Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/86/20241576

Exploring global-local feature fusion for pedestrian re-identification

Jiahao Liu

School of automation, Central South University, 932 Lushan South Road, Changsha, China

caiyu@ldy.edu.rs

Abstract. Pedestrian re-identification (ReID) is a challenging problem in computer vision, crucial for surveillance and security applications. Despite significant advancements, existing methods like Identity Discriminative Embedding (IDE) and Part-based Convolutional Baseline (PCB) have limitations. IDE captures global features but lacks detailed local information, while PCB focuses on local details without considering the global context. To address these issues, this paper attempts a fused model combining the strengths of both approaches. The fused model integrates global features from IDE and part-based features from PCB, creating a comprehensive representation that captures both holistic and localized details. Experimental results on Market-1501 and DukeMTMC datasets demonstrate the fused model's improved performance, significantly outperforming IDE and achieving comparable results to PCB. The fused model offers a balanced solution, apparently enhancing IDE performance and maintaining low computational consumption.

Keywords: Pedestrian Re-identification, feature fusion, identity discriminative embedding, part-based convolutional baseline.

1. Introduction

Pedestrian re-identification (ReID) is a crucial and challenging problem in the field of computer vision, aiming to match an individual's identity across different camera views. This task is fundamental for various applications, including surveillance, security, and intelligent transportation systems, where it is essential to accurately track and identify people across multiple, non-overlapping camera feeds. The ability to re-identify pedestrians reliably has significant implications for public safety, crime prevention, and urban management.

Desite significant advances in recent years, pedestrian ReID remains a difficult problem due to the complexities introduced by variations in lighting, poses, occlusions, and background clutter. These factors can drastically change the appearance of individuals, making it hard for algorithms to maintain consistent recognition performance [1]. Traditional methods primarily relied on handcrafted features and metric learning to achieve this goal. Handcrafted features involve designing specific algorithms to extract pertinent information from images, while metric learning focuses on optimizing distance functions to improve recognition accuracy. However, these methods often fall short when dealing with the high variability in real-world scenarios [2].

DOI: 10.54254/2755-2721/86/20241576

The advent of deep learning has revolutionized the field of pedestrian ReID, offering more powerful and effective approaches to feature extraction and similarity measurement [3]. Convolutional neural networks(CNNs), a subset of deep learning algorithms, have showcased an impressive ability to discern intricate patterns from unprocessed visual inputs autonomously. These models significantly outperform traditional methods, providing robust and scalable solutions to the ReID problem [4].

Although there has been extensive research and comparative analysis in the field of pedestrian ReID, there remains significant research value in the in-depth study of two of the most classic deep learning algorithms in this domain: Identity Discriminative Embedding (IDE) [3] and Part-based Convolutional Baseline (PCB) [5]. Such studies can provide valuable insights and directions for future algorithmic improvements.

For the IDE algorithm, the primary focus is on global features, which means it captures the overall appearance of the pedestrian. This global perspective can sometimes overlook finer details that are crucial for distinguishing between individuals with similar overall appearances. On the other hand, the PCB approach emphasizes extracting detailed information by dividing the pedestrian image into meaningful parts. While this part-based strategy excels in capturing fine-grained details, it often does so at the expense of neglecting the holistic view of the pedestrian's appearance.

This thesis proposes enhancing the IDE method by introducing a part-based approach to further refine its performance. By segmenting pedestrian imagery into its constituent elements and independently extracting distinctive characteristics from each segment, the model can capture finer details crucial for distinguishing between similar-looking individuals. This approach combines the robust global features of IDE with the detailed part-based features, aiming to achieve superior performance in complicated conditions. Meanwhile, this thesis compares the performance of three prominent methods in pedestrian ReID: IDE, PCB, and Transformer-based Re-identification (TransReID) [6]. These comparisons are essential to evaluate the effectiveness of the proposed enhancements to the IDE algorithm. By benchmarking the improved IDE against PCB and TransReID, this study aims to highlight the advantages and potential drawbacks of each approach. This comprehensive evaluation will demonstrate how the integration of part-based features into the IDE framework can lead to superior performance.

