DOI: 10.54254/2755-2721/52/20241611

FIRCNet: Feature-based image reconstruction with classification learning for person re-identification

Junhao Chen

School of Engineer and Applied Science, The George Washington University, 20052

junhao@gwu.edu

Abstract. As an important task in computer vision, the person re-identification problem has two main approaches, identification and verification, in solving this problem. This paper proposed a network associating these two separate methods in training, including a reconstructive model. This design encompasses three core modules: a global feature extractor, a reconstructive decoder, and a classification network. Thus, the total loss function seamlessly integrates triplet loss, cross-entropy loss, and a reconstruction loss, optimizing the network for feature discriminability, accurate identity classification, and faithful image reconstruction concurrently. The global feature plays a pivotal role in both the training and testing phases, aiding in metric learning and ensuring a distinctive representation of identities. To train the decoder and the backbone network, which collectively forms the extractor, simultaneously, the reconstructive model as a part of a wide-range autoencoder can recover the original image from extracted features. Subsequently, the classification module classifies these features, assigning them to distinct person IDs, thereby aiding in precise identity recognition. Experimental results on three major re-id datasets give a demonstration of the notable improvement in the backbone network and the advantages of this approach compared to the state-of-the-art methods.

Keywords: Person Re-identification, Image Reconstruction, Combination Learning, Cross-Dataset Verification

1. Introduction

Person re-identification (often terms of Person Re-id) refers to the challenge of recognizing the same individual across different images, captured by different cameras with various views. As a crucial module in Multi-target Multi-camera (MTMC) tasks, the re-id module's performance decides the MTMC system's ability to associate the information from separate cameras and track a person in multiple views through further steps. As the MTMC system grows larger, more viewpoints present a great challenge to the re-id task. Higher perspective variation means more complex lighting conditions, complicated background information, changeable appearance, etc. making this task intricate.

Benefiting from the deep learning algorithm, the re-id solutions have been prompted from some simple metrics, for example, color distributions and textures [1], to more complex high-dimensional features obtained from neural networks [2]. There are two major ways, which are classification and retrieval models, to apply these features to solve re-id problems [3]. The classification model, as a more straightforward method, takes images as input and outputs the persons' IDs directly through the neural network, and therefore has better inference speed [4]. However, the pre-defined number of output classes makes the classification model very hard to extend the number of outputs in other test scenarios. The retrieval model, in another way, treats the re-id task as a verification problem [5]. It trains deep neural networks to extract feature vectors from the images of different people. By applying metric learning, it uses a designed distance algorithm to retrieve the most similar features from gallery datasets to verify their identities.

This paper proposes a person re-id framework combining the retrieval model with a classification learning module and a Feature-based Image Reconstruction module on the training side. This model aims to blend the classification loss with the loss of retrieval model to enrich the information gained in the training process while avoiding the limitation of the extendibility of the classification model. The reconstructive model, as a part of the autoencoder, recovering the image from the obtained feature to compare with the original images, can supplement the normal feature retrieving in the re-id task.

The main contributions are described as follows:

This paper proposes an image reconstruction network to recover images from extracted features.

A classification module is implemented with the feature extraction network in the training process. The cross entropy as classification loss joins with triplet loss and MSE loss to build the blended loss.

Experiments are conducted comprehensively on four major re-id datasets and achieve significant improvement on rank-1 and mAP.

