Proceedings of CONF-MLA 2024 Workshop: Neural Computing and Applications
DOI: 10.54254/2755-2721/105/2024.TJ18113

Analysis of Different Methods in Pedestrian Re-identification

Kun Li

Information Science and Technology, Xiamen University, Xiamen, China

leovewc@gmail.com

Abstract. This study focuses on the field of pedestrian re-identification (ReID), aiming to enhance the accuracy and efficiency of individual recognition through advanced deep learning models. The article first introduces two deep learning models: Omni-Scale Feature Learning for Person Re-Identification (OSNet) and Multi-Scale Interaction Network (MSINet). OSNet employs Depthwise Separable Convolutions (DSC) and Omni-Scale Residual Blocks to improve the ability to learn features across different scales. MSINet, on the other hand, utilizes neural architecture search technology to design a lightweight network architecture, enhancing feature discrimination and flexible utilization through Twins Contrastive Mechanism (TCM) and Multi-Scale Interaction (MSI). Additionally, a Spatial Alignment Module (SAM) is proposed to enhance the consistency of images under different viewpoints or conditions. The experimental section selects two widely used pedestrian re-identification datasets, Market1501 and MSMT17, for evaluation, and the results show that MSINet outperforms existing methods in terms of accuracy and stability. The article concludes by summarizing the advantages of OSNet and MSINet in multi-scale feature learning and points out their application limitations.

Keywords: Person re-ID,Comparison, MSINet, OSNet.

1. Introduction

With the rapid development of intelligent surveillance systems, Person Re-Identification (ReID) technology is playing an increasingly important role in public safety, intelligent transportation, retail analysis, and other fields [1,2]. The goal of ReID is to identify the same pedestrian across image sequences captured by different cameras, which is crucial for tracking specific individuals in multi-camera surveillance networks. However, due to factors such as viewpoint differences, lighting changes, and occlusion interference in surveillance environments, ReID has become an extremely challenging task [3].

Early ReID methods mainly relied on manual feature extraction, such as SIFT and SURF, which had limitations in feature description capabilities and struggled to cope with pedestrian recognition issues in complex scenes. With the development of deep learning technology, methods based on Convolutional Neural Networks (CNNs) began to be widely applied to ReID tasks, enabling automatic learning of deep image features and significantly improving recognition performance [4,5,6].

In recent years, deep learning methods have made significant progress in the field of ReID. On one hand, researchers have been committed to designing more efficient network architectures to extract more discriminative features [7,8,9]. For example, OSNet enhances the model's ability to learn features across different scales through depthwise separable convolutions and omni-scale residual blocks. On the other hand, technologies such as multi-scale feature fusion, attention mechanisms, and adversarial training

DOI: 10.54254/2755-2721/105/2024.TJ18113

have been introduced into ReID to enhance the model's generalization and robustness. Furthermore, the application of Neural Architecture Search (NAS) technology has made network design more automated and optimized [3]. MSINet is a lightweight network architecture designed using NAS technology, which improves feature discrimination and flexibility through a Twins Contrastive Mechanism and multi-scale interaction mechanism [4].

This article aims to provide an in-depth analysis of the current research progress in the field of ReID and introduce two advanced deep learning models: OSNet and MSINet. These two models are innovative in terms of multi-scale feature learning, feature discrimination enhancement, and lightweight design. The article first introduces the network architectures and key technologies of OSNet and MSINet, and then verifies the effectiveness and superiority of the MSINet model through experiments on the widely used Market1501 and MSMT17 pedestrian re-identification datasets.

2. Method

Two prominent deep learning models in the field of person re-identification are OSNet and MSINet. OSNet enhances feature learning across scales with Depthwise Separable Convolutions and Omni-Scale Residual Blocks, capturing detailed and contextual information essential for identifying individuals in diverse settings. MSINet, designed through Neural Architecture Search, is a lightweight network that uses Twins Contrastive Mechanism and Multi-Scale Interaction to improve feature discrimination and adaptability. The Spatial Alignment Module in MSINet helps maintain recognition consistency despite variations in viewpoints or environmental conditions.

