Journal of Telecommunications and the Digital Economy

CNN-based Occluded Person Re-identification in a Multi Camera Environment

Ali Imran Bin Shahrin
Multimedia University

Noramiza Binti Hashim
Multimedia University

Abstract: In the context of rising global urban security concerns and the growing use of surveillance cameras, this study aims to enhance individual identification accuracy in occlusion scenarios using deep learning. Four CNN-based models for person re-identification are analyzed and put into practice. Additionally, comparative studies are conducted, and the model's performance is assessed using the Market-1501 and Occluded-Reid datasets. We propose the use of ensemble learning and convolutional neural networks (CNNs) to address occlusion issues. Our results show that the ensemble approach performs better in re-identification tasks than traditional deep learning algorithms with an improvement of 1%–2% in mAP and Rank-1 scores, respectively.

Keywords: Person re-identification, deep learning, ensemble deep learning

Introduction

Person re-identification is a crucial discipline in computer vision, tasked with matching a person's identity across disparate surveillance cameras using different types of algorithms. This technology, which is essential for ensuring safety in crowded areas like supermarkets, airports, and cities, faces significant obstacles. Person re-identification in a multi-camera environment involves multiple camera setups to capture person identities in a certain location. As compared to a single camera setup, where person identities are caught in only one perspective, a multi-camera environment must capture the same person identity across multiple camera perspectives. With multiple camera perspectives, significant variations in an individual's appearance, due to viewpoint variation, scaling problems and occlusion, make identifying a person more difficult.

The most challenging of these obstacles is occlusion, the phenomenon where an object or person is completely or partially obscured, as it greatly reduces the visual information in an image, because of interfering elements like pedestrians and moving objects (Wei *et al.*, 2022). Although there are other factors that make this problem worse, like viewpoint and image scale variability, occlusion remains the main problem. Our research is built around the application of deep learning algorithms, with a focus on Convolutional Neural Networks (CNNs) and a straightforward ensemble learning technique. Ensemble learning, which combines predictions from various models, frequently improves performance. It is especially promising for tackling the issue of occlusion in person re-identification. Our study aims to shed light on challenging single-target, multi-camera settings plagued with occlusion issues.

Related Works

Person re-identification involves comparing identities across photos captured at different times and locations, a process referred to as multi-target multi-camera (MTMC) tracking, applied in fields like crowd analysis, traffic management, and municipal security (Shim *et al.*, 2021)

Person re-identification in a single camera and multi-camera setting

Person re-identification can be tackled from two different camera perspectives: one being from single camera settings; the other from multi-camera settings. For a single camera setting, Zhang *et al.* (2019) have managed to perform person re-identification by taking only a single camera perspective for training from a multi-camera dataset, such as Market-1501 or DukeMTMC-reID.

Significant research has been dedicated to resolving person re-identification challenges in multi-camera settings. Numerous CNN-based methods have been proposed, with each introducing innovative techniques to improve re-identification accuracy. For instance, Zhou *et al.* (2019) developed an Omni-Scale Network (OSNet) to facilitate Omni-Scale feature learning. Despite its small model size, OSNet, capable of being trained from scratch on existing re-identification datasets, has shown to outperform larger models like ResNet50 (He *et al.*, 2015) and DenseNet (Huang *et al.*, 2016). On the other hand, Yan *et al.* (2021) proposed a re-identification model that employs a bounded distance loss on occluded data to learn pedestrian features. In addition, the research presented in He *et al.* (2019) integrates pyramid pooling with a Full Convolution Network (FCN), achieving an impressive accuracy of 95.42% on the Market-1501 dataset. Lastly, the work by Wang *et al.* (2020) merges a CNN model with an adaptive direction graph, achieving a promising accuracy of 55.1% on the Occluded-Duke dataset.

Deep-learning-based Person Re-identification approaches

The two main approaches to person re-identification are hand-crafted features and deep learning-based methods. Hand-crafted features encompass low-level, mid-level, and high-level semantic representations but are less utilized recently due to complications with lighting changes and occlusion (Wei *et al.*, 2022).