2. Related Work

2.1. Traditional Approaches

Early methods in ReID primarily centered on designing robust feature descriptors and metric learning algorithms. Handcrafted features, such as color histograms, texture descriptors, and shape features, were extensively used to represent pedestrian images. One notable example is the Local Maximal Occurrence (LOMO) feature, which captures local color and texture information and was considered a significant advancement in the early stages of ReID research. Metric learning techniques, such as Pairwise Constrained Component Analysis (PCCA) and Large Margin Nearest Neighbor (LMNN), were employed to measure the similarity between feature vectors, enhancing the discriminative power of these models [7].

Despite their contributions, these traditional approaches faced significant challenges in dealing with complex variations in pedestrian appearance and environmental conditions. The reliance on handcrafted features often limited their ability to generalize across different datasets and scenarios. These methods struggled to handle variations in lighting, pose, and occlusions, which are common in real-world surveillance settings. As a result, the performance of traditional methods was often inadequate for practical applications, highlighting the need for more advanced techniques.

2.2. Advanced Techniques and Trends

Beyond these foundational methods, advanced techniques have emerged to further boost ReID performance. For instance, attention mechanisms, such as those used in the Harmonious Attention Network (HAN), focus on salient parts of the input image, enhancing fine-grained feature extraction. While effective, these methods can be computationally intensive and may struggle with occlusions.

DOI: 10.54254/2755-2721/86/20241576

Generative Adversarial Networks (GANs) have been explored for data augmentation and domain adaptation, with methods like Domain Adaptation GAN (DAGAN) [8] generating synthetic data to bridge domain gaps. However, GANs often require extensive training and can be sensitive to training instability.

Another direction is the use of unsupervised and semi-supervised learning, aiming to leverage vast amounts of unlabeled data. These approaches show promise but typically achieve lower accuracy compared to supervised methods. This analysis highlights the need for methods that balance complexity, performance, and robustness, motivating the research in this thesis to enhance the IDE algorithm by incorporating part-based features for improved ReID accuracy and reliability.

The evolving landscape of pedestrian ReID continues to evolve from both traditional and novel methodologies. By integrating the robust feature extraction capabilities of CNNs, the detailed local analysis provided by part-based models, and the global context understanding afforded by Transformers, current research aims to develop more accurate and reliable ReID systems. Future advancements will likely focus on further enhancing model generalization and robustness, particularly in unconstrained environments. In summary, the field of pedestrian ReID has made significant strides from its early days of relying on handcrafted features and metric learning to the current state-of-the-art techniques that leverage deep learning, attention mechanisms, GANs, and multi-modality data. These advanced methods provide more reliable technical support for real-world applications of ReID systems.

3. Method

This chapter focuses on the comparative analysis of several prominent methods in pedestrian ReID, specifically IDE, PCB and TransReID [3],[5],[6]. These methods have been selected due to their foundational significance and widespread adoption in the field. Despite their success, each method has certain limitations that warrant further investigation. IDE primarily captures global features and may overlook finer details, while PCB excels in extracting detailed features but often neglects the holistic view of the pedestrian. TransReID leverages advanced transformer architectures but can be computationally demanding and complex to implement.

By critically examining these methods, this study aims to identify their strengths and weaknesses, providing a comprehensive understanding of their performance. This analysis serves as a foundation for the proposed enhancement to the IDE algorithm, which incorporates part-based features to address its current limitations. The goal is to achieve a more balanced and robust ReID system, offering significant improvements in accuracy and reliability under challenging conditions. Through this research, we aim to contribute valuable insights that advance the development of more effective pedestrian ReID solutions.

3.1. IDE feature

IDE is one of the most fundamental deep learning approaches in ReID. This method leverages CNNs to extract robust and discriminative features from pedestrian images, which are then mapped into a high-dimensional feature space. In this feature space, images of the same person (identity) are positioned close together, while images of different people are far apart, facilitating effective re-identification (Figure 1).

IDE typically employs well-known CNN architectures such as ResNet [9] and DenseNet [10]. ResNet, with its residual learning framework, allows for the training of very deep networks by mitigating the vanishing gradient problem, thus enabling the extraction of highly informative features. In contrast, the DenseNet model facilitates a feed-forward interconnection between all layers, thereby amplifying the flow of features and promoting their recurrent utilization. This design strategy culminates in the generation of more concise and potent feature representations. The IDE approach involves training a CNN as a classification model, where each class represents a different identity from the training set. During training, the network learns to differentiate between the various identities, effectively capturing the unique features associated with each person. The features extracted from the final layer of the network, just before the classification layer, are then used for the re-identification task. These features serve as a compact and highly discriminative representation of the pedestrian images.