2.1. Omni-scale feature learning (OSNet) model

OSNet is an innovative CNN architecture specifically designed for omni-scale feature representation learning. Its foundational building block is made up of several convolutional streams, each with distinct receptive field sizes (refer to Figure 1).

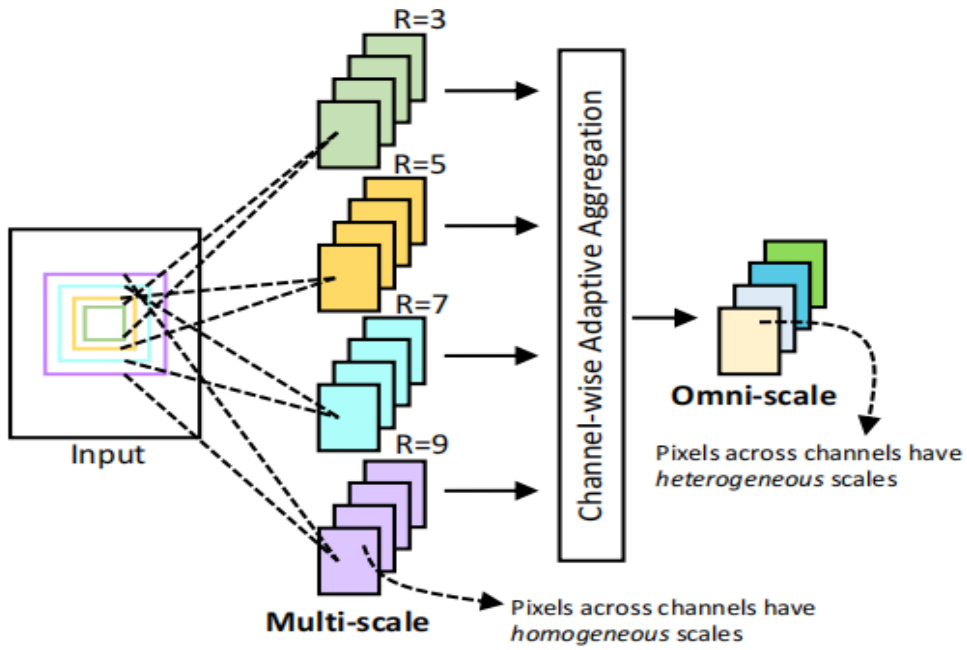

Figure 1. A schematic diagram of the basic building blocks proposed by OSNet [3]. The input image is divided into multiple scales through a multi-scale processing path, and the channels at each scale capture different features. Features are then fused through a channel adaptive aggregation module

2.1.1. Depthwise separable convolutions (DSC) Adopt the DSC to divided a standard convolution into two layers: pointwise convolution and depthwise convolution. A standard convolution is represented by

DOI: 10.54254/2755-2721/105/2024.TJ18113

a 4D tensor $w \in R^{k \times k \times c \times c'}$, where k is the kernel size, c is the depth of the input channels, and c' is the depth of the output channels. The goal is to learn the spatial-channel correlations on the input tensor $x \in R^{h \times w \times c}$, where h represents the height and w represents the width. The convolution operation can be expressed as $x' = \phi(w * x)$, where ϕ is a non-linear mapping (ReLU) and $*$ represents the convolution operation. Biases are omitted for simplicity. Figure 2(a) illustrates the practical implementation of a standard 3x3 convolution layer.