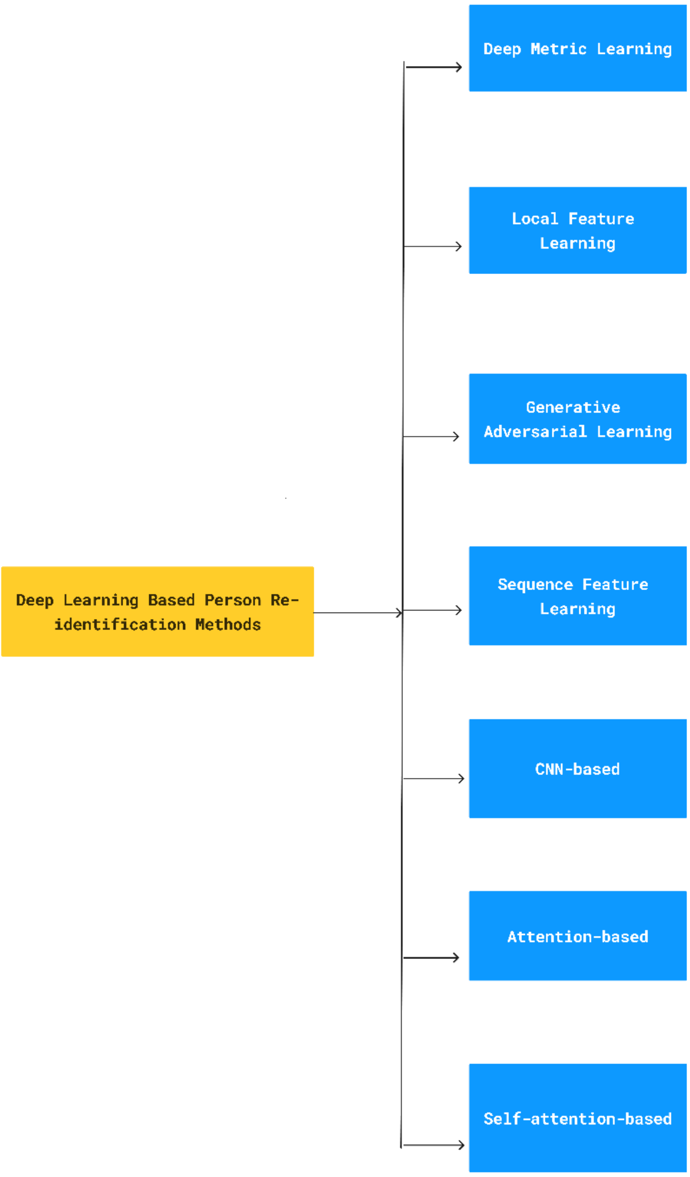
Deep Learning Based Person Re-identification Methods
Deep Metric Learning
Local Feature Learning
Generative Adversarial Learning
Sequence Feature Learning
CNN-based
Attention-based
Self-attention-based

Figure 1. Deep-Learning-Based Person Re-identification Methods

Deep learning, on the other hand, has gained popularity and can be categorized into deep metric learning, local feature learning, generative adversarial learning, and sequence feature learning (Ming *et al.*, 2022), as shown in Figure 1.

Deep metric learning focuses more on training a deep neural network to learn a distance metric, thereby assisting in accurately measuring the similarity between different identities of an image. However, deep metric learning requires a considerable amount of data to perform well and can struggle with overfitting in smaller datasets (Kaya & Bilge, 2019). Local feature learning focuses on robust features to combat certain challenges, such as pose, illumination and clothing changes. Although local feature learning provides consistency in identification on appearance changes, it sometimes still suffers a variation in pose and background clutter (Wang *et al.*, 2018). On the other hand, generative adversarial learning uses a generative model to create new images and a discriminative model to distinguish real from generated images. While generative adversarial learning is useful for data augmentation, it commonly suffers from poor interpretability of neural networks (Wang *et al.*, 2017). Finally, sequence feature learning is designed to extract features from sequences of pedestrian images. It is useful in video-based person re-identification. However, it requires a large amount of sequential data and can be computationally expensive (Wei *et al.*, 2022).

Recently, newer deep learning-based solutions for person re-identification have emerged, including Convolutional Neural Network (CNN) based solutions, attention-based solutions, and self-attention-based solutions. CNN-based solutions leverage the deep learning capabilities of CNNs, using a series of convolutional layers to progressively learn image attributes and their surroundings. However, the performance of CNNs is heavily reliant on image quality (Karahan *et al.*, 2016). For example, if the training data consisted of image degradation such as occlusion, the model tends to perform much worse compared to non-occluded scenarios. Attention-based solutions focus on specific attributes, ignoring less useful background information. While effective at enhancing focus on critical details, attention-based solutions lacked semantic information from the local feature regions that they are extracting from, making it difficult to comprehend (Wei *et al.*, 2022). Self-attention-based solutions employ transformers and multi-head self-attention mechanisms to learn image embeddings. These solutions provide promising results but can be computationally heavy and less effective on smaller datasets due to their high-capacity models.

Datasets for Person Re-identification

In this research, we have decided to focus on the Market-1501 and Occluded-ReID datasets, which examine occlusion in person re-identification. A vast collection of 32,668 images from 1,501 different identities can be found in the exhaustive Market-1501 dataset developed by Zheng *et al.* (2015). In addition, the Occluded-ReID dataset created for occlusion scenarios by Zhuo *et al.* (2018) will be crucial for the needs of this study. The numerous other datasets that are readily available and contribute to the rich diversity in the field of person re-identification must also be mentioned. Examples of this kind include the CUHK01 and CUHK03 datasets by Li, W. *et al.* (2018) and the DukeMTMC-ReID by Ristani *et al.* (2016), all of which vary in the number and variety of images and identities they contain. The MSMT-17 dataset, also by Wei *et al.* (2017), stands out as the largest dataset currently available, while the VIPeR dataset by Gray *et al.* (2007) poses difficulties in terms of real-world conditions despite its smaller size. Each dataset presents distinct characteristics and difficulties, opening the door for numerous studies and developments in the field (see Table 1).