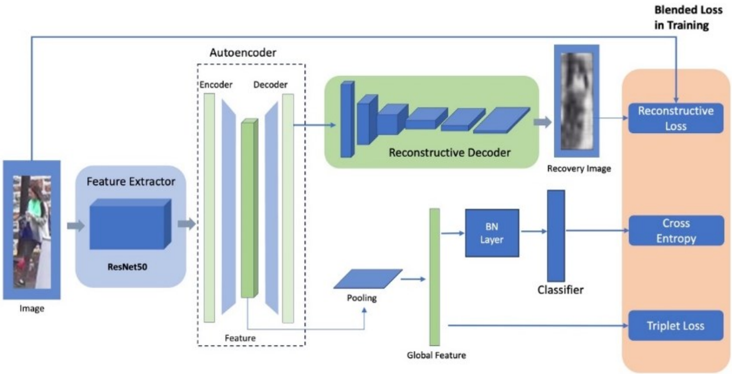

Figure 1. Overview of FIRCNet Framework for Person Re-identification

2. Related Work

2.1. *Retrieval and Classification Model*

The domain of person re-identification has evolved, with significant contributions focusing on the synergistic use of verification and identification models. The exploration of combining retrieval and classification models in the realm of person re-identification began to gain significant attention as early as 2015. This period marked the inception of concerted efforts to blend these two methodologies, aiming to leverage the strengths of each to enhance the performance of re-ID systems. The approach of a four-stream deep learning architecture [6] underscores the integration of identification-verification methodologies, optimizing a newly proposed quartet loss function to enhance person re-ID. Complementing this, there is research that defends the use of classification loss, presenting a method that enriches training information while preserving the model's extendibility for person re-ID [7].

Zheng et al. [8] have explored the fusion of verification and identification models to refine discriminative descriptors for re-ID, which has shown promise in enhancing the robustness of the models against varied challenges such as occlusion and pose variation. Furthermore, cross-modality concerns are addressed through joint middle modality and representation learning, which seeks to map differing modal images to a unified representation, thus aiding in the reduction of modality discrepancy [9].

A retrospective survey of domain-specific challenges and trends in person re-ID, covering a wide spectrum of approaches and datasets, provides a systematic review of the state-of-the-art methods, highlighting the need for a more integrated approach to tackle issues like cross-domain generalization and scalability [10]. This extensive survey delineates the progress made in each challenge area, suggesting a path forward that involves a composite loss function to bridge the gap between research and real-world application efficacy. Wang et al. [11] revisit the fundamentals of person re-identification by conceptualizing it as a combination of retrieval and verification, which challenges the prevailing closed-world image retrieval approach and advocates for a new metric that encapsulates both retrieval and verification processes. This paradigm shift underscores the necessity to balance the retrieval of candidate images with the verification of their true identity, thereby enhancing the accuracy and reliability of re-ID systems.

2.2. Autoencoder in Reid Tasks

Autoencoder, through unsupervised learning and feature extraction capabilities, offers a robust avenue for tackling challenges in re-identification tasks. It helps in not only improving the performance of re-id tasks but also in advancing scalability. Existing re-id methods often require a large training dataset, which can hinder their performance on smaller or different datasets. The use of autoencoders, as part of deep learning techniques, can help improve model generalization, especially when the available training data is limited [12]. Autoencoders are also used for extracting deep features and assisting in image patching which is crucial for person re-identification. They help in managing variations in appearance, illumination, and viewpoints, hence aiding in finding true matches from non-overlapping camera images or videos [13]. To learn a universal domain-invariant feature representation from multiple labeled datasets, the adversarial autoencoder is used to learn a generalized domain-invariant latent feature representation and align the distributions across multiple domains, solving the overfitting problem in the re-id task [14].

The convergence of retrieval and verification not only aligns with but also catalyzes the development of the proposed FIRCNet framework. By integrating retrieval-based metric learning with the verification accuracy of classification models, FIRCNet embodies the essence of re-ID, while also avoiding overfitting with its feature-based image reconstruction component.

3. Method

The FIRCNet is an identification model extracting feature vectors for identifying a person's ID combined with several modules, including the image reconstruction module and classification module (Figure 1), to assist the training process. The Image Reconstruction module recovers the images from extracted features to help decrease the bias by comparing them with the original images. The classification module utilizes the feature to classify the person's identity and generate cross-entropy as supplement information. The proposed model finally blends the feature triplet loss with the classification loss and the Mean Square Error (MSE) between the reconstructive and original images.