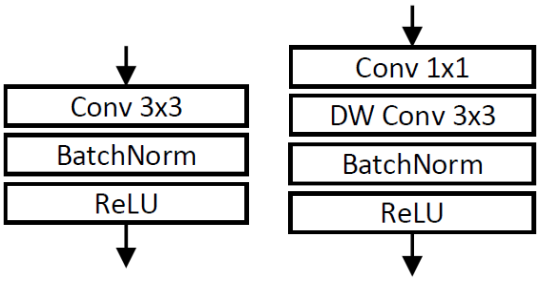

Figure 2. (a) 3 × 3 convolution. (b) 3 × 3 convolution. DW: Depth-Wise. (Photo/Picture credit : Original)

By factorizing w into $v \circ u$, where $u \in R^{1 \times 1 \times c \times c'}$ is a pointwise convolution kernel densely connected along the channel dimension, and $v \in R^{k \times k \times 1 \times c'}$ is a depthwise convolution kernel, the learning of spatial-channel correlations is decomposed. This results in $x = \phi((v \circ u) * x)$, as illustrated in Figure 2(b). This implementation achieves more effective omni-scale feature learning.

2.1.2. Omni-scale residual block The core of the architecture lies in the residual bottleneck, which is equipped with a Lite 3×3 (Figure 3(a)). Given an input x, the bottleneck is structured to acquire a residual through a mapping function F.

$$y = x + \tilde{x}, \quad \text{s.t.} \quad \tilde{x} = F(x), \tag{1}$$

where F represents a Lite 3x3 layer that learns single-scale features (scale=3).To achieve omni-scale feature learning, they introduce a new dimension, exponent t, which represents the scale of the feature. For F^t, where $t > 1$, they stack t Lite 3x3 layers, resulting in a receptive field of size $(2t + 1) \times (2t + 1)$. The residual to be learned, $\tilde{x}$, is the sum of incremental scales of representations up to T:

$$\tilde{x} = \sum_{t=1}^{T} F^t(x), \ \text{s.t.} \ T \geq 1. \tag{2}$$

When $T = 1$, Eq. 2 reduces to Eq. 1. In the paper, the bottleneck is set to $T = 4$ (i.e., the largest receptive field is 9×9), as shown in Figure 3 (b). This shortcut linkage ensures retention of smaller-scale features from the current layer into subsequent layers, supporting comprehensive spatial scale representation in the final features.