DOI: 10.54254/2755-2721/86/20241576

The key strengths of IDE is its simplicity and effectiveness, which have established it as a baseline method in pedestrian ReID tasks. Despite its relatively straightforward implementation, IDE has demonstrated strong performance across various benchmarks, making it a reliable choice for many practical applications. The approach's robustness stems from the powerful feature extraction capabilities of CNNs, which are adept at handling the variations and complexities inherent in real-world pedestrian images, such as changes in lighting, pose, and background.

Moreover, IDE's effectiveness has been proven in numerous studies and competitions, solidifying its status as a cornerstone technique in the field of ReID. Its ability to produce highly discriminative features with relatively simple training procedures makes it an attractive option for both researchers and practitioners aiming to develop efficient and accurate ReID systems. This thesis chooses IDE as one of the focal methods due to its established reliability and effectiveness in real-world scenarios. By building on this well-regarded approach, the research aims to introduce enhancements that address its current limitations, specifically by incorporating part-based features. This will not only improve the granularity of feature extraction but also leverage the robust global features IDE is known for.

This thesis chooses IDE as one of the focal methods due to its established reliability and effectiveness in real-world scenarios. By building on this well-regarded approach, the research aims to introduce enhancements that address its current limitations, specifically by incorporating part-based features. This will not only improve IDE's granularity of feature extraction but also leverage the robust global features.

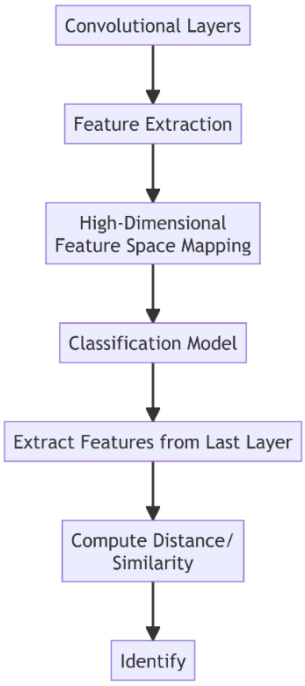

Figure 1. Workflow of IDE [3]

3.2. PCB (Part-based Convolutional Baseline)

PCB represents a significant advancement in ReID by focusing on part-based feature extraction. Unlike traditional methods that rely on global feature extraction, PCB divides the pedestrian image into several horizontal stripes, each of which is processed independently. This approach allows PCB to capture more detailed and localized information, which is crucial for distinguishing between individuals in complex ReID scenarios [5].

Specifically, PCB splits the input image into a fixed number of horizontal parts, typically four to six. Each part is processed through a CNN to extract fine-grained features. For instance, if an image is divided into six parts, each horizontal stripe will go through its own CNN pathway, ensuring that the unique features of each body part are captured effectively (Figure 2). These part-based features are then

DOI: 10.54254/2755-2721/86/20241576

concatenated to form a comprehensive feature representation. By doing this, PCB creates a more robust and detailed feature vector that significantly improves the model's ability to re-identify pedestrians under varying conditions.

This method's effectiveness lies in its ability to capture detailed information while being robust to common challenges in ReID, such as occlusions and pose variations. When parts of a pedestrian's body are occluded, the PCB method can still rely on the unoccluded parts to make accurate identifications. This robustness makes PCB particularly adept at handling occlusions, as it minimizes the impact of occluded regions by focusing on the information available from the visible parts of the body.

PCB's part-based approach makes it highly effective in dealing with pose variations. Different poses can significantly alter a pedestrian's appearance, but by analyzing smaller, more consistent parts of the body, PCB can maintain higher accuracy. This localized focus ensures that even if the overall pose changes, the features from individual parts remain consistent and reliable for identification purposes.

