Region-based convolutional neural networks for occluded person re-identification

Atiqul Islam [a,1,*], Mark Tee Kit Tsun [b,2], Lau Bee Theng [b,3], Caslon Chua [b,4]

[a] Faculty of Engineering, Computing & Science, Swinburne University of Technology Sarawak Campus Kuching 93350, Malaysia
[b] Department of Computer Science and Software Engineering, Swinburne University of Technology, Melbourne 3122, Australia
[1] aislam@swinburne.edu.my; [2] mtktsun@swinburne.edu.my; [3] blau@swinburne.edu.my; [4] cchua@swin.edu.au
* corresponding author

ARTICLE INFO

Article history
Received May 20, 2023
Revised July 1, 2023
Accepted July 18, 2024
Available online February 29, 2024

Keywords
Occlusion
R-CNN
Re-identification
Region re-ranking

ABSTRACT

In a variety of applications, including intelligent surveillance systems, targeted tracking, and assistive human-following robots, the ability to accurately identify individuals even when they are partially obscured is imperative. Such Continuous person tracking is complicated by the close similarity between the appearance of people and target occlusions. This study addresses this significant challenge by proposing a two-step, detection-first approach that uses a region-based convolutional neural network (R-CNN) as the re-identification (re-ID) solution. The model is specifically trained to detect occluded persons at different levels of occlusion before forwarding the image for the re-ID process. Three occluded-specific datasets are selected to evaluate the model's effectiveness in detecting occluded people. There are 379 distinct people in total, and each has five images obstructed from different angles. A sample of the data is taken to simulate various environment settings, and new data points are generated with different degrees of occlusion to assess how well the model performs under varying levels of obstruction. The findings demonstrate that the proposed person re-ID model is reliable in most circumstances, correctly re-identifying at 74% (Rank-1) and 90% (Rank-5). Although there is a decrease in accuracy as the number of distinctive people in the dataset increases, this does not significantly impact the tracking performance in various applications, which are expected to recognize a single person or a small group of individuals. Future works will explore refining similarity matching algorithms by delving into robust image comparison techniques, thereby addressing the challenges presented by occlusions. A critical aspect is to assess the model under diverse lighting conditions and investigate scenarios with multiple individuals in a frame. It is also beneficial to exploit high-resolution datasets, such as DukeMTMC-reID, and integrate finer contextual details, like clothing or carried objects. These collective efforts are essential for optimizing the model's efficacy in practical applications and advancing person re-ID technologies.

1. Introduction

Person Re-Identification (re-ID) is a branch of computer vision study that focuses on successively recognizing people from an existing collection of photos (gallery set) taken by one or multiple cameras [1]. It is the core of many application areas, such as intelligent surveillance systems, targeted tracking, assistive human following robots, etc. Various research efforts have been carried out over the years to improve re-ID in complex scenarios, but it remains one of the most challenging and long-standing research topics in computer vision [2].

Fig. 1. Example of various categories of occluded person images

A robust re-ID system must operate through different viewpoints, low-image resolutions, illumination changes, occlusions, background clutter, etc. [1]. Any traditional deep learning person detection algorithm, such as CNN-based algorithms, can detect a person with an elevated level of accuracy when the person is not occluded and visible consistently. But the introduction of occlusion and background clutter is especially challenging and usually results in failed re-identification because the target person is not detected in the first place [3]–[5]. The failure of detection is because the contemporary vision-based recognition of humans is not consistently successful at generating a bounding box (bbox) when only parts of the target person are visible [5], [6]. It is challenging for the detection techniques to properly define the bounding box of a person when the person is partially or fully occluded by other persons or objects, such as cars, umbrellas, bags etc., in a scene.

Person re-identification (re-ID) consists of person detection, person tracking, and similarity matching [3]–[11]. Detection is a computer vision task identifying and locating humans in images or videos [2]–[5], [7], [12]–[18]. However, most re-ID research focuses on similarity matching or person retrieval [11].

In recent literature, several approaches have been developed to tackle occlusion in person re-identification using deep learning models. One approach, known as pose-guided feature alignment [4], [19]–[21], employs human landmarks to generate attention maps, which guide the model to concentrate on non-occluded regions. Although effective, its performance is limited because it assumes that gallery images are not occluded, which causes a loss in accuracy. Another approach focuses on the reconstruction of occluded images through Generative Adversarial Networks (GANs) [22]–[24], aiming to generate the occluded portion of the human body, and additionally enhance input images by increasing resolution. However, this method is reported to introduce noise; researchers have tried adding additional channels to adjust the noise. The complexity and resource intensiveness of this approach remains a challenge. A more recent development is the partial re-ID approach [25]–[30], which detects and identifies individual body parts. By using spatial location information of visible landmarks, this approach filters noise and handles occluded images more effectively but at the expense of multi-network model architecture. Researchers have incorporated additional context (carry bag, personal items) into this model architecture, which makes this approach the most complex network architecture. Overall, these approaches have made significant strides in occlusion-specific re-ID models but have limitations that leave room for further research and development.

Building upon the existing literature, it is evident that while significant strides have been made in occlusion-specific re-ID models, there are inherent limitations that warrant further exploration and innovation. A key observation is the scarcity of research addressing the re-identification of individuals from various angles of occlusion, especially when only parts of the target are visible [8]. Considering this,

the present study undertakes an exploration into the efficacy of cutting-edge region-based image detection algorithms, with a focus on detecting and re-identifying occluded individuals across various occlusion directions, such as top, bottom, and sideways (as illustrated in Fig. 1). This research introduces a region-based Convolutional Neural Network (CNN) model that is adept at detecting and re-identifying occluded persons in diverse settings. The model employs a Feature Pyramid Network (FPN) to generate multi-scale feature maps, a Region Proposal Network (RPN) to identify prospective regions containing individuals, by manipulating the intersection over union (IoU) and training RPN on segmented images, occluded persons are detected without needing an additional network, effectively simplifying the network architecture, and a Box Head technique for refining these regions and their classifications. Additionally, a Re-ID stack is incorporated, which utilizes an extended version of the Detectron2 framework for feature matching against a gallery of images. In the feature matching phase, Euclidean distance is used to compute the similarity between the feature vectors of the probe image and the gallery images, whereas Jaccard distance is employed to quantify the dissimilarity between sets, thereby enabling the model to effectively re-identify occluded individuals by comparing features in a multi-dimensional space.

