Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/86/20241592

Person Re-ID: Tackling occlusion, labeled data, and privacy

Haoyu Chen

Aberdeen School of Data Science and Artificial Intelligence, South China Normal University, Nanhai Campus, Foshan, Guangdong, China

20213802048@m.scnu.edu.cn

Abstract. Person re-identification (ReID) has become a critical component in various security and surveillance systems, necessitating accurate and robust identification of individuals across different camera views. The motivation for enhancing ReID systems stems from the need to improve public safety, prevent crime, and support forensic investigations. Despite significant advances, ReID faces several key challenges that hinder its deployment in real-world applications. This review paper addresses key challenges in person re-identification (ReID): overcoming occlusion, reducing dependence on labeled data, and addressing privacy concerns. Various techniques have been explored to tackle these issues effectively. For occlusion, part-based models, LCNNs, attention mechanisms, and adversarial training GANs have demonstrated robustness in capturing person features across diverse and complex environments. To mitigate dependence on labeled data, semi-supervised and unsupervised learning approaches leverage unlabeled data and self-learning capabilities. Additionally, multimodal fusion utilizes complementary information from different sources, significantly enhancing model generalization and performance. Privacy concerns are addressed through federated learning, which fosters collaboration across devices while safeguarding individual data privacy. In summary, this paper highlights diverse technological applications in ReID, offering valuable insights and guidance for future research and practical implementations in the field.

Keywords: ReID, occlusion, dependence on labeled data, privacy

1. Introduction

Person re-identification (ReID) is pivotal in computer vision, applied widely across healthcare, transportation, surveillance, security, and forensics. ReID addresses the matching problem—identifying individuals from large image databases—and the identification problem—tracking individuals across multiple camera streams. Despite advancements in deep learning and pattern recognition, deploying ReID systems faces persistent challenges. This paper reviews state-of-the-art methods tackling three critical ReID challenges: occlusion, reliance on labeled data, and privacy concerns [1].

Occlusion is a significant issue in person re-identification (ReID), where parts of an individual's body are obstructed, resulting in incomplete visual information. This is common in real-world environments like crowded public spaces and degrades ReID performance due to compromised feature extraction. Advanced techniques such as part-based models, attention mechanisms, adversarial training GANs, and LCNNs have been developed to address this problem. High-quality labeled data is crucial for training accurate ReID models, which typically rely on supervised learning. However, obtaining such data is labor-intensive and costly, limiting the scalability and adaptability of ReID systems. To reduce dependence on labeled data, techniques like unsupervised learning, semi-supervised learning, and multimodal fusion have been developed.

The privacy issue in ReID arises from handling large volumes of sensitive personal data, posing risks under centralized training methods. Ensuring data privacy is crucial, especially with stringent regulations like the California Consumer Privacy Act (CCPA). Advanced methods such as federated learning have been developed to mitigate these risks.

This comprehensive paper thoroughly reviews and synthesizes the cutting-edge methodologies aimed at tackling the occlusion challenge, the issue of dependence on labeled data, and the privacy concerns that are inherent in the field of ReID. It delves into the intricacies of each problem, analyzing the latest research breakthroughs and critically assessing their respective strengths and limitations. The objective is to offer a detailed and holistic understanding of the advancements made in these critical areas of ReID, thereby facilitating the identification of promising research directions for the scientific community to explore in the future.

The paper not only examines the technical aspects of each approach but also considers the practical implications and potential impact on the broader application of ReID technology. It discusses how current methods address the challenges posed by occlusion, which often leads to incomplete or distorted visual data, affecting the accuracy of identification. Additionally, it explores strategies to mitigate the heavy reliance on extensive labeled datasets, a significant barrier to the scalability and generalizability of ReID systems. Furthermore, the paper addresses the urgent need to protect individual privacy in the face of increasingly stringent data protection regulations (Figure 1).