Table 1. Datasets for image-based person re-identification

Dataset	**Number of identities**	**Number of images**	**Number of cameras**
Market-1501 (Zheng *et al.*, 2015)	1501	32,668	6
DukeMTMC-ReID (Ristani *et al.*, 2016)	1,852	22,515	8
CUHK01 (Li, W., *et al.*, 2018)	971	3,884	2
VIPeR (Gray *et al.*, 2017)	632	1,264	2
CUHK03 (Li, W., *et al.*, 2018)	1,360	13,614	6
MSMT-17 (Wei *et al.*, 2017)	4,101	126,441	15
Occluded-ReID (Zhuo *et al.*, 2018)	200	2,000	1

Table 2. Datasets for video-based person re-identification

Dataset	**Number of identities**	**Number of images**	**Number of cameras**
DukeMTMC-VideoReID (Wu *et al.*, 2018)	702	2,196 training videos and 2,636 test-related videos	8
MARS (Zheng *et al.*, 2015)	1,261	1,191,003	8
iLIDS-VID (Li, M., *et al.*, 2018)	300	22,000	2

Several important datasets in the field of video-based person re-identification deserve to be noted as well. A sizable collection of 2,196 training videos and 2,636 test-related videos for 702 identities are available in the DukeMTMC-VideoReID dataset (Wu *et al.*, 2018). The MARS (Zheng *et al.*, 2015) dataset is notable for its size, containing over 1.19 million images from 1,261 distinct identities, captured from 8 camera viewpoints. The iLIDS-VID dataset (Li. M., *et al.*, 2018) contains 22,000 images spread across 300 identities, recorded from 2 camera views, and presents a condensed yet difficult environment for video-based re-identification studies. Each dataset (see Table 2) adds unique features and insights that broaden the scope of research on video-based person re-identification.

Challenges in Person Re-identification

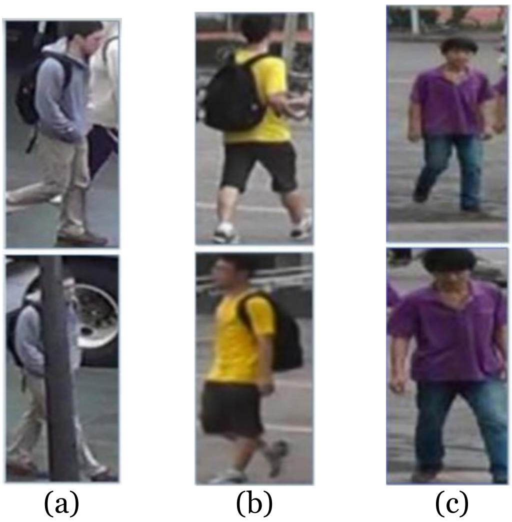

Figure 2. Challenges in Person Re-identification (a) Occlusion; (b) Viewpoint variation; (c) Scale issues

Despite being heavily researched, person re-identification presents challenges, such as the need for more camera perspectives, high cost of data labelling for unsupervised approaches, accuracy of manual labelling, and handling appearance features, like clothing similarity or attire changes. In person re-identification, there still exist many issues that affect the performance of person re-identification, which we will highlight. One of them is occlusion, as seen in Figure 2(a).

Occlusion happens when a person is blocked by an overlapping object that can lead to inaccurate information from an image. In a person re-identification scenario, people are often occluded by environmental objects, such as a traffic sign board, vehicles in a parking lot or simply by other pedestrians. When a person is occluded, or when a portion of their body is hidden from view, the features that are extracted from the entire image may contain some distracting information and result in errors if the model cannot differentiate between the individual region and the occluded region (Zahra *et al.*, 2022).

Another problem that arises for an image when conducting person re-identification is viewpoint variation. Viewpoint variation is when a person appears in different positions of the capturing camera, as illustrated in Figure 2(b). It is challenging to create a model with great generalizability, because a person's visual appearance can change depending on the angle and proximity from which they are photographed by various cameras.

The final issue that arises when conducting person re-identification is scale differences. Scale difference is when an object appears to be in a smaller or larger form, as seen in Figure 2(c). The issue of scale difference is a challenging one to resolve, because image appearances are greatly influenced by a given camera setting. Instead of using a multiple-scale technique, most person re-identification methods use a fixed-scale approach (Ahmed *et al.*, 2015; Li *et al.*, 2014).

Deep ensemble learning

Deep ensemble learning has become an effective machine learning technique in recent years, providing a solid way to improve model performance (Serbetci & Akgul, 2020; Yang *et al.*, 2018). By utilizing the strengths of multiple learning models, this method lowers the risk of overfitting to training data, increasing robustness and generalizability. As a result, deep ensemble learning has demonstrated to be helpful in several challenging tasks, including person re-identification.

In the context of person-reidentification, ensemble approaches are employed to take advantage of the strengths of multiple person re-identification models or methods to overcome some of the limitations of individual methods. An ensemble of deep learning-based person re-identification models can contain a mixture of CNN models, attention mechanism models, self-attention mechanism models, and other types of models (Mauldin *et al.*, 2019). By strategically combining these different approaches, an ensemble model can potentially overcome some of the limitations of the individual approaches. For instance, while CNNs can sometimes overlook critical local features, attention mechanisms can help focus on these details. Conversely, where attention mechanisms may overemphasize certain regions and neglect others, the global feature learning capability of CNNs can balance this out. The self-attention mechanism, with its focus on relationship between all parts of an image, can further enhance this balance, adding a comprehensive understanding of the image context.

One key consideration when designing an ensemble model is the issue of diversity. Having a diverse set of person re-identification models can make the ensemble more robust and improve its generalization capabilities. This is because different types of models can excel in different aspects of a person re-identification task and be able to compensate for each other's weaknesses. However, ensemble learning comes with its own challenges. It is more prone towards overfitting (Li *et al.*, 2019) and more computationally expensive compared to single-model person re-identification approaches. Additionally, ensemble learning also requires careful design and tuning to ensure that the base person re-identifications are complementary with each other and that their outputs are combined in an effective way. Some of the methods of approaching ensemble learning are dropout ensemble, snapshot ensemble and boosting.