The scope of this research is deliberately confined to settings involving groups of 2-3 individuals, owing to the availability of relevant occlusion-oriented datasets for training and testing [3], [4], [31]. The datasets utilized in this study comprise CVC05-Part occlusion [16], Pascal VOC2007 [32], Occluded REID [33], Partial_REID & PartialiLIDS [31], jhu-crowd, Crowd-Human, and Market1501 [30]. It is pertinent to note that except for jhu-crowd and CrowdHuman, the datasets primarily feature single individuals or groups not exceeding three members, which is the focus of this study. Jhu-crowd and CrowdHuman, in contrast, contain images of larger crowds, such as audiences at football games or concerts, and are not the central focus of this research.

The significance and potential impact of this research include but are not limited to (a) the creation of a robust region-based CNN model for re-identifying occluded individuals, enhancing the reliability of intelligent surveillance systems, (b) supporting targeted tracking in various conditions, crucial for continuous monitoring of individuals even when partially hidden, and (c) providing an advanced re-ID algorithm to improve assistive human-following robots in indoor environments where occlusion by objects or other individuals is common.

2. Method

The proposed occluded person re-id module consists of four parts: the backbone network, also known as the feature pyramid network (FPN), a region proposal network (RPN), a box head module (mask generator) and lastly, the re-ID stack. The model architecture for FPN, RPN, and box head is chosen because the current state-of-the-art image detection framework detectron2 [cite] utilizes this approach. It is observed in the literature that the state-of-the-art in occluded re-ID is moving towards partial re-ID, and recent development included contextual information such as bags, hats, umbrellas etc., related to a person for better feature embedding [cite]. As described in the previous section, the existing methods use multiple networks to achieve this. The proposed method uses a single pipeline to create the multiscale feature map with FPN and use the generated feature set to identify the possible region of the occluded person (RPN), generate a segmentation mask (contextual understanding), and calculate the distance between a searched image and a gallery image for re-identification. The complete architecture of the proposed model can be viewed in Fig. 2.

2.1. Feature Pyramid Network

Recognizing people in surveillance systems, where there is often a distance between people and CCTV cameras, can be challenging, especially when detecting subjects of various sizes. The feature pyramid network (FPN) is a technique that helps address the challenge of detecting objects at different scales in surveillance systems. FPN enhances object recognition by creating a pyramid of multi-scale feature maps that capture information at various levels of detail. This enables more accurate detection of objects, including both large and small ones [34].

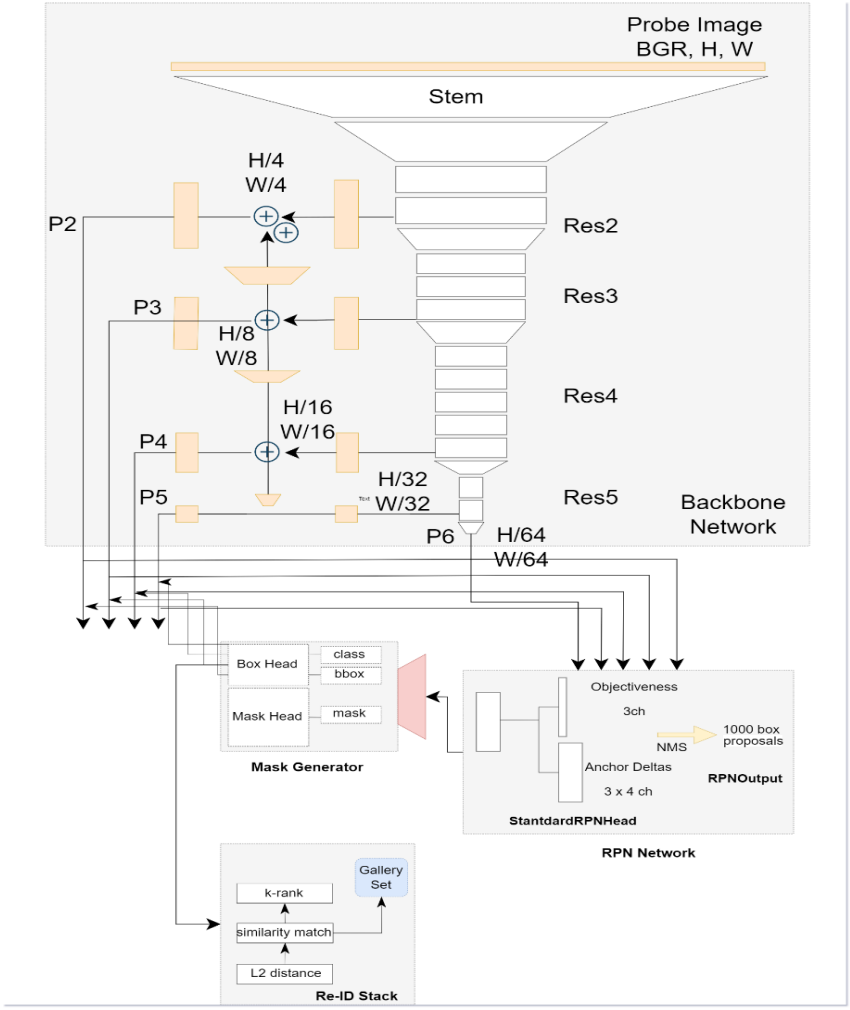

Fig. 2. The architecture of the proposed model consists of four parts: the backbone network (FPN), the region proposal network (RPN), the mask generator, and re-ID stack

The backbone network used in the FPN is based on the ResNet-101 architecture. The stride used in the initial convolutional layer is 2. This results in downscaling the input resolution by a factor of 2, as is explained later as the relationship between res2, res3, res4, and res5. The ResNet-101 provides the necessary depth and capacity to capture complex features from the input images.

The feature map generated at res2, res3, res4, and res5 (Fig. 2) refers to the output feature maps at distinct levels of the backbone network. res2 represents the highest resolution and lowest depth, while res3 is obtained by downsampling res2 by a factor of 2 along both the height and width dimensions, res4 is derived by downsampling res3 by a factor of 2, and res5 is obtained by downsampling res4 by a factor of 2, representing the lowest resolution and highest depth among them. This hierarchical relationship between res3, res4, and res5 ensures that the feature maps at each level capture progressively larger receptive fields while retaining spatial information from the previous levels.