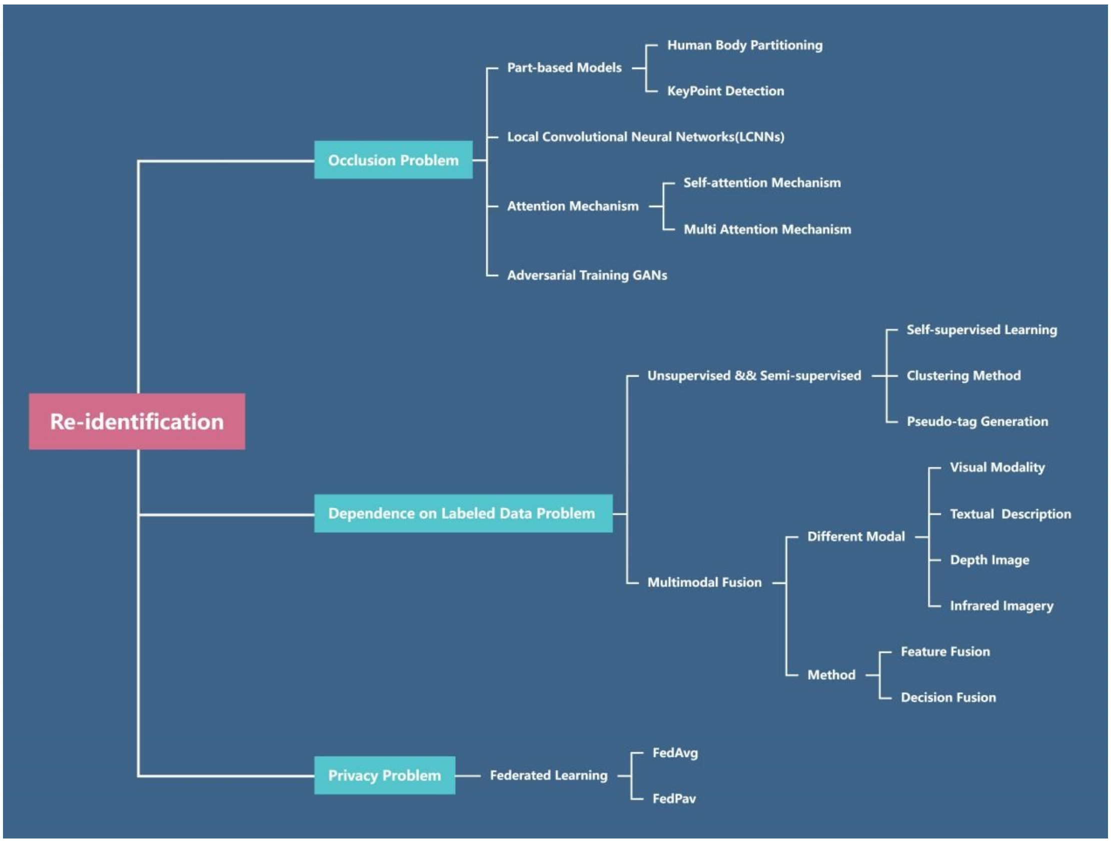

Figure 1. Essay Architecture (Photo/Picture credit : Original)

2. Occlusion Problem

2.1. Part-based Model

To address the occlusion problem in ReID task, people can optimize for the first step of local feature extraction. The Part-based Model was first proposed in 2016. Part-based models separate the human body into discrete components and then take individual parts' attributes. This approach leverages the idea that even if some parts of the body are occluded, the remaining visible parts can still provide enough discriminative information to identify individuals accurately. There are two main approaches to this advanced technology: Human Body Partitioning and Keypoint Detection.

The principle of Human Body Partitioning is to divide an image of a person into different areas, usually corresponding to different body parts, such as trunk, arms, head, legs and so on. This segmentation method uses advanced image processing techniques and deep learning techniques to perform, and then classifies each pixel in the image into a predefined category. The features are extracted separately in each segment and aggregated to form a comprehensive feature vector, so as to deal with the problem of partial occlusion. Keypoint detection, as a sub-task of pose estimation, aims to locate the keypoints on a person's body, usually corresponding to anatomical positions such as the head, shoulders, elbows, wrists, knees, ankles, etc. It forms a skeletal representation of a person based on the positions of joints, thereby inferring the person's position and posture. After obtaining a high-level abstraction of human posture, the areas around the keypoints that are not obscured are extracted to capture important information, thus solving the occlusion problem..

Human Body Partitioning focuses on individual body parts, while Keypoint Detection can estimate occluded key points based on visible key points, both of which have good robustness to occlusion. Additionally, the former provides a more detailed and localized representation, while the latter provides a high-level abstraction of the human body with a more compact representation. However, Keypoint Detection has higher computational efficiency and requires less annotated data, and is suitable for real-time applications and scenarios with limited computing resources, while Human Body Partitioning is more suitable for scenarios where specific body part information is required.

The table 1 provides a specific comparison of the algorithms of several mainstream Part-Based models.

Table 1. Human Body Partitioning Algorithm

Algorithm	FCNs	Mask R-CNN	DeepLab
Principle	FCNs use only convolutional layers, allowing them to generate spatial maps of predictions that correspond directly to the input image's resolution.	Mask R-CNN is adding a branch for predicting segmentation masks on each detected object.	DeepLab captures multi-scale contextual data via atrous (dilated) convolutions without sacrificing spatial resolution.
Architecture	1.Encoder-Decoder Architecture: Encoder: convolutional layers 2.Skip Connections	1.Backbone 2.RPN 3. Rol Aign 4.Segmentation branch	1.Backbone 2 Atrous Convolutions 3 ASPP
Key Components	1.Convolutional layers 2.Transposed convolutions 3.Skip connetections	1.Backbone Network 2.RPN 3.Rol Align 4.Segmentation branch	1.Atriys Convolutions 2.ASPP 3.CRFs(optional)
Advantage	1.Simplicity 2.Dense prediction tasks(well) 3 Flexibility in input size 4 Efficient Computation	1.high mask accuracy 2.Accurate instance segmentation	1.Captures multi-scale context 2.Accurate boundary delineation
Disadvantage	1.Fixed receptive field size 2.struggle with multi-scale context	1.Computationally intensive 2.Requires accurate RoI proposals	1.complex architecture 2.Computationally expensive

In 2023, Vladimir Somers et al. [2] proposed a body part-based representation learning approach to address occlusion issues in person re-identification. This method focuses on extracting features from individual body parts, ensuring that the model can still identify individuals even when some parts are occluded. By learning robust representations for each body part separately, the model enhances its ability to handle occlusions effectively. However, the model's performance may be sensitive to the choice of body parts and their respective weights in the feature representation [2].