DOI: 10.54254/2755-2721/105/2024.TJ18113

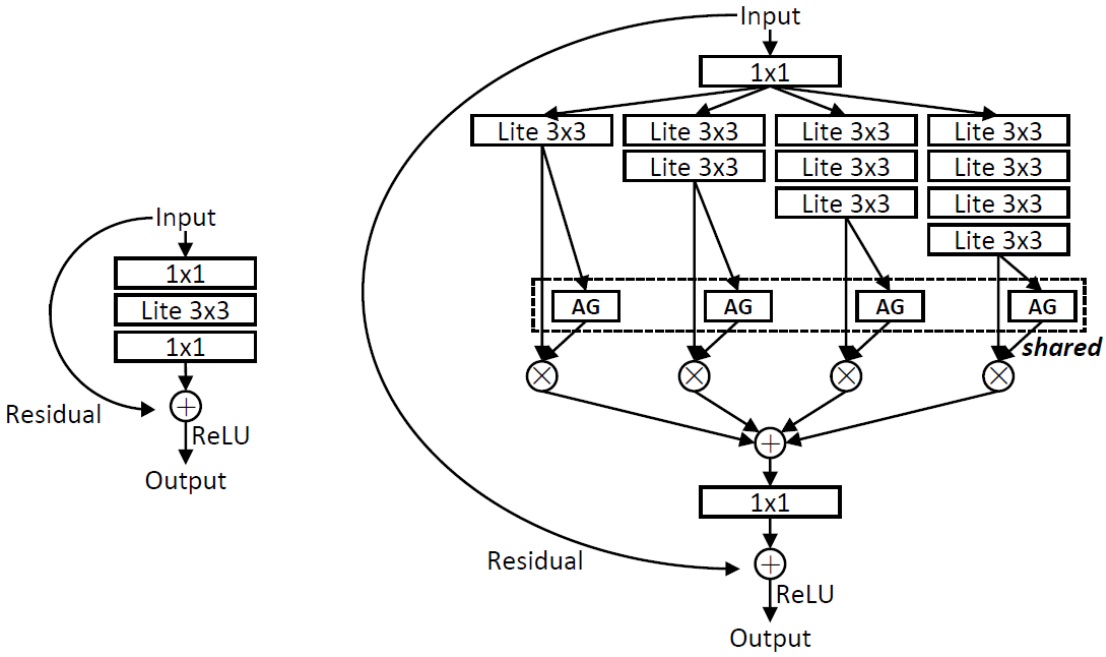

Figure 3. (a) Baseline bottleneck. (b) Proposed bottleneck. AG: Aggregation Gate. The first/last 1 × 1 layers are used to reduce/restore feature dimension [3]

2.1.3. Unified aggregation gate (UAG) To enable omni-scale feature learning, they suggested a dynamic combination of outputs from multiple streams, adjusting weights across scales according to the input image, instead of setting them statically post-training. In particular, dynamic fusion is achieved using a novel Aggregation Gate (AG), which is a learnable neural network. Defining x^t denote $F^t(x)$, the omni-scale residual $\tilde{x}$ is obtained by Equation (3):

$$\tilde{x} = \sum_{t=1}^{T} \alpha_t x^t = \sum_{t=1}^{T} G(x^t) \odot x^t, \text{ s.t. } x^t \triangleq F^t(x) \tag{3}$$

where α_t represents a vector covering the full channel dimension of x^t, with $\odot$ indicating the Hadamard product operation. The sub-network G is configured to produce outputs compressed via a sigmoid function. Specifically, the sub-network G includes a global average pooling layer, succeeded by two fully connected (FC) layers. Within this architecture, the AG is applied uniformly across all feature streams in the corresponding omni-scale residual block (see the dashed box in Figure 3(b)). This design parallels the parameter-sharing concept of convolutional filters in CNNs, yielding multiple benefits. Suppose the network is supervised by a differentiable loss function L, and the gradient $\frac{\partial L}{\partial x'}$ can be computed which is:

$$\frac{\partial \mathcal{L}}{\partial G} = \frac{\partial \mathcal{L}}{\partial \tilde{x}} \frac{\partial \tilde{x}}{\partial G} = \frac{\partial \mathcal{L}}{\partial \tilde{x}} \left(\sum_{t=1}^{T} x^t \right) \tag{4}$$

2.2. Multi-scale interaction network (MSINet)

MSINet leverages Neural Architecture Search (NAS) technology to create a lightweight and efficient network architecture tailored for re-identification tasks. The authors introduce a Twins Contrastive Mechanism (TCM) and Multi-Scale Interaction (MSI), which enhance feature differentiation and enable more flexible use of features. As illustrated in Figure 4, the network primarily consists of MSI cells and down-sampling blocks, making its structure broadly similar to that of OSNet.

DOI: 10.54254/2755-2721/105/2024.TJ18113

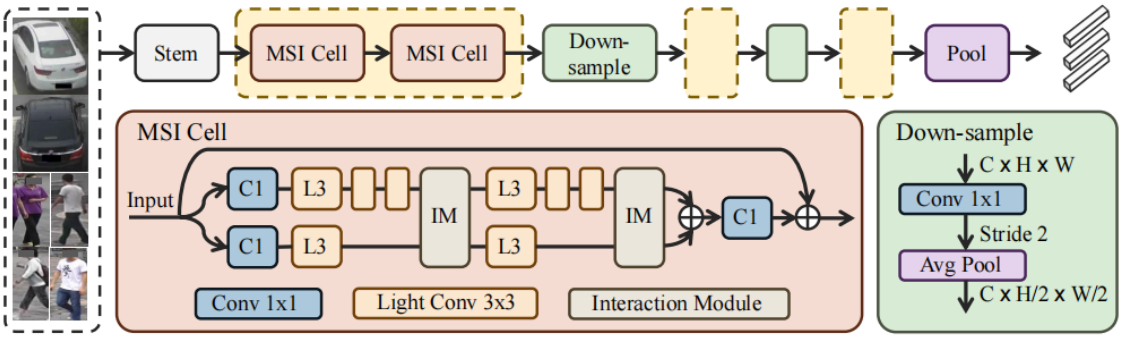

Figure 4. The architectural design of MSINet allows for both person and vehicle inputs. Within each cell, the input is split into two branches, each having distinct receptive field scales. An interaction module facilitates information exchange between the branches, while an architecture search algorithm autonomously determines the optimal interaction for each cell. [4]