These feature maps are then used as input to the top-down pathway in the FPN architecture. The top-down pathway combines the feature maps from the backbone network with up-sampled feature maps from coarser levels to generate multi-scale feature maps (P2, P3, P4, and P5), as shown in Fig. 2. The multi-scale feature maps P2, P3, P4, and P5 from the feature maps Res2, Res3, Res4, and Res5 in the FPN architecture can be expressed as follows: P2 = Res2, P3 = Res3 + Upsample(P4, scale_factor=2), P4 = Res4 + Upsample(P5, scale_factor=2), and P5 = Res5

For the purpose of person re-identification with occlusion, a fine-tuned feature extraction and similarity measurement module is applied to the FPN pipeline for re-ID. For example, a weighted feature

combination merges multiple features or representations, considering their relative importance or contribution. In the context of multi-scale features from an FPN, the weighted feature combination combines the features obtained from diverse levels (P2, P3, P4, and P5) using a set of weights (w2, w3, w4, w5). Given the FPN output feature maps P2, P3, P4, and P5, and a set of weights w = {w2, w3, w4, w5}, the combined feature F is computed as:

$$F = (w2 * P2 + w3 * P3 + w4 * P4 + w5 * P5) / (w2 + w3 + w4 + w5) \quad (1)$$

The multi-scale features and weighted features obtained from the FPN are used to learn discriminative and robust person representations invariant to changes in pose, illumination, and camera viewpoint.

2.2. Region Proposal Network

After the backbone network generates feature maps, these feature maps are utilized by two critical components: the Region Proposal Network (RPN) and the Region of Interest Pooling (RoIP). The RPN takes the multi-scale feature maps F from the backbone network as input and uses a sliding window approach to generate region proposals, and potential bounding boxes containing objects of interest. For each position (i, j) in the sliding window, the RPN generates k anchor boxes $A = \{a1, a2, \ldots, ak\}$ with different aspect ratios and scales, which are scored based on their likelihood of containing an object (objectness score S) and refined to generate the final region proposals. Objectness score S is computed as:

$$S = Signoid\ (W_s * F[i, j]) \quad (2)$$

The bounding box refinement is calculated as

$$AB = W_b * F[i, j] \quad (3)$$

where W_s and W_b are learned weights for objectness and bounding box refinement, respectively.

To adapt the Detectron2 framework to focus on detecting people, the RPN is modified to concentrate on person-specific anchors during training. By configuring the dataset to include only person annotations as ground truth labels and filtering out other object classes, the RPN is trained to generate region proposals primarily for people.

After obtaining the region proposals from the RPN, extracting features corresponding to each proposal using the Region of Interest Pooling (RoIP) technique is next. RoIP converts the features inside each region proposal into a fixed-size feature map M by dividing each region proposal into an H x W grid of equally sized subregions and pooling the features within each subregion using max-pooling or average pooling. For each subregion (h, w) in the RoIP grid, M $[h, w]$ is computed as:

$$max\_pool(F\ [r\_top: h\_top, r\_left: w\_right]) \quad (4)$$

$$avg\_pool(F\ [r\_top: h\_top, r\_left: w\_right]) \quad (5)$$

where r_top, h_top, r_left, and w_right are the coordinates of the subregion in feature map F.

The Jaccard Index is a measure of similarity between two sets, and it is calculated as the ratio of the intersection of the sets to the union of the sets. When applied to object detection tasks, the Jaccard Index is called IoU, which measures the overlap between the predicted bounding box and the ground truth bounding box. A higher IoU threshold requires a greater overlap between the predicted bounding box and the ground truth bounding box to be considered a true positive detection. Conversely, a lower IoU threshold would allow for more leniency in the overlapping requirement, potentially leading to more false positives. Therefore, the partially occluded person is more likely to be detected by relaxing the IoU threshold. Based on the region of interest proposal (RoIP) from the RPN network, the ground truth preparation module is responsible for assigning, generating, and matching the bounding boxes. For this

research, the IoU threshold for the RPN network is set to .5, which results in a higher number of partially visible bodies to be detected as a positive result, essentially achieving what the partial re-ID techniques achieve with training on concrete occluded person data. Fig. 3 shows an example of IoU segmentations.

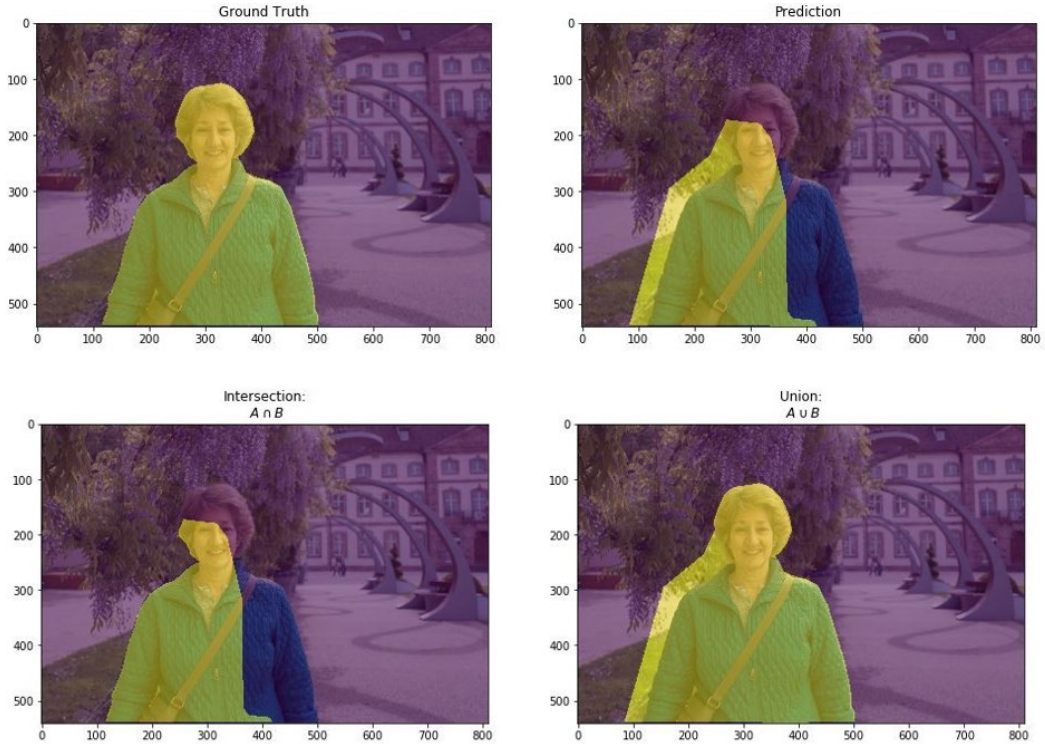

Fig. 3. Example of IoU during image segmentation

The Jaccard distance (D) is calculated as:

$$D(A,B) = 1 - J(A,B) \tag{6}$$

2.3. Re-ID Stack

In the final development phase of the proposed person re-identification (re-ID) system, the previous two phases of feature maps and region proposals are integrated. The similarity calculation in the re-ID task requires a gallery set (an array of pre-existing images) and a probe set (images containing the person of interest), where the probe image may or may not be present in the gallery set. Given the gallery set G and probe set P, the objective is to generate an array of ranked matches. The expected input for the similarity calculation includes; Probe_feature: All feature vectors of the query set, with shape $(image\_size, feature\_dim)$; Gallery_feature: All feature vectors of the gallery set, with shape $(image\_size, feature\_dim)$; $k1, k2, lambda$: Parameters that, in the original paper by Zhong et al. [35], are set to (k1=20, k2=6, lambda=0.3).

Given feature vectors of probe and gallery images, the algorithm refines the similarity distances between them. It starts by computing the original pairwise distances. Then, it performs re-ranking by considering reciprocal nearest neighbours and expanding the neighbours iteratively. Next, the Jaccard distance is calculated based on the weighted similarities. Finally, the Jaccard distance is combined with the original distances using a weight factor to obtain the final refined similarity distances. The re_ranking function, described in the pseudocode, implements this process and returns the final distance matrix.

The Fig. 4 shows the finalized re-ranking algorithm that uses the generated feature maps in the previous section.

Algorithm: producing a re-ranked list using the feature maps generated in the previous section

Input: Probe feature vectors, Gallery feature vectors

Output: Refined similarity distances

1. Take the feature vectors of a probe and gallery images as inputs.
2. Compute the pairwise Euclidean distances between all pairs of feature vectors:
   Calculate distances: $distances = pairwise\_distances(probe\_vectors, gallery\_vectors)$
3. Square the distances to obtain the squared Euclidean distances:
   Square distances: $squared\_distances = distances ** 2$
4. Normalize the distances by dividing each distance value by the maximum distance value in the column:
   Normalize distances: $normalized\_distances = squared\_distances / np.max(squared\_distances, axis = 0)$
5. Initialize an empty matrix V to store weighted similarities.
6. For each feature vector:
   Sort the distances in ascending order and find the k1 nearest neighbors:
   Sort distances: $sorted\_indices = np.argsort(normalized\_distances, axis = 1)$
   Get nearest neighbors: $nearest\_neighbors = sorted\_indices[:, :k1]$
   Iterate through the k1 nearest neighbors of the current feature vector:
   Expand the neighbor set by considering the k1 nearest neighbors of each neighbor:
   Expand neighbors: $expanded\_neighbors = np.unique(np.concatenate (nearest_{neighbors[neighbor]}] \; for \; neighbor \; in \; nearest\_neighbors))$
   Calculate the weights for the expanded neighbors based on the exponential of the negative squared distances:
   Calculate weights: weights = $np.exp(-normalized\_distances[current\_feature\_vector, expanded\_neighbors])$
   Normalize the weights by dividing each weight value by the sum of weights:
   Normalize weights: $normalized\_weights = weights / np.sum(weights)$
7. Compute the Jaccard distance between the feature vectors using the weighted similarities:
   Compute Jaccard distances: $jaccard\_distances = 1 - temp\_min / (2 - temp\_min)$
8. Combine the Jaccard distance and the original distances using a weighted sum, where the weight factor is $lambda\_value$:
   Combine distances: $final\_distances = jaccard\_distances * (1 - lambda\_value) + normalized\_distances * lambda\_value$

Return the final refined similarity distances, excluding the distances within the probe set

Fig. 4. Algorithm for re-ranked list of re-ID feature map set

2.4. Implimentation Details

Person re-identification is a multidimensional area of study, and researchers have examined gait, key points, and posture estimates for improved re-identification in various scenarios throughout the years. Detectron2 image detection framework provides an uncomplicated way to implement different head modules for several types of detection algorithms. Detectron2 has pre-trained weights available for download that are either trained on MS COCO or ImageNet datasets. Both datasets are annotated for object segmentation. For the purpose of this research, the backbone network is trained on ImageNet for feature extraction. The pre-trained model for the RPN network is trained on a subset of the MS COCO dataset and Open Images Dataset with only annotated person images. The person images ranged from occluded person to person with additional artifacts. The subset dataset has 70k images of people in different scenarios. As the RPN network tries to generate the segmentation mask for each person detected, the intersection over union (IoU) threshold is set at 50% to detect possible occluded bodies.

- Training Process : The initial weight is initialized from COCO weights available on the Detectron2 GitHub repository for faster training. The number of the target class is set to “Person,” as that is our primary focus in a scene. The learning rate is fixed at 0.001, the number of workers set as two and the batch size as 16—the network trained for a total of 2.25 × 105 iterations.

2.5. Occluded re-ID dataset and detection testing

To evaluate the proposed two-step model for occluded person re-identification, three occluded specific datasets are chosen: OccludedREID, Partial_REID and PartialiLIDS datasets. The number of unique individual images from each dataset is maintained at 100 images for indoor testing. The total 300 images contain 60 unique person images. Each unique person has 5 occluded images from top, bottom, and either side. Fig 1 shows sample images from all three of the datasets. The detection performance of the proposed occluded re-ID model is shown in Table 1.

Table 1. The detection performance of the proposed model

Dataset	Average Precision (AP)	AP at IoU 0.5 (AP50)
OccludedREID	76.36	89.02
Partial_REID	78.00	93.10
PartialiLIDS	69.70	88.90
Average AP	74.69	90.34

The average precision (AP) across all three datasets is 74.69, and the AP with 0.5 IoU achieved a result of 90.34 precision. This demonstrates the effectiveness of the proposed model in identifying occluded persons regardless of occlusion type.