Building on this concept, Gang Yan et al. [3] introduced an innovative technique for addressing occlusions through part-based representation enhancement. Similar to Somers et al., this approach systematically extracts and refines features from individual body parts to maintain model effectiveness despite occlusions. By focusing on distinct body regions and enhancing their representations, Yan et al. significantly boosted the robustness and accuracy of ReID systems. Nonetheless, their method may encounter limitations in scenarios with heavy occlusions where key body parts are entirely obscured [3].

Further advancing these methodologies, Wang et al. [4] presented Quality-Aware Part Models (QPM), a pioneering approach notable for its robustness against occlusions. Unlike the previous approaches, QPM eschews reliance on external tools for visibility inference, opting for an end-to-end learning strategy. This strategy involves the joint learning of part features and the prediction of part quality scores, which help mitigate the impact of occluded regions. Additionally, QPM includes an adaptive global feature extraction technique that focuses on consistently non-occluded regions, ensuring semantic consistency. However, this reliance on part quality metrics could introduce additional computational complexity and may require fine-tuning for different datasets or camera conditions [4].

2.2. Local Convolutional Neural Networks

To address the occlusion problem in ReID tasks, people can optimize for the second step of local feature extraction process. Local Convolutional Neural Networks solve the problem by focusing on the local areas of the image and enhancing the feature extraction of the non-occluded parts even if some body parts are partially occluded, thereby improving accuracy and solving the occlusion problem.

Local Convolutional Neural Networks (LCNNs) do not target the entire image, but are designed to process local images or segmented image blocks. LCNNs consist of multiple convolution layers and pooling layers. Each of these convolution layers applies a set of learned filters to the input patches, generating feature maps that highlight specific patterns, thereby extracting local features and maintaining spatial hierarchy. Using techniques like max pooling and average pooling, the pooling layer is utilized to minimize the spatial dimension of the feature map and capture the most significant characteristics, thus reducing the computational complexity. First, LCNNs use various strategies such as the above key point detection to divide the input image into multiple overlapping or non-overlapping small blocks, and then LCNNs will normalize each patch to make it have a consistent mean and standard deviation. Then, as each patch passes through a convolutional neural network composed of multiple convolution layers and pooling layers, separate local feature extraction is carried out, making the network more focused on local region extraction, enhancing the ability to capture fine-grained details, and thus reducing the impact of occlusion on accuracy.

Compared to traditional CNNs, LCNNs focus more on local areas, allowing them to capture more details and be more robust to occlusion problems. Secondly, LCNNs has better flexibility and scalability, which can expand the image size by adjusting the number of patches.In terms of computational efficiency, LCNNs are more efficient when using overlapping patches due to localized processing.

2.3. Attention Mechanism

When facing the occlusion problem in ReID task, people can optimize in the step of feature fusion, and the Attention Mechanism is one of the mainstream optimization techniques. The Attention Mechanism dynamically focuses on the part with the most information in the image and simultaneously reduces the weight of the blocked area, thereby reducing the influence of the occlusion problem on ReID. The Attention Mechanism will be presented in detail based on this classification, which separates it into the Self-attention Mechanism and the Multi-attention Mechanism.

2.3.1. Self-attention Mechanism Self-attention Mechanism is a way to enhance the representation of a model by learning global dependencies within a sequence. The Self-attention Mechanism allows each element or pixel in the sequence to focus on all the rest, dynamically assign different levels of importance to various parts of the image, and then generate a weighted sum of all elements to ensure that the image stands out even when it is partially obscured.

After the input image is processed by CNN, a set of feature maps can be extracted. the Self-attention module calculates three vectors for each pixel or feature point, namely, the query, the key and the value. The query represents the pixel for which we are computing the attention, the key represents the elements against which the query is compared, and the value represents the elements containing the actual information. The attention score between a query and a key is calculated using a compatibility function, such as the dot product. This attention score indicates how much attention the query should pay to the corresponding key, so that the entire model pays different attention to different parts of the image. Finally, all the parts are weighted to get the sum, which is the value vectors. Here is the formula for Scaled Dot-Product Attention, by applying the softmax function to get an extended score to obtain attention weight.

$$Attention(Q,K,V) = \mathrm{softmax}(\frac{\mathrm{QK}^T}{\sqrt{d_k}})V \tag{1}$$

In 2021, Sun Mengzhe et al. [5] presented a sophisticated person ReID network that combines channel attention and self-attention mechanisms to enhance feature representation. The self-attention mechanism is particularly noteworthy, because it permits the model to dynamically concentrate on the areas of the input image that are most pertinent, capturing intricate dependencies within the data. This approach allows the network to prioritize critical features while minimizing the impact of less informative regions. By integrating self-attention with channel attention, the method achieves a balanced and comprehensive feature extraction process, leading to superior performance on multiple benchmark ReID datasets [5].

