for aerial data. *Pattern Recognit. Track. XXXII* **2021**, *11735*, 161–175. [CrossRef]
50. Specker, A.; Moritz, L.; Sommer, L.W. Deep learning-based video analysis pipeline for person detection and re-identification in aerial imagery. *Count. Crime Fight. Forensics Surveill. Technol.* **2021**, *11869*, 86–95. [CrossRef]