3. Results and Discussion

The deep learning model was created, executed, trained, and evaluated on a desktop PC with Nvidia RTX 2080 Super GPU and Pytorch, and Detectron2 python library. The performance of the re-identification is evaluated on OccludedREID, Partial_REID, and PartialiLIDS datasets. Rank1, Rank3, and Rank5 metrics are used for the evaluation. Rank refers to the position or order of an item within a ranked list or set of items. In evaluation metrics, ranks signify the position of a matching candidate in a ranked list, reflecting the accuracy or similarity between compared entities. For this research, Rank1 refers to the top-ranked or exact match. Rank3 and Rank5 indicate a correct match within the top 3 or top 5 results, respectively.

The proposed Occluded Person re-ID model demonstrated strong performance in detecting occluded individuals across occlusion-specific datasets. However, when compared to other state-of-the-art re-ID models that focused on a single dataset, its re-identification performance was relatively lower. The drop in performance can be explained by the use of three different occlusion-specific datasets for testing instead of fine-tuning the model for only one specific dataset. The detection module achieved an impressive 90% accuracy in correctly identifying occluded images, while the re-identification module achieved a Rank-1 accuracy of 74% and a Rank-5 accuracy of 90%. The advantage of the proposed model is that the bottleneck is easily identifiable. It is evident that the bottleneck lies in the re-identification module, where even with correct detection, the similarity matching fails to recognize the person accurately. Therefore, in occluded re-ID, occlusion of a person is not the primary cause of a misrecognition, unless the obstruction is over 80%.

3.1. Re-ID Performance

The re-ID module of the proposed model is designed to work after a successfully detected person image is forwarded in the person re-ID pipeline. This module receives the region-of-interest-pooled feature map of the query person image and has access to the feature map set of the gallery database.

At this stage, the query feature map is compared with each image in the gallery set and the result is presented as a similarity score. This similarity score is calculated using the Jaccard index. Lastly, the similarity score is ranked from the closest match to the least close match. Table 2 shows an overview of the re-id performance of the three datasets.

Table 2. Occluded person dataset configuration

Dataset	Total Images	Unique Person	Resolution	Rank-1	Rank-5
OccludedREID	1000	200	128x64	62.78	81.33
Partial_REID	600	60	58x165	59.66	70.00
PartialiLIDS	238	119	128x420, 128x256	55.34	

- **OccludedREID dataset:** The evaluation results of the proposed re-ID model and other approaches with the Occluded REID dataset are shown in Table 3.

Table 3. Performance of occluded person re-ID model with 1000 samples

Approaches	Rank-1	Rank-5
XQDA [36]	36.61	65.11
KCVDCA [37]	32.48	59.10
GOG [38]	40.50	63.16
Null Space [39]	46.47	75.36
SVDNet [40]	63.13	85.13
REDA [41]	65.79	87.88
ResNet+AFPB [42]	68.14	88.29
Occluded Person re-ID (2022)	62.78	81.33

- **Partial_REID dataset:** The result is shown in Table 4. The table shows Rank 1 = exact match, and Rank 3 = top 3 results. Rank 5 is not included for Partial_REID because the existing comparison in the literature does include it.

Table 4. Performance of Occluded Person re-ID model

Approaches	Rank-1	Rank-3
MTRC [43]	23.70	27.30
AMC+SWM [44]	37.30	46.00
DSF [45]	50.70	70.00
SFR [46]	56.90	78.50
VPM [47]	67.70	81.90
SCPNet [44]	68.30	n/a
FastReID [48]	82.70	n/a
Occluded Person re-ID (2022)	59.66	70.00

- **ParlialiLIDS dataset:** PartialiLIDS is a collection of simulated partial people based on the Imagery Library for Intelligent Detection Systems (iLIDS). Both the gallery and query images are taken from behind. This dataset's images are noisier than the other two datasets evaluated in this research.

The evaluation results of the Occluded Person re-ID model and other approaches are shown in Table 5. The evaluation resulted in a similar result to the previous two datasets.

Table 5. Performance of occluded person re-ID model on PartialiLIDS dataset showing rank 1 (exact match) and top 3 matches as rank 3

Approaches	Rank-1	Rank-3
MTRC [43]	17.65	26.05
AMC+SWM [44]	17.65	26.05
DSR [45]	58.82	67.23
SFR [46]	63.87	74.79
VPM [47]	65.50	74.80
FastReID [48]	73.1	n/a
Occluded Person re-ID (2022)	55.34	64.38

The evaluation resulted in a similar result to the previous two datasets. In addition to the previously identified issue of correctly detecting a person but incorrectly re-identifying, the graininess of the images also caused incorrect recognitions.

3.2. Discussion

The proposed Occluded Person re-ID model demonstrated effective detection performance across all three datasets. This research was the first to perform instance segmentation on occluded specific datasets, and the result of instance segmentation is consistent, showing occlusion itself is not the failing point for any re-ID pipeline. The re-identification performance is low compared to other state-of-the-art re-ID models. However, it must be noted that the other models only focused on a single dataset for training and testing purposes, whereas the proposed re-ID model has been applied to three different datasets. This is a well-documented issue in the re-ID research community that the same model applied to a different dataset loses performance immediately [49], [50], [35]. Therefore, performance fluctuations between datasets were expected. The advantage of the proposed two-step model is that the bottleneck for the complete process is identifiable. The detection result showed, on average, 90% of the occluded images being correctly detected. The re-identification module proved to be the bottleneck, where even if a person is correctly detected, the re-identification (similarity matching) fails to recognize the person correctly.

Three sets of new datasets were created from the OccludedREID dataset to test the observation, with the lower body cropped at 50%, 70%, and 80%. The dataset is created from the gallery set and tested on the original gallery set to ensure a consistent viewpoint. Fig. 5 illustrates the generated images at different cropped ratios. The newly developed controlled dataset is evaluated using the re-ID model, and the result is tabulated in Table 6.