2.3.2. Multi Attention Mechanism The Multi Attention Mechanism is an extension of the Attention Mechanism that works by assigning multiple sets of attention weights to different parts of the input image, ensuring that even if some parts are obscured, other areas of information are still highlighted. Secondly, each mechanism uses a hierarchical structure, and the attention mechanisms at different levels focus on the image features at different scales, so that a wider range of data can be obtained. Finally, by combining attention weights from multiple regions, the model can integrate contextual information and enhance its ability to distinguish between different individuals.

The Multi Attention Mechanism operates by running multiple Self-Attention operations in parallel to assign weights, which is similar to the previous method. When multiple attention heads are obtained, the outputs of all attention heads are joined together to form a single tensor. The final attention output is produced by aggregating these into a comprehensive feature, which may require additional layers or operations.

In contrast, the Self-attention Mechanism consists of a single attention mechanism that uses a set of queries, keys, and values to capture relationships in the data, making it simpler in complexity. Secondly, the Multi Attention Mechanism allows the model to perceive information from different subspaces, providing a more diversified and comprehensive understanding of data, and capturing richer information than Self-attention Mechanism. In terms of performance, Multi Attention Mechanism has more parameters and higher computational complexity, but it significantly improves the model's ability to capture all aspects of data relationships. In terms of performance, the Multi Attention Mechanism has more parameters and higher computational complexity, but it significantly improves the model's ability to capture all aspects of data relationships.

In 2022, Yang Cong et al. [6] explored the use of unsupervised domain adaptation to tackle the domain shift problem in person re-identification. They introduce a multi-scale attention mechanism that enhances feature extraction by focusing on different scales of the input images. This multi-attention approach allows the model to capture detailed and contextual information simultaneously, crucial for improving cross-domain generalization. The method effectively aligns feature distributions between source and target domains without relying on labeled target data, demonstrating significant performance gains on several benchmark datasets. However, the model's sensitivity to the choice of hyperparameters could impact its robustness and generalizability [6].

Compared with Yang Cong, in 2022, MA XIAO et al. [7] added the innovative double-stream person ReID method of multi-scale feature fusion on the basis of attention mechanism. The approach utilizes a multi-attention mechanism to focus on various aspects of the input data, gathering global contextual information as well as minute details. By integrating multi-scale feature fusion, the method effectively combines features at different levels of abstraction, enhancing the model's ability to handle diverse and complex visual patterns. This dual-stream architecture, with its emphasis on attention and feature fusion, significantly improves re-identification accurate across several challenging benchmark datasets. However, it may lead to higher computational and memory requirements, which could be a limitation for real-time applications or systems with limited resources [7].

2.4. Adversarial Training GANs

Adversarial Training GANs effectively address the occlusion problem in ReID tasks. These GANs comprise a generator and a discriminator. The generator creates synthetic images or features from random noise or conditional inputs, while the discriminator distinguishes between real and synthetic data. Both are trained adversarially: the generator aims to deceive the discriminator with realistic images, and the discriminator attempts to identify real versus generated images. Using loss functions like adversarial loss and reconstruction loss, GANs generate realistic images from occluded datasets. The generator learns to fill occluded areas, producing complete images. The discriminator then differentiates between original and synthesized images, guiding the generator to improve until the discriminator can no longer tell them apart. This process enhances the ReID model's robustness against occlusion. Additionally, the generator creates synthetic occlusions in training images, enriching the dataset and improving generalization to various occlusion scenarios.

In facing the occlusion problem in ReID tasks, Adversarial Training GANs can generate occlusion-free images and enhance feature representation, improving the model's robustness and accuracy against occlusion. Secondly, GANs can customize image generation according to the objectives of ReID tasks, adapting to different occlusion problems through image completion and feature enhancement, providing high flexibility and adaptability. Lastly, adversarial training increases overall identification accuracy by making the model learn robust features that are less susceptible to occlusions.

In 2023, Houjing Huang et al. [8] introduced an innovative approach using adversarially occluded samples. The study generates these samples through adversarial training techniques, enabling the model to learn to identify individuals despite partial occlusions. This method significantly enhances the robustness and performance of ReID systems under real-world occlusion conditions, as demonstrated by improved results on standard ReID benchmarks. However, the reliance on GANs for generating samples may lead to increased training times and the potential for mode collapse, where the GAN fails to generate diverse samples [8]. Aiming at the above problems, Xingyue Shi et al. [9] presented a novel method to enhance person re-identification performance by combining viewpoint contrastive learning with adversarial training. Viewpoint contrastive learning enables the model to learn robust features by contrasting different viewpoints of the same individual, while adversarial training generates challenging samples that improve the model's resilience to variations in appearance. The accuracy of the model under occlusion is improved through the dual method. But the complexity introduced by combining viewpoint contrastive learning with adversarial training could increase the difficulty of tuning the model's hyperparameters [9].