2.2.1. Twins contrastive mechanism (TCM) In typical NAS approaches, training and validation datasets have matching categories, with loss calculated through linear classification. This is suitable for conventional image classification tasks, as the categories remain consistent across training and validation phases. However, this approach is not suitable for ReID tasks because ReID is an open set task and the categories in the training and validation sets are usually different. Accordingly, They propose TCM for NAS ReID training. The TCM unbinds the categories of the training and validation sets, allowing for dynamically adjustable overlap ratio. And two auxiliary memories are introduced for training and validation data to store embedded features, helping in calculating contrastive classification loss and stabilizing model and architecture updates:

$$\mathcal{L}_{tr}^{c\ell s} = -\log \frac{\exp\left(f \cdot c_{tr}^{j}/\tau\right)}{\sum_{n=0}^{N_{tr}^{c}} \exp\left(f \cdot c_{tr}^{n}/\tau\right)} \tag{5}$$

Where C_{tr} and C_{val} refer to embedded features for training and validation data, f represents an embedded feature with category label j, c_{tr}^{n} signifies the memorized feature of category n, N_{tr}^{c} is the total categories in the training dataset, and τis the temperature parameter setting as 0.05.

2.2.2. Multi-scale interaction (MSI) In previous ReID works, the utilization of the local perspective and multi-scale features is monotonous and restrained. Therefore, MSI is proposed to establish for ReID to enhance the interaction between multi-scale features for improved feature extraction. A novel MSI search space is designed to identify the most suitable interaction operations between features of different scales, including none, feature exchange, channel gate, and cross attention. Thereby different network layers could collaborate effectively by dynamically adjusting the interaction across shallow and deep layers, ensuring better utilization of local and global features.

2.2.3. Spatial alignment module (SAM) SAM is used to improve consistency in attention for images taken from different viewpoints or under different conditions such as poses, illumination and occlusion. Through calculating the spatial correlation between feature maps within a mini-batch:

$$a(i,j) = \max_{\dim=1} A\,(i,j) \tag{6}$$

And aligns the spatial positions across images from various sources, ensuring that the model consistently focuses on important areas during recognition and enhancing its robustness when varying viewpoints and lighting conditions.

DOI: 10.54254/2755-2721/105/2024.TJ18113

In comparison, OSNet is advantageous in applications requiring feature learning across multiple scales. Its lightweight design: OSNet use separable convolutions so that the building blocks and subsequently the entire network are lightweight. It omni-scale capability, and efficient use of the AG contribute to its effectiveness. However, its only focus on single-instance fine-grained recognition, which can limit its generalizability. On the other hand, MSINet particularly suitable for scenarios with limited resources. With the MSI and SAM, MSINet is capable of capturing fine-grained features and achieving high recognition accuracy while maintaining low computational complexity, making it suitable for general object detection and recognition. However, MSINet is heavily rely on feature interaction and higher training data requirements are notable drawbacks for applications needing fast and efficient deployment.

3. Result and discussion

3.1. Dataset

In this experiment, we selected two widely used pedestrian re-identification datasets: Market1501 and MSMT17. The Market1501 dataset consists of 12,936 training, 3,368 query, and 15,913 test images featuring 751 identities, capturing variations in pedestrian appearances across different camera views. The MSMT17 dataset is more extensive and challenging, comprising 32,248 training, 11,659 query, and 82,161 test images with 1,041 identities, and includes more variations in lighting, backgrounds, and camera angles, emphasizing the need for robust cross-domain performance.