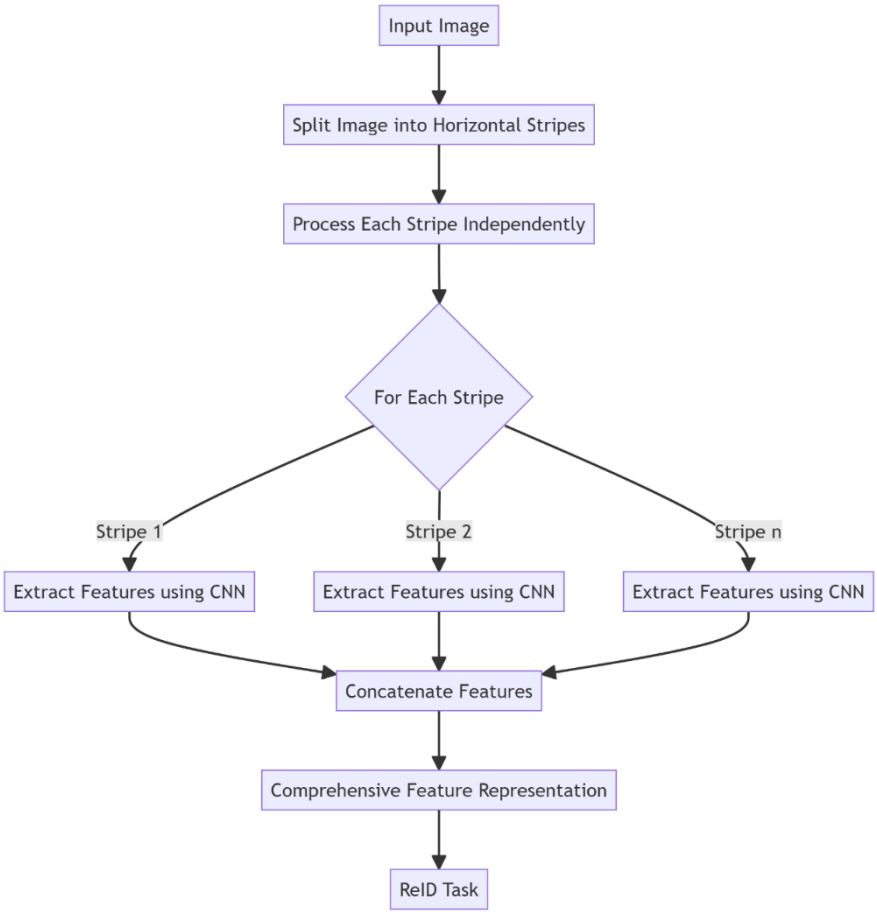

Figure 2. Workflow of PCB[5]

3.3. TransReID

TransReID introduces the use of Transformer architecture to the ReID domain, marking a significant departure from traditional CNNs. Unlike CNNs, which rely on convolutional filters to extract local features, the Transformer architecture employs self-attention mechanisms, enabling the capture of extensive interdependencies present across the dataset, offering a robust means for global context modeling [5]. This ability to capture global relationships within an image makes Transformers particularly well-suited for complex tasks like pedestrian re-identification.

TransReID begins by dividing the input image into smaller patches, treating each patch as a token. These patches are then linearly embedded into vectors, which are subsequently processed through

DOI: 10.54254/2755-2721/86/20241576

multiple layers of self-attention within the Transformer. This method enables the model to learn both global and local features, enhancing its robustness to common challenges in ReID such as occlusions and pose variations. By processing patches independently and then integrating information through self-attention, TransReID effectively captures the intricate relationships between different parts of the image.It integrates part-level feature extraction modules that further refine the feature representation. The approach integrates the complementary advantages of CNNs and Transformer models, capitalizing on the fine-grained feature detection proficiency inherent to CNNs alongside the expansive contextual comprehension provided by Transformers. This hybrid approach has demonstrated significant promise in handling complex surveillance environments where both fine-grained details and overall context are crucial (Figure 3).

One of the key innovations of TransReID is its ability to maintain high performance in the presence of occlusions. Since the self-attention mechanism can weigh the importance of each patch, the model can focus on unoccluded regions and still make accurate identifications. This flexibility makes TransReID highly effective in real-world surveillance scenarios where occlusions are common.

The architecture of TransReID sets a new direction for future research in the field of ReID and beyond. By showcasing the potential of Transformer-based models in addressing complex computer vision tasks, TransReID inspires further exploration and application of this architecture in other research areas.