3. Dependence on Labeled Data Problem

3.1. Unsupervised Learning

Unsupervised learning techniques can significantly reduce dependence on labeled data in Re-ID tasks. Unlike supervised learning, which requires large labeled datasets, unsupervised learning learns feature representations directly from unlabeled datasets. Unsupervised learning encompasses two primary techniques: self-supervised learning and clustering methods.

Self-supervised learning generates pseudo-labels or creates proxy tasks from unlabeled data, leveraging inherent data structures to develop useful representations without manual labeling. The goal of this method is to create reliable, generalized feature representations that can be used straight away for jobs that come after or adjusted as needed. By enhancing model robustness and generalization in the absence of labeled data, self-supervised learning improves performance on unseen data and diverse scenarios. Pretext tasks such as contrastive learning, instance discrimination, and rotation prediction generate supervisory signals directly from data. Contrastive learning distinguishes between similar and dissimilar samples by creating pairs of augmented positive and negative samples. Instance discrimination treats each instance as a separate class, generating multiple augmented views per instance to train the model to recognize different instances. Rotation prediction rotates images at fixed angles (0°, 90°, 180°, 270°) and trains models to predict these angles, thereby learning invariant features across rotations. These techniques use methods like cross-entropy loss for effective training and classification.

Clustering methods, a form of unsupervised learning, reduce reliance on labeled data in ReID tasks by grouping similar data points based on inherent patterns. This maximizes intra-cluster similarity and minimizes inter-cluster similarity. Pseudo-labels are assigned to data points based on their clusters and iteratively refined to train a high-accuracy ReID model. Popular clustering methods include K-means, DBSCAN, and hierarchical clustering, each using different criteria for forming clusters. The table 2 below compares these algorithms in detail.

Table 2. Clustering Algorithm

Algorithm	Principle	Advantage	Disadvantage	Application
K-means	divide the dataset into k clusters, with the nearest mean for each data point indicating which cluster it belongs to.	1.Simple 2.Efficient 3.Spherical clusters	1.Require predefined number of clusters 2.sensitive to initialization	Initial exploratory analysis, when clusters are known in advance
DBSCAN	combines densely populated points and identifies sites that are isolated in low-density areas as outliers.	1.No need to specify number of clusters 2.Can handles noise 3.Can find clusters of arbitrary shapes	1.Parameter Sensitivity about ϵ and MinPts 2.Poor Scalability	Non-uniform data distributions, geographical data, noisy datasets
Hierarchical Clustering	creates a dendrogram (tree) of clusters by repeatedly joining or dividing preexisting clusters according to a distance metric.	1.No Need for pre-specified number of clusters 2.A complete hierarchy of clusters	1.Computational complexity(large dataset) 2.Irreversible merges/splits	Non-uniform data distributions, geographical data, noisy datasets

In 2022, Xuefeng Tao et al. [10] proposed an unsupervised learning framework focusing on minimizing domain gaps to enhance person re-identification (ReID) across varying camera settings. Their approach eliminates the need for labeled target domain data, offering a robust solution for domain transfer challenges in ReID. However, it may encounter difficulties with substantial domain gaps or highly dissimilar feature spaces between source and target domains [10].

Building on this, in 2024, Jianfeng Weng et al. [11] proposed a federated learning method for unsupervised ReID, emphasizing decentralized data processing to address privacy concerns. This approach enables knowledge transfer across different datasets without centralized supervision. Nonetheless, its effectiveness could be constrained in datasets with significant diversity or pronounced domain gaps [11].

3.2. Semi-supervised Learning

Compared to unsupervised learning, semi-supervised learning provides a balanced approach to reducing the dependence of ReID tasks on labeled data. Semi-supervised learning leverages unlabeled data to gain additional insights into data distribution and structure, thereby enhancing model training with a limited amount of labeled data. Specifically, semi-supervised approach involves pseudo-labeling, where the model assigns provisional labels to unlabeled data and treats them as genuine labels during training, thereby expanding the effective labeled dataset and mitigating the reliance on annotated data in ReID tasks.

In order to conduct semi-supervised learning, a dataset comprising a greater proportion of unlabeled data and a smaller amount of labeled data is gathered. First, pseudo-labels for the unlabeled data are created by training a baseline model on the labeled data. The augmented training set, which is used iteratively to enhance the model, is created by combining the labeled and pseudo-labeled data. Consistency loss functions are employed to minimize discrepancies between predictions on original and pseudo-labeled data, refining the model's accuracy.

This approach optimizes the use of both labeled and unlabeled data, addressing data scarcity challenges. By integrating unlabeled data, it helps learn more generalized and robust feature representations, improving the generalization of the model to new data. Additionally, semi-supervised learning scales effectively to large datasets without requiring a proportional increase in labeled data, making it suitable for practical applications in scientific research and beyond.