Deep ensemble learning has been investigated for person re-identification in a few notable studies. In Srivastava *et al.* (2014) and Singh *et al.* (2016), a regularization technique called dropout is often used and seen as a form of ensemble learning. During training, dropout involves randomly "dropping out" or deactivating a proportion of neurons in the network. Apart from that, snapshot ensemble can also be seen being used by Garipov *et al.* (2018). Snapshot ensemble involves saving model parameters at several points during training, and then averaging their predictions. Boosting is also another ensemble technique that involves training several models sequentially, where each model learns from the errors of its predecessor. Li *et al.* (2019) utilized boosting by adaptively assembling features and metrics from a ranking perspective.

Research gap

Currently, the training strategies employed for ensemble learning models often necessitate the independent training of multiple deep learning networks. This can prove to be computationally expensive and raises the question of whether there could be more efficient approaches to use ensemble deep learning. In addition, despite noteworthy advances in the field, occlusion remains a substantial problem in person re-identification. The existing research field is notably insufficient when it comes to addressing occlusion issues through ensemble deep learning. While past studies (Serbetci & Akgul, 2020) have demonstrated that ensemble learning can significantly boost the performance of person re-identification in general datasets like Market-1501, its implementation in occlusion-based datasets, such as Occluded-Reid, remains in a premature stage. This underlines an urgent need for more intensive research in this area. Therefore, we propose a baseline solution for a simple ensemble learning method called model averaging for other researchers to implement this technique and possibly improve their own person re-identification models.

Research Methodology

Data collection

For this research, we use publicly available datasets. The Market-1501 dataset (Zheng *et al.*, 2015) is used for image-based person re-identification. It comprises 1,501 IDs and 32,668 bounding boxes. As an attempt at solving the problem of occlusion in person re-identification, the Occluded-Reid dataset (Zhuo *et al.*, 2018) will be used. This dataset consists of 2000 images of 200 occluded identities. Each identity contains 5 full body person images and 5 occluded person images.

Data pre-processing and feature extraction

Data splitting and feature extraction will both be applied to the chosen datasets. Each dataset has a distinct split that was established by the dataset's creators, and this study will abide by that split. Consider the 32,668 image Market-1501 dataset as an illustration. Of these images, 12,936 will be used for training, 3,368 as the query set, and the remaining 15,913 will be used as the gallery set. The Occluded-Reid dataset uses occluded person images as query, full-body person images as the gallery, and randomly divided identities with half going to training and the other half going to testing.

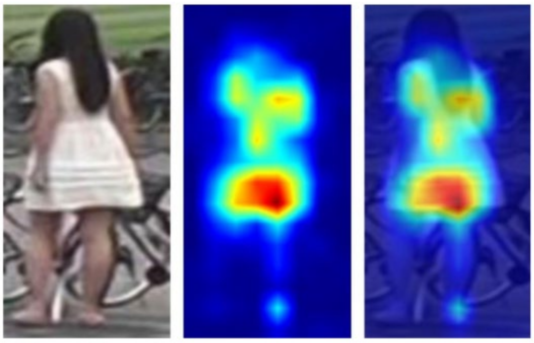

Figure 3. Feature extraction using OSNet (Zhou *et al.*, 2019)

Following the data splitting procedure, feature extraction is done to extract important attributes from the source data. The research's feature extraction procedure is dependent on the model being used. For example, if we are choosing the OSNet model, it performs two functions: it extracts features; and acts as a predictive model. In other words, the model does not require any additional feature extraction algorithms or manual intervention to extract pertinent features from the input data directly. It is crucial to remember that the features that are extracted have an unbreakable connection to the particulars of the selected model. For instance, based on Figure 3, depending on the nature and depth of its architecture, it might extract features from the input data such as the clothing information of a specific identity. In this method, the internal layers of the loaded model's architecture are used to automatically identify and extract significant features from the data. The subsequent phases of model training and testing make use of this extracted data.

Model training and testing

We implement the deep learning method known as CNN for model training and testing. The goal is to uniquely classify person images based on the specified dataset. Model configuration also includes choosing an optimizer and loss function. In the model training and testing phase, a selection of baseline models provided by the torchreid library (Zhou *et al.*, 2019), including ResNet50 (He *et al.*, 2015; Xie *et al.*, 2016), OSNet (Zhou *et al.*, 2019) and its variations, DenseNet (Huang *et al.*, 2016), and others, will be utilized.

The primary reasons for selecting these models are threefold. Firstly, these models have been extensively validated in the literature and have demonstrated high performance in a range of computer vision tasks, including person re-identification. Their robust performance is attributed to their unique architectures that allow for the extraction of hierarchical, multi-scale features, which are crucial in the context of person re-identification. Secondly, these models provide a diverse set of architectures for comparison and ensemble learning. ResNet50, for instance, introduces a residual learning framework to ease the training of networks, while DenseNet connects each layer to every other layer in a feed-forward fashion, enabling feature reuse. On the other hand, OSNet incorporates omni-scale feature learning into deep neural networks, making it highly effective for person re-identification. The diversity in architecture not only allows for comprehensive comparison but also increases the potential for ensemble learning to achieve more robust results. Finally, these models are readily available in the torchreid library (Zhou *et al.*, 2019), making them convenient to employ in our research framework. The availability of these models, along with the necessary training and evaluation utilities, reduces implementation time and allows for a focus on the experimental design and results analysis.