The results show that with 50% occlusion scenarios, the model achieved 90.1% rank-1 accuracy, 70% occlusion dipped the accuracy below 50%, but the rank-5 accuracy was still at 74.7%. Only at the extreme occlusion (80% of the body not visible) the performance dropped to 6%. This experiment shows that occlusion at a moderate level (below 70%) does not significantly impact re-ID if the viewpoint/camera angle is the same. From these observations, it can be confirmed that a more robust image comparison or matching algorithm is necessary for better accuracy.

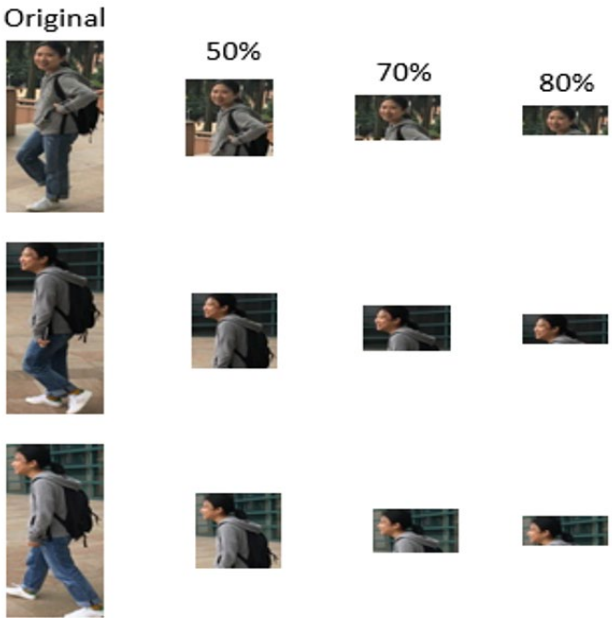

Fig. 5. Example of the generated dataset from OccludedREID dataset

The newly developed controlled dataset is evaluated using the re-ID model to confirm at what level of occlusion the model fails to identify the query correctly. The result is tabulated in Table 6.

Table 6. Performance of Re-ID model on generated OccludedREID dataset

Dataset	Rank-1	Rank-5
OccludedREID_50	90.1	97.1
OccludedREID_70	48.6	74.7
OccludedREID_80	6	16.3

Another observed pattern is that as the number of images increases, the model must compare more data points for a perfect match. The first dataset's model evaluation followed an incremental approach to test the assumption that an increase in gallery image drops the accuracy rate. The results are shown below in Table 7. It can be observed from the table that as the gallery size rises, both Rank-1 and Rank-5 steadily decline.

Table 7. Comparison of accuracy performance at 100, 500, and 1000 gallery images

Dataset	Number of gallery images	Number of unique people	Rank-1	Rank-5
OccludedREID	100	20	74.0	90.0
OccludedREID	500	50	66.25	86.33
OccludedREID	1000	100	62.78	81.33

Lastly, the effect of multiple people or groups in a scene has not been tested thoroughly. Evaluating the model's performance on higher-quality occluded person datasets, such as DukeMTMC-reID, would benefit future improvements.

In summary, A variety of factors influenced the model's performance. Notably, the size of the dataset and the extent of occlusion affected the re-identification accuracy. As the extent of occlusion exceeded 70%, the accuracy significantly dropped. Similarly, as the gallery size increased, both Rank-1 and Rank-5 accuracies declined. Further, dataset characteristics such as the graininess of images and the viewpoint of the camera also had an impact on the model's performance.

4. Conclusion

This research investigated region-based convolution neural networks for re-ID tasks. This has been successfully achieved by extending the current state-of-the-art image detection framework to integrate with the feature comparison algorithm and re-ranking, effectively proposing a two-step occluded person re-ID model. Unlike previous methods, this model streamlined complex architectures into a singular network by extending the detectron2 image detection framework to serve as a person detector while deriving additional context. The proposed model showed robust performance in detecting occluded individuals, with a detection accuracy of 90%, and re-identification Rank-1 and Rank-5 accuracies of 74% and 90%, respectively. However, the re-identification performance was lower when compared to other state-of-the-art re-ID models, due to the diverse range of occlusion-specific datasets employed for testing.

Acknowledging the study's limitations, several future research directions have been identified. These include improving similarity matching algorithms by investigating more robust image comparison or matching techniques that can help address the challenges posed by occlusion and other complex scenarios, thus improving the overall Re-ID performance. Assessing the model's performance under various lighting conditions and incorporating techniques to address these variations in the detection and re-ID processes will lead to consistent performance in various environments. Another critical area to explore is the group re-ID scenarios. Future research could explore how the model performs when multiple individuals impede each other in a scene and develop strategies to enhance its performance in such complex situations.

Furthermore, testing the model on high-quality occluded person re-ID datasets, such as DukeMTMC-reID, can provide valuable insights into its effectiveness in detecting and re-IDing occluded persons in high-quality images, guiding further refinements to the model to ensure optimal performance across various image resolutions. Additionally, incorporating the pixel-level understanding of a scene by integrating contextual information, such as identifying whether a person is wearing a hat or carrying a bag or luggage, can provide valuable context to improve the model's performance in occluded person re-ID tasks. In conclusion, by addressing these research directions, the proposed model can be further refined and optimized for practical applications in various contexts, ultimately contributing to the advancement of person re-ID technology.

Acknowledgment

Malaysian Ministry of Higher Education funded the work under the Fundamental Research Grant Scheme, FRGS/1/2020/ICT02/SWIN/03/01.

Declarations

Author contribution. All the listed authors have written the work based on their expertise
Funding statement. Malaysian Ministry of Higher Education funded the work under the Fundamental Research Grant Scheme, FRGS/1/2020/ICT02/SWIN/03/01, 2020-2023.
Conflict of interest. The authors declare no conflict of interest.
Additional information. No additional information is available for this paper.

References

[1] A. Islam, M. K. T. Tee, and B. T. Lau, "Development of an Improved Occluded Person Re-Identification System Using Deep Learning," in *2022 6th High Performance Computing and Cluster Technologies Conference (HPCCT)*, Jul. 2022, pp. 44–50, doi: 10.1145/3560442.3560449.

[2] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep Learning for Person Re-Identification: A Survey and Outlook," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 2872–2893, Jun. 2022, doi: 10.1109/TPAMI.2021.3054775.