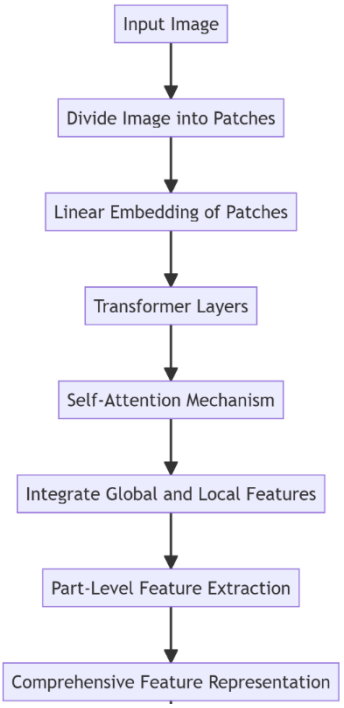

Figure 3. Workflow of TransReID [6]

3.4. Feature Fusion Approach

The experiments in this paper mainly try to combine the respective features of IDE global feature fusion-based and part-based convolutional methods, and to combine the fused global features and the component features extracted in chunks. The IDE model is known for its effective global feature representation, which captures the overall appearance of a person, while the PCB model excels at learning detailed, part-based features that help distinguish between similar-looking individuals by focusing on local body parts. The fused model aims to integrate these two approaches by extracting global features in the IDE manner and part-based features as in PCB, subsequently fusing them to create a comprehensive representation that captures both global and local characteristics.

DOI: 10.54254/2755-2721/86/20241576

The fused model integrates the strengths of both the IDE and PCB approaches to enhance performance in person re-identification tasks. This hybrid model leverages the comprehensive representation capabilities of global features from IDE and the fine-grained discrimination power of part-based features from PCB. The base architecture of the model uses pre-trained ResNet-50 extracting features up to the penultimate convolutional block to retain high-dimensional feature maps. From these base features, the model derives global features using global average pooling to reduce spatial dimensions to a single vector. If necessary, a fully connected layer reduces the feature dimension, followed by batch normalization to stabilize training and optional dropout to prevent overfitting. Intermediate features are also extracted from specific layers within the final convolutional block, capturing localized information that complements the global context.

Simultaneously, the model extracts intermediate features from specific layers within the final convolutional block of the backbone network. These intermediate features are intended to capture localized information from different parts of the image, which is crucial for distinguishing between individuals with similar global appearances. Each intermediate feature map undergoes adaptive average pooling to reduce its spatial dimensions, followed by batch normalization. These intermediate features are then averaged with the global feature vector, creating a fused feature that encompasses both holistic and localized details. The fusion process involves expanding the fused feature and splitting it into stripes corresponding to different parts of the person image, mimicking the part-based approach of the PCB model. Each stripe is processed separately through normalization, dropout, and batch normalization to ensure stability (Figure 4). Individual classifiers, specific to each stripe, then process these part-based features, allowing the model to recognize different body parts and enhance its robustness to variations in pose and occlusion.

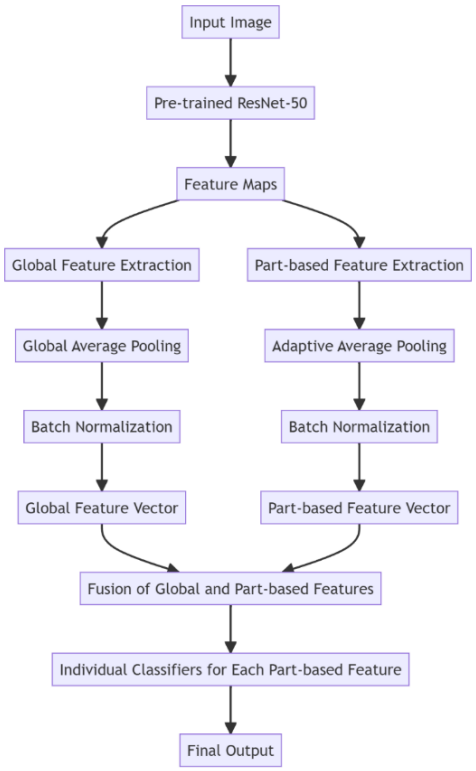

Figure 4. Workflow of fused model (Photo/Picture credit : Original)

DOI: 10.54254/2755-2721/86/20241576

4. Experiment

4.1. Datasets

This thesis's experiments are constructed around two widely-used datasets within the realm of person re-identification: the Market-1501 and the DukeMTMC-reID. The Market-1501 dataset is comprised of 32,668 labeled bounding boxes, representing a total of 1,501 unique individuals, while DukeMTMC-reID comprises 36,411 images of 1,404 identities [1]. The performance evaluation metrics used are rank-1 accuracy and mean Average Precision (mAP), which are standard in pedestrian ReID research.