In 2019, Ding Guodong et al. [12] introduced a semi-supervised learning approach for person ReID, focusing on leveraging feature affinity to assign pseudo labels to unlabeled data. The method effectively expands the training dataset and demonstrates superior performance under limited labeled data conditions. However, its reliance on feature affinity metrics may limit its adaptability to diverse real-world datasets [12]. Building on this theme, in 2023, Ancong Wu et al. [13] proposed a reward-based learning mechanism aimed at enhancing model generalization by utilizing unlabeled data effectively. Their approach specifically addresses challenges posed by sparse identity crossings, improving system robustness and accuracy. While innovative, this method faces challenges in accurately estimating rewards and balancing the contributions of labeled and unlabeled data sources .

3.3. Multimodal fusion

One of the key elements affecting the model's accuracy and resilience in ReID tasks is the quantity of datasets, and dependence on labeled data may affect the number and coverage of the datasets. Multimodal fusion combines data from different modalities, such as visual modality, textual descriptions, depth images, infrared imagery, etc. It creates a comprehensive representation of features by integrating rich information to enhance model accuracy and mitigate the challenges of relying on labeled data. Compared with single modality, multimodal fusion provides richer feature representation and can enhance the model's ability to adapt to new environments and improve its generalization ability by integrating various types of information. Next, we will briefly introduce different modalities and analyze two multimodal related methods.

3.3.1. Different Modals Visual modality uses RGB images captured by standard cameras to provide rich color, texture, and spatial information about individuals, such as clothing and facial features. CNNs and other technologies extract hierarchical features from these images. Given the ubiquity of RGB cameras, collecting ample visual data is easier, making this modality crucial in ReID systems (Figure 2). However, accuracy can be compromised by human factors, environmental conditions, and lighting variations, leading to blurred or partial data loss.

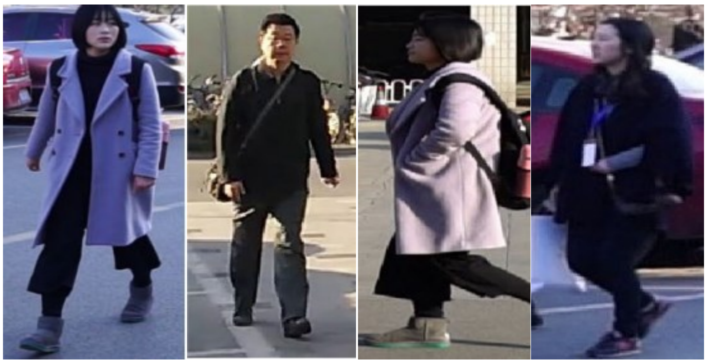

Figure 2. RGB images [7]

Textual descriptions narrate an individual's appearance, detailing clothing color, accessories, and other specifics. Natural Language Processing converts these descriptions into digital characteristic symbols, capturing details often missed by images. Text descriptions remain useful when visual data is insufficient but may introduce variability and biases due to differing linguistic expressions and observer inconsistencies. Depth images, processed by 3D-CNNs, capture spatial features, aiding in understanding structure and shape without being affected by lighting (Figure 3). However, they have lower resolution compared to RGB images and require specific conditions for effective fusion with other modalities.

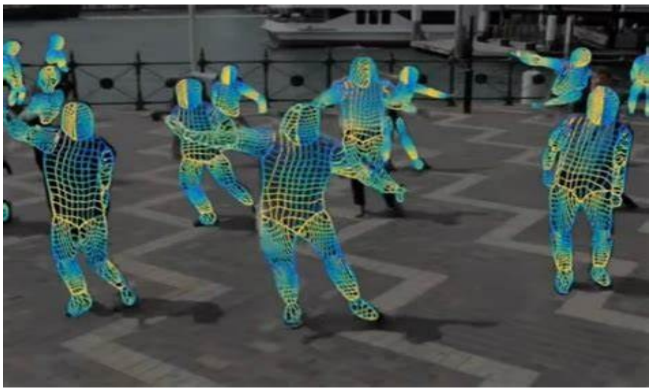

Figure 3. Depth Image [14]

Infrared imagery captures thermal information from objects and bodies to generate relevant datasets, distinguishing individuals based on unique thermal patterns even in low light or darkness (Figure 4). Processed by thermal CNNs, this modality provides effective data in varied lighting conditions but lacks the detail of RGB imagery and may vary with environmental and individual states.

Figure 4. Infrared Imagery [15]

3.3.2. Methods of feature fusion and decision fusion Feature fusion involves integrating features from multiple modalities at the feature level to create a unified representation. This approach generates a richer feature set, thereby enhancing the accuracy of ReID models. After collecting multi-modal data and independently extracting features, we obtain feature vectors from different modalities. Subsequently, we normalize and align these feature vectors using techniques such as standardization to ensure compatibility. Then, fusion techniques merge the eigenvectors of all modalities into a single high-dimensional vector. By combining information in this manner, the model acquires a detailed feature set for training and learning. Moreover, it reduces reliance on annotated data, enhancing model robustness.