For evaluation, we utilized standard metrics: cumulative matching characteristics (CMC) and mean average precision (mAP), focusing on the Rank-1 accuracy and mAP to assess model precision and reliability.

3.2. Experimental settings

For the training of Market1501 and MSMT17 datasets, we followed the standard classification training paradigm, and each pedestrian identity was regarded as a unique class. The loss functions used included cross entropy loss (with label smoothing) and triplet loss. For the model trained from scratch, the learning rate was initially 0.065 and decreased by 0.1 at 150, 225, and 300 epochs. Data augmentation included random flipping, random cropping, and random erasing. For the pre-trained model, we used the AMSGrad optimizer and froze the pre-trained backbone network in the first 10 epochs, training only the randomly initialized classifier.

3.3. Experimental results

When comparing the performance of QN-MSINet and OSNet across two datasets (Market1501 and MSMT17), we can observe that MSINet consistently demonstrates impressive capabilities in pedestrian re-identification tasks.

3.3.1. Analysis of the market1501 Dataset In the Market1501 dataset, MSINet achieves a Rank-1 accuracy of 94.8%, which is equal to that of OSNet, and it also excels in Rank-5 and Rank-10 accuracy, reaching 98.2% and 99.0%, respectively (Table 1). These metrics indicate that MSINet is capable of accurately identifying most queried pedestrians and effectively distinguishing them from others. Moreover, the mean Average Precision (mAP) for MSINet is 87.9%, significantly higher than OSNet’s 86.0%. mAP is a comprehensive metric that effectively reflects a model's overall performance in classification tasks. Therefore, MSINet's performance on this dataset is not only stable but also superior to that of OSNet.

This improvement in performance can be attributed to MSINet's architectural and training optimizations. MSINet likely employs advanced feature extraction techniques or a deeper network architecture, enabling it to capture the features of pedestrian images more sensitively in complex environments, thereby achieving good recognition results across various scenarios. In contrast, while

DOI: 10.54254/2755-2721/105/2024.TJ18113

OSNet performs adequately, it appears to struggle with learning and generalizing certain detailed features.

3.3.2. Analysis of the MSMT17 Dataset On the MSMT17 dataset, MSINet shows even greater superiority, with a Rank-1 accuracy of 94.9%, significantly outpacing OSNet's 78.7% (Table 2). OSNet's performance drops markedly on this dataset, indicating that the network faces considerable challenges in handling the features present in the MSMT17 dataset. The Rank-5 and Rank-10 accuracies for MSINet are 97.9% and 98.8%, while OSNet's figures are only 83.6% and 89.7%, clearly reflecting a disadvantage. In terms of mAP, MSINet scores 85.4%, compared to OSNet's mere 52.9%. This significant gap illustrates that MSINet can maintain high recognition performance even in a more complex and diverse data environment, while OSNet shows insufficient adaptability to the dataset's characteristics.

This phenomenon may be linked to the unique features of the dataset. MSMT17 contains a broader variety of samples and environments that could potentially impact OSNet's performance. In contrast, MSINet's design enhancements enable it to optimize its performance in terms of diversity and complexity. This underscores the necessity for deep learning algorithms to consider the different features of various datasets when constructing pedestrian re-identification models, effectively boosting the model's generalization capacity and accuracy.

In summary, MSINet outperforms OSNet in both datasets, reflecting its advantages in feature extraction and algorithm design while demonstrating adaptability to environmental changes. Strong metrics such as Rank-1, Rank-5, Rank-10, and mAP can provide reliable support for practical applications, particularly in fields like security surveillance, intelligent transportation, and business analytics. Future optimizations in model design, combined with reinforcement learning techniques, could further enhance accuracy and efficiency, providing robust tools for pedestrian re-identification.