4.2. Hyperparameters and training details

The fused model was trained and evaluated on the Market-1501 and DukeMTMC-reID datasets. Each dataset provides a challenging environment with diverse images captured from multiple cameras. Training involved 60 epochs using data augmentation techniques such as random cropping, flipping, and normalization. The SGD optimizer with a dynamic learning rate scheduler was used to facilitate effective learning, and the loss function Cross Entropy Loss for classification to boost feature discriminativeness.

For the model, the backbone network is based on ResNet-50, with a feature dimension set to 256 and a dropout rate of 0.5. The images are resized to a height of 256 and a width of 128 pixels. The batch size is set to 64, with four workers handling data loading. Techniques for augmenting data, including methods like randomized horizontal inversion, selective cropping, and the random removal of information, are utilized to bolster the resilience of the model. The training configuration encompasses a total of 60 epochs, initiated with a learning rate set at 0.1, subject to a reduction by an order of magnitude every 40 epochs.The final stride of the network is set to 2, and label smoothing regularization (LSR) is optionally applied.

The training regimen for these models is executed through the application of the Stochastic Gradient Descent (SGD) optimization algorithm, characterized by a momentum coefficient of 0.9 and a weight decay rate of 5e-4. The experiments are conducted on an NVIDIA GPU with 16GB memory, using the PyTorch deep learning framework.

4.3. Performance comparison

An evaluation of the performance of various methodologies on the Market-1501 and DukeMTMC-reID datasets is summarized in the following Table 1 and Table 2.

Assess results show that the highest performance achieved by the TransReID method on both datasets, with 85.90% mAP and 94.20% Rank-1 on Market-1501, and 88.01% mAP and 94.70% Rank-1 on DukeMTMC-reID. TransReID leverages advanced transformer-based architectures, which excel at modeling long-range dependencies and contextual information, providing a more comprehensive understanding of the input images. The Fused_model, which combines global and part-based features, significantly outperforms the baseline IDE method. This improvement demonstrates the effectiveness of integrating global and part-based features, allowing the model to capture finer details that are crucial for distinguishing between similar-looking individuals.

Despite the significant improvements over IDE, the Fused_model's performance is still slightly lower than that of the PCB method, which achieves 76.52% mAP and 92.40% Rank-1 on Market-1501, and 77.73% mAP and 92.40% Rank-1 on DukeMTMC-reID. This difference can be attributed to PCB's more sophisticated part-based feature extraction, which may capture more nuanced details and improve the model's robustness to variations in pose and appearance. Simple fusion of global and local features may not fully utilize them, and more complex fusion is needed. Additionally, the PCB method's explicit focus on part-based features might offer better spatial alignment, further enhancing its discriminative power.

In conclusion, while the Fused_model shows a significant performance boost over IDE by combining global and part-based features, there remains room for further improvement to match or surpass the capabilities of PCB and TransReID, particularly in terms of capturing fine-grained details and achieving better spatial alignment.

DOI: 10.54254/2755-2721/86/20241576

Table 1. The result from different methods on Market1501

Market1501	mAP	Rank-1
LOMO	36.47%	50.10%
IDE	67.51%	85.93%
Fused_model	74.93%	90.77%
PCB	76.52%	92.40%
TransReid	85.90%	94.20%

Table 2. The result from different methods on DukeMTMC

DukeMTMC	mAP	Rank-1
LOMO	30.10%	41.32%
IDE	56.36%	77.56%
Fused_model	66.93%	80.24%
PCB	68.24%	82.68%
TransReid	78.70%	88.50%

4.4. Computational Efficiency

The computational efficiency of each method is assessed by measuring the time per epoch and the total time consumption. From the result(Table 3), IDE is the most computationally efficient, with a time per epoch of 18.17 seconds and a total training time of 1090.2 seconds. The Fused_model, which introduces additional part-based features, incurs a slight increase in computational cost, with a time per epoch of 18.31 seconds and a total training time of 1098.6 seconds. This minor increase is expected due to the additional complexity in feature extraction but is still comparable to IDE, indicating that the proposed enhancements do not significantly hinder training efficiency.