Decision fusion combines decisions from each model after independently processing and classifying or scoring their respective modalities. Unlike feature fusion, the preceding steps are consistent across different modalities. In the third step, each specific modality model makes its own decisions. Voting, averaging, or stacking methods are then employed to merge these decisions. Ultimately, the most relevant modalities combine decisions specific to their modalities to derive the final recognition results. One stacking approach involves using a meta-classifier that takes outputs from individual models as inputs and makes final decisions based on these outputs. This modular approach allows customization of extraction methods for specific data types, thereby improving feature extraction accuracy for individual modalities. However, combining decisions across different modalities requires complex fusion techniques and may be influenced by technical factors.

In 2023, Wenshan Shen et al. [16] introduced a bidirectional multi-modal attention mechanism that effectively fuses visible and infrared data, capitalizing on the complementary nature of these modalities. This fusion strategy is designed to enhance feature representation by capturing both global and local discriminative cues, thereby improving the robustness of the Re-ID system against variations in lighting conditions and occlusions. The empirical evaluation demonstrates the superior performance of their proposed model, but the computational complexity associates with the dual attention mechanism, which may hinder real-time applications [16].

In addition, in 2023, Ziling He et al. [17] introduced a Low-rank Fusion Network (LFN) that adeptly combines features from multiple modalities, such as visible light and thermal imagery, to achieve a more robust representation of individuals. The LFN is underpinned by a low-rank decomposition technique that distills the essence of each modality while mitigating the impact of noise and irrelevant variations, and it sets a new benchmark for multimodal Re-ID systems. However, the network's performance may be sensitive to the quality and alignment of the input modalities [17].

4. Privacy Problem

In person ReID tasks, training typically requires the centralization of a vast number of personal image data, which poses significant privacy risks. In certain nations, these hazards have even resulted in the halting of individual ReID research initiatives. [1]. Federated learning has emerged as a solution to alleviate these privacy concerns. By decentralizing the training process and keeping data local, FL significantly reduces the potential for privacy breaches, making it a promising approach for addressing the privacy challenges in ReID systems.

4.1. Federated Learning

Federated Learning is a machine learning system that preserves privacy and may successfully address the issue of privacy in ReID jobs. Instead of transferring data to a central server, federated learning trains machine learning models on local data stored on multiple devices, such as smartphones, and continuously uploads the shared local model to a central server, ultimately generating a global model. In this way, since the edge shares model updates with the server instead of training data, it can effectively reduce the risk of potential privacy breaches [1] (Figure 5).

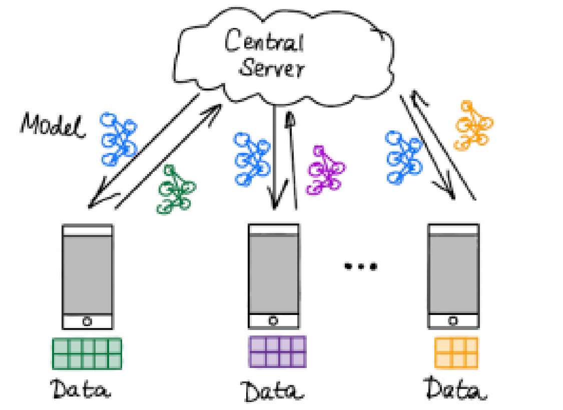

Figure 5. Federated Learning [1]

Federated learning starts by allowing each device to train its local model on its dataset for a set number of epochs. Each device then updates its local model and sends these updates to a central server, which aggregates them using algorithms like FedAvg or FedPav to update the global model. After that, the devices receive this updated global model in preparation for the upcoming training session. This iterative process continues until the model converges or meets performance goals.

Federated learning maintains decentralized data to solve privacy concerns., supports collaboration across devices with diverse datasets to enhance model robustness, and ensures scalability and efficiency. It minimizes data transfer and communication overhead while improving model generalization across different environments. Moreover, it addresses data sovereignty and regulatory compliance issues effectively.

4.2. Federated Averaging and Federated Partial Averaging

Federated Averaging (FedAvg) trains global models across decentralized devices without sharing raw data. The central server aggregates local model updates using weighted average aggregation, proportional to each device's dataset size.

$$\boldsymbol{w_{t+1}} = w_t - \eta \sum_{k=1}^{K} \frac{n_k}{n} \nabla L_k(w_t) \quad (2)$$

Federated Partial Averaging (FedPav) differs from FedAvg by selecting subsets of devices for each round of model aggregation, unlike FedAvg which aggregates updates from all devices every round. The central server randomly selects a subset of devices, which train locally on their private datasets and upload updated models. This iterative process involves different subsets of devices in each round.

$$w_{t+1} = w_t - \eta \sum_{k \in S_t} \frac{n_k}{\sum_{j \in S_t} n_j} \nabla L_k(w_t) \quad (3)$$

FedPav reduces communication and computing load compared to FedAvg by not requiring updates from all devices every round, making it suitable for resource-constrained scenarios. However, FedAvg generally achieves faster convergence and scalability for large-scale deployments. Its inclusive approach across all participants in each round helps mitigate biases that selective subsets might introduce in FedPav. FedAvg is preferred in high-throughput applications, while FedPav is ideal for environments with limited computational resources.