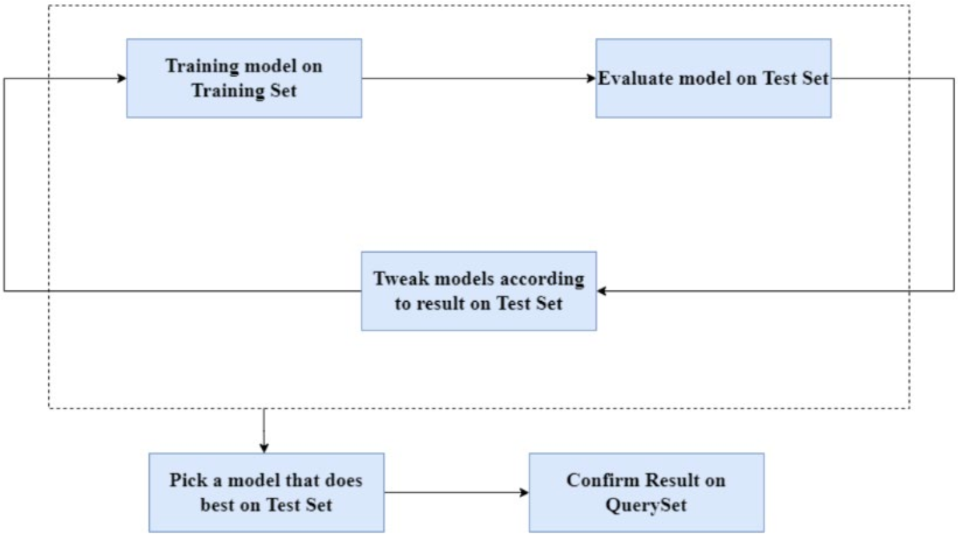
Training model on Training Set
Evaluate model on Test Set
Tweak models according to result on Test Set
Pick a model that does best on Test Set
Confirm Result on QuerySet

Figure 4. Framework of the model training and testing process

Loss function and optimizer

We used the Triplet loss engine, which is renowned for its effectiveness in person re-identification tasks, during the training and testing phases of our models (Hermans *et al.*, 2017). By enhancing anchor-positive-negative triplets, this function increases the capacity of learned embeddings for discrimination. The Adam optimizer (Kingma & Ba, 2014) was used, and its learning rate was set to 0.0003, because it has a track record of success (Zhou *et al.*, 2019) and is computationally efficient for deep-learning model training. These hyperparameter selections, along with our choice of deep learning models, were made with the goal of creating a robust and effective system for person re-identification.

Simple ensemble learning

This study makes use of Model Averaging (Srivastava *et al.*, 2014), a type of ensemble learning technique. The predictions of various models are combined in this method, which is renowned for its effectiveness in improving model performance. This is accomplished by averaging the results from each model in a way that lowers variance, encourages generalizability, and minimizes the risk of overfitting.

Model averaging is used in this study in a novel way. We combine the unique predictive power of two distinct models by averaging the features they extract, as opposed to averaging predictions. By encapsulating each model's unique strengths in a new feature space, this process increases the robustness of the resulting ensemble model. While not explicitly tailored to occlusion, viewpoint variation, or scale differences individually, we have conducted comprehensive evaluations, including testing on an occlusion-focused dataset. The results, presented in a subsequent section, demonstrate notable effectiveness in addressing challenges posed by occluded images. The Euclidean metric is used to calculate a distance matrix after averaging the features. This matrix measures how different each pair of features from the query set and the gallery set are from one another. The foundation for subsequent evaluation tasks is provided by this integral computation, demonstrating how model averaging can be creatively used to support person re-identification efforts.

Integration of learners from CNN

As we highlight the use of model averaging, we also need to explain how this ensemble approach is specifically customized for our use case. We purposefully include CNNs as the base learners in the group. Notably, the superior individual performance of OSNet models is the main factor in CNN selection. OSNet models, as proposed by Zhou *et al.* (2019), exhibit remarkable capabilities in extracting intricate features from complex data. These models have demonstrated robustness and high accuracy as standalone entities. In our ensemble framework, we leverage the inherent strengths of OSNet models, combining OSNet with its sub-models, such as OSNet_x0_75, with OSNet_x1_0, which we can observe in the result subsection later. The individual strengths of OSNet models—which demonstrate high mAP scores, Rank-1, Rank-5, Rank-10, and Rank-20 accuracy metrics on the Market-1501 dataset—are the main justification for using them. Our goal is to leverage the complementary features of two different OSNet models to improve the performance of the ensemble model in a synergistic way. As the following model comparison section shows, this strategic combination works especially well for addressing issues like occlusion.

Model evaluation

Two metrics are employed for evaluating the performance of the models: the CMC curve and the mAP score.

- Cumulative Match Characteristic (CMC) Curve: The CMC curve provides a visual representation of the identification job's performance by comparing the number of candidates returned with the probability of accurate identification (Paisitkriangkrai *et al.*, 2015). The curve is generated by determining the accuracy of the top-k retrieved images in the test set containing the query identity. Based on Figure 5, the accuracy of Rank-1 will increase generally because the first image is a correct prediction based on the query. But as the rank gradually increases, for example in Rank-5, the first 5 images will be taken to evaluate the prediction based on the query. As we can observe, the first 5 images contain 3 wrong predictions, making the accuracy at Rank-5 decrease.