[3] X.-T. Vo and K.-H. Jo, "Accurate Bounding Box Prediction for Single-Shot Object Detection," *IEEE Trans. Ind. Informatics*, vol. 18, no. 9, pp. 5961–5971, Sep. 2022, doi: 10.1109/TII.2021.3138336.

[4] J. Miao, Y. Wu, P. Liu, Y. Ding, and Y. Yang, "Pose-Guided Feature Alignment for Occluded Person Re-Identification," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019, vol. 2019-Octob, pp. 542–551, doi: 10.1109/ICCV.2019.00063.

[5] D. Wu *et al.*, "Random Occlusion Recovery for Person Re-identification," *J. Imaging Sci. Technol.*, vol. 63, no. 3, pp. 030405-1-030405–9, May 2019, doi: 10.2352/J.ImagingSci.Technol.2019.63.3.030405.

[6] X. Liu, Y. Jiang, P. Jain, and K.-H. Kim, "TAR: Enabling Fine-Grained Targeted Advertising in Retail Stores," in *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*, Jun. 2018, pp. 323–336, doi: 10.1145/3210240.3210342.

[7] C. B. Nalty *et al.*, "A Brief Survey on Person Recognition at a Distance," in *2022 56th Asilomar Conference on Signals, Systems, and Computers*, Oct. 2022, vol. 2022-Octob, pp. 145–152, doi: 10.1109/IEEECONF56349.2022.10051819.

[8] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "VRSTC: Occlusion-Free Video Person Re-Identification," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, vol. 2019-June, pp. 7176–7185, doi: 10.1109/CVPR.2019.00735.

[9] X. Zhang, Y. Yan, J.-H. Xue, Y. Hua, and H. Wang, "Semantic-Aware Occlusion-Robust Network for Occluded Person Re-Identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 7, pp. 2764–2778, Jul. 2021, doi: 10.1109/TCSVT.2020.3033165.

[10] C. Zhao, X. Lv, S. Dou, S. Zhang, J. Wu, and L. Wang, "Incremental Generative Occlusion Adversarial Suppression Network for Person ReID," *IEEE Trans. Image Process.*, vol. 30, pp. 4212–4224, 2021, doi: 10.1109/TIP.2021.3070182.

[11] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person Re-identification: Past, Present and Future," *arXiv Computer Vision and Pattern Recognition*, Oct. 10, pp. 1-20, 2016. https://arxiv.org/abs/1610.02984v1.

[12] Q. Leng, M. Ye, and Q. Tian, "A Survey of Open-World Person Re-Identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 4, pp. 1092–1108, Apr. 2020, doi: 10.1109/TCSVT.2019.2898940.

[13] A. Zahra, N. Perwaiz, M. Shahzad, and M. M. Fraz, "Person re-identification: A retrospective on domain specific open challenges and future trends," *Pattern Recognit.*, vol. 142, p. 109669, Oct. 2023, doi: 10.1016/j.patcog.2023.109669.

[14] R. Chalapathy, N. L. D. Khoa, and S. Chawla, "Robust Deep Learning Methods for Anomaly Detection," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Aug. 2020, pp. 3507–3508, doi: 10.1145/3394486.3406704.

[15] L. Liu *et al.*, "Deep Learning for Generic Object Detection: A Survey," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 261–318, Feb. 2020, doi: 10.1007/s11263-019-01247-4.

[16] J. Marin, D. Vazquez, A. M. Lopez, J. Amores, and L. I. Kuncheva, "Occlusion Handling via Random Subspace Classifiers for Human Detection," *IEEE Trans. Cybern.*, vol. 44, no. 3, pp. 342–354, Mar. 2014, doi: 10.1109/TCYB.2013.2255271.

[17] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004, doi: 10.1023/B:VISI.0000029664.99615.94.

[18] Y. Hu, S. Liao, Z. Lei, D. Yi, and S. Z. Li, "Exploring Structural Information and Fusing Multiple Features for Person Re-identification," in *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Jun. 2013, pp. 794–799, doi: 10.1109/CVPRW.2013.119.

[19] G. Xie *et al.*, "Pose-guided feature region-based fusion network for occluded person re-identification," *Multimed. Syst.*, vol. 29, no. 3, pp. 1771–1783, Jun. 2023, doi: 10.1007/s00530-021-00752-2.

[20] D. K. Dastur *et al.*, "The B-vitamins in malnutrition with alcoholism: A model of intervitamin relationships," *Br. J. Nutr.*, vol. 36, no. 2, pp. 143–159, Sep. 1976, doi: 10.1017/S0007114500020158.

[21] J. Miao, Y. Wu, and Y. Yang, "Identifying Visible Parts via Pose Estimation for Occluded Person Re-Identification," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 33, no. 9, pp. 4624–4634, Sep. 2022, doi: 10.1109/TNNLS.2021.3059515.

[22] Z. Zhao, R. Song, Q. Zhang, P. Duan, and Y. Zhang, "JoT-GAN: A Framework for Jointly Training GAN and Person Re-Identification Model," *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 18, no. 1s, pp. 1–18, Feb. 2022, doi: 10.1145/3491225.

[23] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person Transfer GAN to Bridge Domain Gap for Person Re-identification," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 79–88, doi: 10.1109/CVPR.2018.00016.

[24] C. Zhang, L. Zhu, S. Zhang, and W. Yu, "PAC-GAN: An effective pose augmentation scheme for unsupervised cross-view person re-identification," *Neurocomputing*, vol. 387, pp. 22–39, Apr. 2020, doi: 10.1016/j.neucom.2019.12.094.

[25] T. He, X. Shen, J. Huang, Z. Chen, and X.-S. Hua, "Partial Person Re-identification with Part-Part Correspondence Learning," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2021, pp. 9101–9111, doi: 10.1109/CVPR46437.2021.00899.

[26] Y. Li, J. He, T. Zhang, X. Liu, Y. Zhang, and F. Wu, "Diverse Part Discovery: Occluded Person Re-identification with Part-Aware Transformer," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2021, pp. 2897–2906, doi: 10.1109/CVPR46437.2021.00292.

[27] X.-P. Lin and Y.-B. Yang, "An Adaptive Part-Based Model For Person Re-Identification," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2021, vol. 2021-June, pp. 1965–1969, doi: 10.1109/ICASSP39728.2021.9415086.