In 2022, Zhuang et al. [1] introduced FedPav, a novel approach addressing statistical heterogeneity in person re-identification. FedPav enables partial model aggregation across clients, enhancing knowledge transfer and convergence in decentralized learning. They established FedReIDBench, integrating nine diverse datasets to simulate real-world heterogeneity. Their experiments highlight FedPav's efficacy in improving model performance and convergence, particularly with non-IID and unbalanced data volumes. However, varying data heterogeneity among nodes may challenge consistent model performance, impacting system reliability [1].

5. Conclusion

In the rapidly evolving field of person re-identification (ReID), addressing the challenges of occlusion, dependence on labeled data, and privacy concerns is paramount for developing robust and practical systems. The review explores strategies aimed at mitigating occlusion effects using part-based models, LCNNs, attention mechanisms, and adversarial training. It also examines approaches to reduce dependence on labeled datasets through unsupervised learning, semi-supervised learning, and multimodal fusion. Furthermore, the review discusses methods to protect personal privacy, focusing on federated learning. Through this exploration, the review underscores the importance of interdisciplinary collaborations and ongoing research efforts in enhancing the reliability and scalability of ReID systems across diverse real-world scenarios. Looking ahead, continued innovation in algorithm development, dataset augmentation techniques, and privacy-preserving methodologies will play pivotal roles in advancing the field towards more efficient, inclusive, and secure person re-identification solutions.

References

[1] Wen Z. Yong W., Xin Z., et al. Performance Optimization for Federated Person Re-identification via Benchmark Analysis. In Proc. of the 28th ACM Int. Conf. on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA. ACM, New York, NY, USA, 2020, 14 pages.

[2] Somers, V . VleeschouwerC. D and A. Alahi, Body Part-Based Representation Learning for Occluded Person Re-Identification, 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2023, pp. 1613-1623.

[3] Yan G., Wang Z., Geng S., Yu Y., Guo Y., Part-Based Representation Enhancement for Occluded Person Re-Identification, IEEE Transactions on Circuits and Systems for Video Technology, vol. 33, no. 8, pp. 4217-4231.

[4] Wang P., Ding C., Shao Z., Hong Z., Zhang S., Tao D., Quality-Aware Part Models for Occluded Person Re-Identification, IEEE Transactions on Multimedia, vol. 25, pp. 3154-3165, 2023.

[5] Sun M., Wang Z., A Person Re-Identification Network Based upon Channel Attention and Self-Attention, 2021 IEEE 6th International Conference on Signal and Image Processing (ICSIP), Nanjing, China, 2021, pp. 61-65.

[6] Yang C., Guo Y., Zhang W., Unsupervised Domain Adaptation for Person Re-Identification Based on Multi-Scale Attention Mechanism, 2024 4th International Conference on Consumer Electronics and Computer Engineering (ICCECE), Guangzhou, China, 2024, pp. 105-110.

[7] Ma X., Lv W., Zhao M., A Double Stream Person Re-Identification Method Based on Attention Mechanism and Multi-Scale Feature Fusion, IEEE Access, vol. 11, pp. 14612-14620, 2023.

[8] Huang H., Li D., Zhang Z., Chen X., Huang K., Adversarially Occluded Samples for Person Re-identification, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 5098-5107.

[9] Shi X., Liu H., Shi W., Zhou Z., Li Y., Boosting Person Re-Identification with Viewpoint Contrastive Learning and Adversarial Training, ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023, pp. 1-5.

[10] Tao X., Kong J., Jiang M., Liu T., Unsupervised Domain Adaptation by Multi-Loss Gap Minimization Learning for Person Re-Identification, IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 7, pp. 4404-4416, July 2022.

[11] Weng J., Hu K., Yao T., Wang J., Wang Z., Robust Knowledge Adaptation for Federated Unsupervised Person ReID, 2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA), Sydney, Australia, 2022, pp. 1-8.

[12] Ding G., Zhang S., Khan S., Tang Z., Zhang J., Porikli F., Feature Affinity-Based Pseudo Labeling for Semi-Supervised Person Re-Identification, IEEE Transactions on Multimedia, vol. 21, no. 11, pp. 2891-2902, Nov. 2019.

[13] Wu A., Ge W., Zheng W.-S., Rewarded Semi-Supervised Re-Identification on Identities Rarely Crossing Camera Views, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 12, pp. 15512-15529, Dec. 2023.

[14] Zhihu Column. 2018, June 28. Article Image. Retrieved from https://zhuanlan.zhihu.com/p/38597956

[15] Iyiou. 2020, . Article Image. Retrieved from https://www.iyiou.com/news/20200214123501

[16] Shen W., Huang Q., Bidirectional Multi-modal Attention for Visible-Infrared Person Re-Identification, 2023 9th International Conference on Computer and Communications (ICCC), Chengdu, China, 2023, pp. 1799-1803.

[17] He Z., Shi H., Wu Y., Tu Z., Low-rank Fusion Network for Multi-modality Person Re-identification, 2023 8th International Conference on Intelligent Computing and Signal Processing (ICSP), Xi'an, China, 2023, pp. 1578-1581