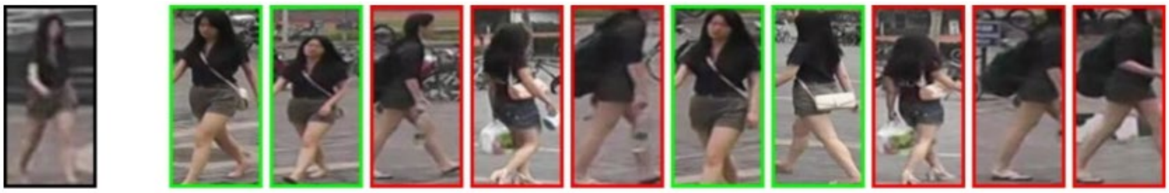

Figure 5. Predictions made by a model based on query image.

- Mean Average Precision (mAP) Score: The mAP score calculates the average precision for each query image. It considers the number of correct person retrievals, helping gauge the model's performance at retrieving more accurate people and ranking the correct person higher in the list of retrieved images.

Calculation Results and Discussion

When examining the person re-identification task on the non-occluded dataset, the OSNet ensemble model surpasses other individual models, achieving a mAP score of 80.6%, and Rank-1, Rank-5, Rank-10, and Rank-20 accuracies of 93.1%, 97.5%, 98.5%, and 99.1%, respectively. In addition, when compared to single architecture models, our model seems to be the best performing among them all. Our model's performance is likely due to the ensemble model's ability to leverage the distinct strengths of multiple models, thereby enhancing its adaptability and generalization capabilities (Perin *et al.*, 2020). Such strengths make it an effective approach in handling the unique challenges posed by person re-identification tasks.

Table 3. Comparison of rank 1/5/10/20 on the Market-1501 dataset
(*Result highlighted in bold is the best performing in our experimentation)

Dataset	**Market-1501**				
Model Name	mAP score	Rank-1	Rank-5	Rank-10	Rank-20
DenseNet121 (Huang *et al.*, 2016)	64.0%	83.3%	92.4%	94.9%	96.7%
DenseNet169 (Huang *et al.*, 2016)	65.8%	84.2%	92.8%	95.2%	96.7%
DenseNet201 (Huang *et al.*, 2016)	63.5%	81.9%	91.8%	94.4%	96.2%
ResNet101 (He *et al.*, 2015)	67.8%	84.1%	93.9%	95.9%	97.4%
ResNet50 (He *et al.*, 2015)	67.6%	84.6%	93.5%	96.3%	97.8%
ResNext50 (Xie *et al.*, 2016)	67.9%	84.8%	93.4%	95.7%	97.6%
OSNet_x0_75 (Zhou *et al.*, 2019)	76.4%	91.3%	97.1%	98.3%	98.7%
OSNet_x1_0 (Zhou *et al.*, 2019)	79.5%	92.5%	97.2%	98.3%	98.9%
OSNet Ensemble	**80.6%**	**93.1%**	97.5%	98.5%	99.1%

Table 4. Comparison of rank 1/5/10/20 for state-of-art models on the Market1501 dataset

Dataset	**Market-1501**				
Model Name	mAP score	Rank-1	Rank-5	Rank-10	Rank-20
OSNet Ensemble	80.6%	93.1%	97.5%	98.5%	99.1%
PAT (Li *et al.*, 2021)	88.0%	95.4%	-	-	-
GASM (He & Liu, 2020)	84.7%	95.3%	-	-	-

However, when compared to sophisticated state-of-the-art architecture, our ensemble model tends to fall off in terms of mAP score and Rank-1. This might be because GASM leverages spatial features that can help extract rich information from the input data, creating a more effective person re-identification model. Apart from that, PAT might perform better because of its dealing with occlusion. The model addresses occlusion using part discovery, which is useful in crowded scenarios, making the model learn and recognizing individual parts.

For this occlusion-based dataset, ensemble learning also helps improve the general performance of the person re-identification model. For comparative analysis, we have benchmarked our results against other state-of-the-art CNN methods, such as AFBP (see Table 5). Numerically, the ensemble model exhibits the highest mAP score at 56.0% compared to the individual models, suggesting superior average precision across all queries.

Furthermore, when compared to more sophisticated methods, such as AFBP, it achieves a comparable performance across Rank-n accuracy, attaining Rank-1, Rank-5, Rank-10, and Rank-20 scores of 62.8%, 81.8%, 88.4%, and 93.4% respectively. This can be because ensemble learning can reduce model variance (Rajaraman *et al.*, 2019) by optimally combining predictions from multiple models.