Table 3. The results of computational efficiency

Methods	Time per epoch	Total training time(60epoch)
IDE	18.17s	1090.2s
Fused_model	18.31s	1098.6s
PCB	32.14s	1928.4s
TransReid	21.42s	1285.2s

In contrast, PCB and TransReID both show higher computational costs. PCB has a time per epoch of 32.14 seconds and a total training time of 1928.4 seconds, due to its sophisticated part-based feature extraction and alignment mechanisms that require more processing power and time. TransReID, leveraging transformer-based architectures, also has a higher computational cost compared to IDE and the Fused_model, with a time per epoch of 21.42 seconds and a total training time of 1285.2 seconds. Despite the higher computational costs, PCB and TransReID provide superior performance, suggesting that the additional time and resources are justified by their significant gains in accuracy.

These findings highlight the trade-off between computational efficiency and accuracy, with the Fused_model providing a balanced approach that enhances performance without a substantial increase in training time.

5. Conclusion

This thesis explored the performance and computational efficiency of three prominent methods in pedestrian ReID: IDE, PCB, and TransReID. Additionally, we attempted a Fused_model that combines global and part-based features to enhance IDE's performance. Our experiments conducted on the Market-1501 and DukeMTMC-reID datasets reveal that the integrated model substantially surpasses the

DOI: 10.54254/2755-2721/86/20241576

foundational IDE approach, attaining superior mean Average Precision (mAP) and Rank-1 precision metrics. However, while the Fused_model offers a balanced trade-off between accuracy and computational efficiency, it still slightly lags behind PCB and TransReID in terms of performance.

The computational efficiency analysis showed that IDE and the Fused_model are more efficient than PCB and TransReID, with the latter two methods incurring higher training times due to their sophisticated feature extraction mechanisms. These findings underscore the importance of balancing accuracy with computational demands, particularly for real-world applications where resources may be limited.

Looking forward, future research could focus on further optimizing the Fused_model to close the performance gap with PCB and TransReID while maintaining or even improving its computational efficiency. Additionally, exploring advanced techniques such as attention mechanisms and domain adaptation could provide further enhancements. By continuing to refine these models, we can develop more robust and efficient pedestrian ReID systems, contributing to advancements in surveillance, security, and intelligent transportation systems.

References

[1] Zheng, L., et al. 2015 "Scalable person re-identification: A benchmark." Proceedings of the IEEE international conference on computer vision. 2015.

[2] Wu, D., et al. 2017 "Deep learning-based methods for person re-identification: A comprehensive review." Neurocomputing 337: 354-371.

[3] Zheng, Z., Liang Z., and Yi Y.. 2017 "A discriminatively learned cnn embedding for person reidentification." ACM transactions on multimedia computing, communications, and applications (TOMM) 14.1 : 1-20.

[4] Wang, G., et al. 2018"Learning discriminative features with multiple granularities for person re-identification." Proceedings of the 26th ACM international conference on Multimedia. 2018.

[5] Sun, Y., et al. 2018 "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)." Proceedings of the European conference on computer vision (ECCV). 2018.

[6] He, S., et al. 2021"Transreid: Transformer-based object re-identification." Proceedings of the IEEE/CVF international conference on computer vision. 2021.

[7] Mignon, Alexis, and Frédéric Jurie. 2012 "Pcca: A new approach for distance learning from sparse pairwise constraints." 2012 IEEE conference on computer vision and pattern recognition. IEEE, 2012.

[8] Radford, Alec, Luke Metz, and Soumith Chintala. 2015 "Unsupervised representation learning with deep convolutional generative adversarial networks." arXiv preprint arXiv:1511.06434.

[9] He, K., et al. 2016 "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 1-11.

[10] Huang, Gao, et al. 2017 "Densely connected convolutional networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 1-9.