As a great feature-extracting network, ResNet is widely used in feature extraction to solve re-id problems. The FIRCNet also uses ResNet50 as the backbone. To reconstruct images, this paper takes the features before adaptive average pooling in ResNet50 so that the features can have higher dimensions and make it easier for reconstruction. The autoencoder shown in Figure 1 aims to map the feature to the global space to prevent overfitting [14] while connecting the extractor with the classifier and reconstruction decoder. All these sub-networks will be presented in detail in the following subsections.

3.1. Global Feature

The global feature is the only feature that metric learning refers to during the testing process. Through adaptive average pooling, this global feature is generated by sampling from the feature inside the autoencoder. In the training process, the global feature is used in the classifier, and also directly used for triplet. After the training process, the network uses the Euclidean distance of the global feature to verify the identity of the input images. this paper set the (f = 2048, 1024, 512) with the dimension of the feature. Based on the size of the feature vector, the encoder changes its depth (d = 1, 2, 3) accordingly.

3.2. Reconstructive Decoder

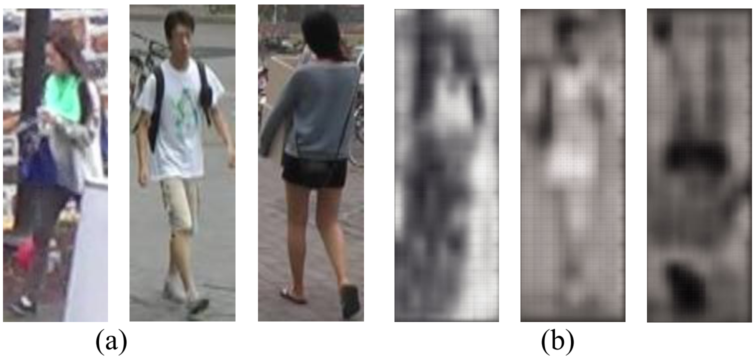

Figure 2. Original Images and Recovery Images

This network reconstructs the image through the output feature from the decoder. Thus, this network can also be regarded as a part of the decoder in autoencoder, minimizing the difference between input and output. As mentioned above, the dimension of the global feature vector is pre-defined. Therefore, before the reconstructive decoder, a decoder is needed to adjust the dimension of the feature vectors, making it suitable for the reconstructive decoder input. To reconstruct the images from those compressed features, this paper adapts transposed convolution layers to up-sample the features $f_{r,i}$ and recover the images, where the final output $f_{r,7}$ is the recovery image I_r. The difference between the original images I_o and recovery images I_r is shown in Figure 2. Relu is applied as the connection between transposed convolution layers. Thus, the feature equation for i^{th} layer TC_i is denoted as below

$$f_{r,i+1} = TC_{i+1}(Relu(f_i)), where\ f_1 = TC_1(f_0), i \in [2, 6] \tag{1}$$

3.3. Classification Network

The classification module is a crucial component designed to assign a person ID to the extracted global features. This ID classification aids in person recognition by categorizing the global features into distinct person IDs. The process involves a two-stage transformation: a bottleneck layer $BN()$ followed by a softmax layer $softmax()$. The global feature f_g, obtained after various feature extraction processes, is first passed through a bottleneck layer. This layer is designed to reduce the dimensionality of the global feature, making it more manageable and focusing on the most critical and distinguishing aspects of the data. The output from the bottleneck layer, now a compressed representation of the global feature, is fed into a SoftMax layer. The SoftMax layer is designed to classify the feature into one of the possible person IDs, according to the output probability distribution p. Given the dataset's nature, the total number of categories (or classes) in the softmax layer corresponds to the total number of unique person IDs present in the dataset.

$$p = softmax\left(BN(f_g)\right) \tag{2}$$

3.4. Loss Design

Combining different parts of the sub-network in training, this work constructs the total loss function including losses from separate tasks above. The total loss in the network is a composite of three distinct loss components, each catering to a specific module. The triplet loss is derived from the global feature f_g, ensuring that the distance between an anchor feature $f_{g,a}$ and a positive $f_{g,p}$ (same identity) image in the feature space is minimized, while simultaneously maximizing the distance to a negative (different identity) feature $f_{g,n}$. This loss ensures that the global features are discriminative and can accurately distinguish between different identities with the margin m.