Table 1. Results on the market1501

Method	Rank-1	Rank-5	Rank-10	mAP
MSINet	94.8%	98.2%	99.0%	87.9%
OSNet	94.8%	98.1%	98.7%	86.0%

Table 2. Results on the msmt17

Method	Rank-1	Rank-5	Rank-10	mAP
MSINet	94.9%	97.9%	98.8%	85.4%
OSNet	78.7%	83.6%	89.7%	52.9%

4. Conclusion

This paper discusses two advanced deep learning models in the field of ReID: OSNet and MSINet. The OSNet model, through the design of depthwise separable convolutions and omni-scale residual blocks, achieved effective learning of features across different scales. This design not only enhanced the expressive power of the features but also dynamically integrated features of various scales through the UAG, bolstering the model's ability to recognize pedestrian identities. Moreover, the lightweight design of OSNet allows it to operate efficiently on devices with limited resources, which is significant for practical deployment. The MSINet model, leveraging neural architecture search technology, crafted a lightweight and efficient network architecture. This model enhanced its interaction and fusion capabilities with features of different scales through the TCM and MSI mechanisms, thereby improving feature discrimination. The introduction of the SAM further strengthened the model's recognition capabilities under images with varying perspectives and lighting conditions. Through experiments on the Market1501 and MSMT17 datasets, we verified that the MSINet model outperforms existing methods in terms of recognition accuracy and stability. These results indicate that MSINet has a

DOI: 10.54254/2755-2721/105/2024.TJ18113

significant performance advantage when dealing with pedestrian re-identification tasks in complex environments.

Nevertheless, this paper also noted some limitations of OSNet and MSINet in practical applications. For instance, OSNet primarily focuses on fine-grained recognition of single instances, which restricts its generalization capabilities in broader scenarios. While MSINet excels in feature interaction, its high dependency on training data may affect the model's performance under data-limited conditions.

Future research can explore several avenues: firstly, how to further optimize the model structure to enhance its generalization capabilities under different camera and environmental conditions; secondly, how to balance the model's accuracy and computational complexity so that it can operate efficiently on resource-constrained devices; thirdly, how to integrate privacy protection requirements to develop ReID technology that can effectively recognize pedestrian identities while safeguarding personal privacy; and finally, how to utilize multimodal data, such as video and depth information, to further enhance ReID performance.

References

[1] Zheng, L., Wang, S., & Wang, J. (2016). Person Re-identification: A Benchmark. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(1), 1-20. doi:10.1109/TPAMI.2016.2593860.

[2] Wu, Y., & Zheng, Y. (2019). A Comprehensive Review on Person Re-identification: From Feature Representation to Matching Strategy. ACM Computing Surveys, 54(3), 1-35. doi:10.1145/3299900.

[3] Zhou, K., Yang, Y., Cavallaro, A., & Xiang, T. (2019). Omni-Scale Feature Learning for Person Re-Identification. arXiv preprint arXiv:1905.00953

[4] Gu, J., Wang, K., Luo, H., Chen, C., Jiang, W., Fang, Y., Zhang, S., You, Y., & Zhao, J. (2023). MSINet: Twins Contrastive Search of Multi-Scale Interaction for Object ReID. arXiv preprint arXiv:2303.07065

[5] Sun, Y., Zheng, L., & Wang, S. (2018). Beyondpart Models for Person Re-Identification. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 1-15. doi:10.1109/CVPR.2018.00456.

[6] Li, W., Zhang, X., & Wang, J. (2018). Hetero-Feature Fusion for Person Re-Identification. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 1-10. doi:10.1109/CVPR.2018.00368.

[7] Chen, Y., & Wang, G. (2019). Relation Network for Person Re-identification. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 1-10. doi:10.1109/CVPR.2019.00289.

[8] Zhao, R., Zhang, L., & Wang, H. (2017). Person Re-identification by Deep Learning: A Review. Journal of Multimedia, 12(6), 493-505. doi:10.18178/jmm.2017.12.6.410.

[9] Huang, K., & Wang, Y. (2020). Person Re-identification: A Survey of the State of the art. IEEE Transactions on Image Processing, 29, 8846-8863. doi:10.1109/TIP.2020.2993405.