Table 5. Comparison of rank 1/5/10/20 on the Occluded-Reid dataset
(*Result highlighted in bold is the best performing in our experimentation)

Dataset	**Occluded-Reid**				
Model Name	mAP score	Rank-1	Rank-5	Rank-10	Rank-20
DenseNet121 (Huang *et al.*, 2016)	49.6%	54.2%	73.2%	81.2%	86.6%
DenseNet169 (Huang *et al.*, 2016)	55.0%	61.8%	78.8%	85.0%	90.0%
DenseNet201 (Huang *et al.*, 2016)	55.4%	60.2%	80.4%	86.6%	90.8%
ResNet101 (He *et al.*, 2015)	41.1%	43.8%	62.4%	72.2%	79.8%
ResNet50 (He *et al.*, 2015)	45.7%	49.8%	67.8%	78.2%	86.6%
ResNext50 (Xie *et al.*, 2016)	50.4%	55.0%	74.2%	80.8%	87.8%
OSNet_ibn_x1_0 (Zhou *et al.*, 2019)	53.6%	58.8%	77.4%	84.4%	90.4%
OSNet_x1_0 (Zhou *et al.*, 2019)	51.8%	60.4%	76.6%	84.0%	90.8%
OSNet Ensemble	**56.0%**	**62.8%**	81.8%	88.4%	93.4%
AFBP (Zhuo *et al.*, 2018)	-	68.14%	88.29%	-	-

Conclusion

For this occlusion-based dataset, ensemble learning also helps improve the general performance of the person re-identification model. For comparative analysis, we have benchmarked our results against other state-of-the-art CNN methods such as AFBP. Numerically, the ensemble model exhibits the highest mAP score at 56.0% compared to the individual models, suggesting superior average precision across all queries. Furthermore, when compared to more sophisticated methods such as AFBP it achieves a comparable performance across Rank-n accuracy, attaining Rank-1, Rank-5, Rank-10, and Rank-20 scores of 62.8%, 81.8%, 88.4%, and 93.4% respectively. This can be because ensemble learning can reduce model variance (Rajaraman *et al.*, 2019) by optimally combining predictions from multiple models.

Acknowledgement

A version of this paper was presented at the third International Conference on Computer, Information Technology and Intelligent Computing, CITIC 2023, held in Malaysia on 26–28 July 2023.

References

Ahmed, E., Jones, M., & Marks, T. K. (2015). An Improved Deep Learning Architecture for Person Re-Identification. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 2015, pp. 3908–3916. https://doi.org/10.1109/CVPR.2015.7299016

Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D., & Wilson, A. G. (2018). *Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs*. http://arxiv.org/abs/1802.10026

Gray, D., Brennan, S., & Tao, H. (2007). *Evaluating Appearance Models for Recognition, Reacquisition, and Tracking*. https://api.semanticscholar.org/CorpusID:15225312

He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Deep Residual Learning for Image Recognition*. http://arxiv.org/abs/1512.03385

He, L., & Liu, W. (2020). Guided Saliency Feature Learning for Person Re-identification in Crowded Scenes. In Vedaldi, A., Bischof, H., Brox, T., Frahm, J. M. (eds), Computer Vision – ECCV 2020. ECCV 2020. Lecture Notes in Computer Science, vol. 12373. Springer, Cham. https://doi.org/10.1007/978-3-030-58604-1_22

He, L., Wang, Y., Liu, W., Zhao, H., Sun, Z., & Feng, J. (2019). Foreground-aware Pyramid Reconstruction for Alignment-free Occluded Person Re-identification. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019, pp. 8449–8458. https://doi.org/10.1109/ICCV.2019.00854

Hermans, A., Beyer, L., & Leibe, B. (2017). *In Defense of the Triplet Loss for Person Re-Identification*. http://arxiv.org/abs/1703.07737

Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2016). *Densely Connected Convolutional Networks*. http://arxiv.org/abs/1608.06993

Karahan, S., Yildirim, M. K., Kirtac, K., Rende, F. S., Butun, G., & Ekenel, H. K. (2016). How Image Degradations Affect Deep CNN-based Face Recognition? 2016 International Conference of the Biometrics Special Interest Group (BIOSIG), Darmstadt, Germany, 2016, pp. 1–5. https://doi.org/10.1109/BIOSIG.2016.7736924

Kaya, M., & Bilge, H. Ş. (2019). Deep metric learning: A survey. *Symmetry*, *11*(9), 1066. https://doi.org/10.3390/sym11091066

Kingma, D. P., & Ba, J. (2014). *Adam: A Method for Stochastic Optimization*. http://arxiv.org/abs/1412.6980

Li, M., Zhu, X., & Gong, S. (2018). *Unsupervised Person Re-identification by Deep Learning Tracklet Association*. http://arxiv.org/abs/1809.02874

Li, W., Zhao, R., Xiao, T., & Wang, X. (2014). DeepReID: Deep filter pairing neural network for person re-identification. Proceedings of the IEEE Computer Society Conference on

Computer Vision and Pattern Recognition, pp. 152–159. https://doi.org/10.1109/CVPR.2014.27

Li, W., Zhu, X., & Gong, S. (2018). Harmonious Attention Network for Person Re-identification. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2285–2294. https://doi.org/10.1109/CVPR.2018.00243

Li, Y., He, J., Zhang, T., Liu, X., Zhang, Y., & Wu, F. (2021). Diverse Part Discovery: Occluded Person Re-identification with Part-Aware Transformer. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021, pp. 2897–2906 https://doi.org/10.1109/CVPR46437.2021.00292

Li, Z., Han, Z., Xing, J., Ye, Q., Yu, X., & Jiao, J. (2019). High performance person re-identification via a boosting ranking ensemble. *Pattern Recognition*, *94*, 187–195. https://doi.org/10.1016/j.patcog.2019.05.022