[28] Z. Yao, X. Wu, Z. Xiong, and Y. Ma, "A Dynamic Part-Attention Model for Person Re-Identification," *Sensors*, vol. 19, no. 9, p. 2080, May 2019, doi: 10.3390/s19092080.

[29] L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-Learned Part-Aligned Representations for Person Re-identification," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, vol. 2017-Octob, pp. 3239–3248, doi: 10.1109/ICCV.2017.349.

[30] W.-S. Zheng, X. Li, T. Xiang, S. Liao, J. Lai, and S. Gong, "Partial Person Re-Identification," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015, pp. 4678–4686, doi: 10.1109/ICCV.2015.531.

[31] i-LIDS Team, "Imagery Library for Intelligent Detection Systems (i-LIDS) A Standard for Testing Video Based Detection Systems," in *Proceedings 40th Annual 2006 International Carnahan Conference on Security Technology*, Oct. 2006, pp. 75–80, doi: 10.1109/CCST.2006.313432.

[32] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Development Kit," *Pattern Analysis, Statistical Modelling and Computational Learning, Tech. Rep*, pp. 1-45, 2012. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=9d0df3b123a78c34f6ca874d51a321b33a9f1199.

[33] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning Discriminative Features with Multiple Granularities for Person Re-Identification," in *Proceedings of the 26th ACM international conference on Multimedia*, Oct. 2018, pp. 274–282, doi: 10.1145/3240508.3240552.

[34] J. Dang, X. Tang, and S. Li, "HA-FPN: Hierarchical Attention Feature Pyramid Network for Object Detection," *Sensors*, vol. 23, no. 9, p. 4508, May 2023, doi: 10.3390/s23094508.

[35] Y. Wang, S. Yang, S. Liu, and Z. Zhang, "Cross-Domain Person Re-identification: A Review," in *Lecture Notes in Electrical Engineering*, vol. 653, Springer Science and Business Media Deutschland GmbH, 2021, pp. 153–160, doi: 10.1007/978-981-15-8599-9_19.

[36] S. Liao, Y. Hu, Xiangyu Zhu, and S. Z. Li, "Person re-identification by Local Maximal Occurrence representation and metric learning," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015, vol. 07-12-June, pp. 2197–2206, doi: 10.1109/CVPR.2015.7298832.

[37] Y.-C. Chen, W.-S. Zheng, and J. Lai, "Mirror Representation for Modeling View-Specific Transform in Person Re-Identification," in *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)*, 2016, pp. 3402–3408. [Online]. Available at: https://www.ijcai.org/Proceedings/15/Papers/479.pdf.

[38] N. Perwaiz, M. M. Fraz, and M. Shahzad, "Hierarchical Refined Local Associations for Robust Person Re-Identification," in *2019 International Conference on Robotics and Automation in Industry (ICRAI)*, Oct. 2019, pp. 1–6, doi: 10.1109/ICRAI47710.2019.8967389.

[39] L. Zhang, T. Xiang, and S. Gong, "Learning a Discriminative Null Space for Person Re-identification," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, vol. 2016-Decem, pp. 1239–1248, doi: 10.1109/CVPR.2016.139.

[40] Y. Sun, L. Zheng, W. Deng, and S. Wang, "SVDNet for Pedestrian Retrieval," in *Proceedings of the IEEE International Conference on Computer Vision*, 2018, pp. 3820–3828, doi: 10.1109/ICCV.2017.410.

[41] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random Erasing Data Augmentation," *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 07, pp. 13001–13008, Apr. 2020, doi: 10.1609/aaai.v34i07.7000.

[42] C. Yan, G. Pang, J. Jiao, X. Bai, X. Feng, and C. Shen, "Occluded Person Re-Identification with Single-scale Global Representations," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 11855–11864, doi: 10.1109/ICCV48922.2021.01166.

[43] Q. Yang, P. Wang, Z. Fang, and Q. Lu, "Focus on the Visible Regions: Semantic-Guided Alignment Model for Occluded Person Re-Identification," *Sensors*, vol. 20, no. 16, p. 4431, Aug. 2020, doi: 10.3390/s20164431.

[44] X. Zhong, M. Wang, W. Liu, J. Yuan, and W. Huang, "SCPNet: Self-constrained parallelism network for keypoint-based lightweight object detection," *J. Vis. Commun. Image Represent.*, vol. 90, p. 103719, Feb. 2023, doi: 10.1016/j.jvcir.2022.103719.

[45] L. He, J. Liang, H. Li, and Z. Sun, "Deep Spatial Feature Reconstruction for Partial Person Re-identification: Alignment-free Approach," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 7073–7082, doi: 10.1109/CVPR.2018.00739.

[46] H. Luo, W. Jiang, X. Fan, and C. Zhang, "STNReID: Deep Convolutional Networks With Pairwise Spatial Transformer Networks for Partial Person Re-Identification," *IEEE Trans. Multimed.*, vol. 22, no. 11, pp. 2905–2913, Nov. 2020, doi: 10.1109/TMM.2020.2965491.

[47] Y. Sun *et al.*, "Perceive Where to Focus: Learning Visibility-Aware Part-Level Features for Partial Person Re-Identification," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, vol. 2019-June, pp. 393–402, doi: 10.1109/CVPR.2019.00048.

[48] L. He, X. Liao, W. Liu, X. Liu, P. Cheng, and T. Mei, "FastReID: A Pytorch Toolbox for General Instance Re-identification," in *Proceedings of the 31st ACM International Conference on Multimedia*, Oct. 2023, pp. 9664–9667, doi: 10.1145/3581783.3613460.

[49] Z. Pang, J. Guo, W. Sun, Y. Xiao, and M. Yu, "Cross-domain person re-identification by hybrid supervised and unsupervised learning," *Appl. Intell.*, vol. 52, no. 3, pp. 2987–3001, Feb. 2022, doi: 10.1007/s10489-021-02551-8.

[50] H. Zhang, S. Wang, N. Wang, S. Liu, and Z. Zhang, "Efficiency Evaluation of Deep Model for Person Re-identification," in *Lecture Notes in Electrical Engineering*, vol. 653, Springer Science and Business Media Deutschland GmbH, 2021, pp. 130–136, doi: 10.1007/978-981-15-8599-9_16.