Mauldin, T. A., Ngu, A., Metsis, V., Canby, M. E., & Tesic, J. (2019). Experimentation and Analysis of Ensemble Deep Learning in IoT Applications. *Open Journal of Internet of Things (OJIOT)*, *5*(1), 133–149. https://api.semanticscholar.org/CorpusID:264249126

Ming, Z., Zhu, M., Wang, X., Zhu, J., Cheng, J., Gao, C., Yang, Y., & Wei, X. (2022). Deep learning-based person re-identification methods: A survey and outlook of recent works. *Image and Vision Computing*, *119*. https://doi.org/10.1016/j.imavis.2022.104394

Paisitkriangkrai, S., Shen, C., & Van Den Hengel, A. (2015). Learning to rank in person re-identification with metric ensembles. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 2015, pp. 1846–1855. https://doi.org/10.1109/CVPR.2015.7298794

Perin, G., Chmielewski, Ł., & Picek, S. (2020). Strength in numbers: Improving generalization with ensembles in machine learning-based profiled side-channel analysis. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, *2020*(4), 337–364. https://doi.org/10.13154/tches.v2020.i4.337-364

Rajaraman, S., Jaeger, S., & Antani, S. K. (2019). Performance evaluation of deep neural ensembles toward malaria parasite detection in thin-blood smear images. *PeerJ*, *7*. https://doi.org/10.7717/PEERJ.6977

Ristani, E., Solera, F., Zou, R. S., Cucchiara, R., & Tomasi, C. (2016). *Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking*. http://arxiv.org/abs/1609.01775

Serbetci, A., & Akgul, Y. S. (2020). End-to-end training of CNN ensembles for person re-identification. *Pattern Recognition*, *104*, 107319. https://doi.org/10.1016/j.patcog.2020.107319

Shim, K., Yoon, S., Ko, K., & Kim, C. (2021). Multi-Target Multi-Camera Vehicle Tracking for City-Scale Traffic Management. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Nashville, TN, USA, 2021, pp. 4188–4195. https://doi.org/10.1109/CVPRW53098.2021.00473

Singh, S., Hoiem, D., & Forsyth, D. (2016). *Swapout: Learning an ensemble of deep architectures*. http://arxiv.org/abs/1605.06465

Srivastava, N., Hinton, G., Krizhevsky, A., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, *15*(56), 1929–1958. http://jmlr.org/papers/v15/srivastava14a.html

Wang, G., Yang, S., Liu, H., Wang, Z., Yang, Y., Wang, S., Yu, G., Zhou, E., & Sun, J. (2020). *High-Order Information Matters: Learning Relation and Topology for Occluded Person Re-Identification*. https://arxiv.org/abs/2003.08177

Wang, K., Gou, C., Duan, Y., Lin, Y., Zheng, X., & Wang, F. Y. (2017). Generative adversarial networks: Introduction and outlook. *IEEE/CAA Journal of Automatica Sinica*, *4*(4), 588–598. https://doi.org/10.1109/JAS.2017.7510583

Wang, K., Wang, H., Liu, M., Xing, X., & Han, T. (2018). Survey on person re-identification based on deep learning. *CAAI Transactions on Intelligence Technology*, *3*(4), 219–227. https://doi.org/10.1049/trit.2018.1001

Wei, L., Zhang, S., Gao, W., & Tian, Q. (2017). *Person Transfer GAN to Bridge Domain Gap for Person Re-Identification*. http://arxiv.org/abs/1711.08565

Wei, W., Yang, W., Zuo, E., Qian, Y., & Wang, L. (2022). Person re-identification based on deep learning — An overview. *Journal of Visual Communication and Image Representation*, *82*, 103418. https://doi.org/10.1016/j.jvcir.2021.103418

Wu, Y., Lin, Y., Dong, X., Ya, Y., Ouyang, W., & Yang, Y. (2018). Exploit the Unknown Gradually: One-Shot Video-Based Person Re-Identification by Stepwise Learning. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 5177–5186. https://doi.org/10.1109/CVPR.2018.00543

Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2016). *Aggregated Residual Transformations for Deep Neural Networks*. http://arxiv.org/abs/1611.05431

Yan, C., Pang, G., Jiao, J., Bai, X., Feng, X., & Shen, C. (2021). Occluded Person Re-Identification with Single-scale Global Representations. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 11855–11864. https://doi.org/10.1109/ICCV48922.2021.01166

Yang, Y., Liu, X., Ye, Q., & Tao, D. (2018). Ensemble learning-based person re-identification with multiple feature representations. *Complexity*, *2018*, 5940181. https://doi.org/10.1155/2018/5940181

Zahra, A., Perwaiz, N., Shahzad, M., & Fraz, M. M. (2022). *Person Re-identification: A Retrospective on Domain Specific Open Challenges and Future Trends*. http://arxiv.org/abs/2202.13121

Zhang, T., Xie, L., Wei, L., Zhang, Y., Li, B., & Tian, Q. (2019). *Single Camera Training for Person Re-identification*. http://arxiv.org/abs/1909.10848

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

Zhou, K., Yang, Y., Cavallaro, A., & Xiang, T. (2019). *Omni-Scale Feature Learning for Person Re-Identification*. http://arxiv.org/abs/1905.00953

Zhuo, J., Chen, Z., Lai, J., & Wang, G. (2018). *Occluded Person Re-identification*. http://arxiv.org/abs/1804.02792