$$\mathscr{L}_t(x, W) = \max\left\{0, \left||f_{g,a} - f_{g,p}|\right|_2 - \left||f_{g,a} - f_{g,n}|\right|_2 + m\right\} \quad (3)$$

The cross-entropy loss from the classification module gauges the difference between the predicted person ID probabilities and the actual labels, promoting accurate categorization of identities. For the global feature f_g mentioned above, it is then passed through a bottleneck layer $BN()$ to enhance discriminative details and reduce overfitting. The compressed features are then fed into a softmax layer, which maps them to a probability distribution p across all possible classes, where p_i is the possibility of i-th class among class set K in probability distribution p. Finally, the cross entropy $\mathcal{L}_c$ is denoted as follow

$$\mathscr{L}_c(x, W) = \sum_{i \in K} -p_i \log p_i \quad , where\ p = softmax\left(BN\left(f_g\right)\right) \quad (4)$$

The reconstructive loss stems from the image reconstruction module. This loss measures the discrepancy between the original input images and their reconstructed versions from the autoencoder, pushing the network to maintain the integrity and quality of the image during the encoding-decoding process. To measure the difference between the original I_o and recovery image I_r, this paper denotes the mean square error as the reconstructive loss $\mathcal{L}_r$ for dataset D which has in total $|D|$ images.

$$\mathscr{L}_r(x, W) = \frac{1}{|D|} \sum_{i \in D} \left|\left| I_{o,i} - I_{r,i}\right|\right|_2 \quad (5)$$

The aggregate of these losses from the task set ($T = \{c, r, t\}$) mentioned above forms the total loss, guiding the network in harmonizing feature extraction, accurate classification, and faithful image reconstruction. For the weight W and each input x, the total loss $\mathscr{L}_{total}$ is defined as follows.

$$\mathscr{L}_{total} = \sum_{i \in T} \lambda_i \mathscr{L}_i(x, W) \quad (6)$$

4. Experiment

In this section, the paper will describe the details of conducting the experiments. To evaluate the proposed model, this paper tests the model on three wide-reaching re-id datasets, including Market-1501 [15], Duke-MTMC [16], and MSMT-17 [17] in the experiments.

4.1. Dataset

The Market-1501 dataset is a widely recognized benchmark for person re-identification tasks, featuring 1,501 distinct identities captured across six different cameras. It consists of a total of 32,668 bounding box images, with an average of 3.6 images per identity for each viewpoint. The dataset is unique for utilizing a Deformable Part Model (DPM) as the pedestrian detector, and it provides both true positive bounding boxes and false alarm detection results. This dataset was collected in front of a supermarket at Tsinghua University, employing five high-resolution cameras and one low-resolution camera with overlapping coverage among them.

DukeMTMC-reID (Duke Multi-Tracking Multi-Camera Re-Identification) dataset is a subset of the larger DukeMTMC dataset designed for image-based person re-id. It comprises images from high-resolution videos captured by eight different cameras, making it one of the largest pedestrian image datasets. The dataset includes images of 1,852 individuals, with 1,413 unique identities having 22,515 bounding boxes appearing in more than one camera. In total, the dataset consists of 1,404 identities, 2,228 queries, 17,661 gallery images, and 16,522 training images, demonstrating a significant resource for re-id research and development.

MSMT17 (Multi Scene Multi Time dataset for person re-id) is a large-scale re-id dataset consisting of 180 hours of videos captured by 15 cameras (12 outdoor and 3 indoor) during 12 different time slots across four days. It includes 126,441 bounding boxes of 4,101 identities, making it one of the largest re-id datasets. The diverse collection environment and the extensive data volume provide a robust benchmark for evaluating person re-identification algorithms under various real-world conditions.

The model is implemented in PyTorch and the experiments are conducted through a single V100 GPU. The comprehensive setting is designed to assess the model's performance across different configurations and the impact of individual components on its effectiveness. During the training phase, this paper varied the loss functions and feature sizes to examine their effects on the model's accuracy and convergence speed. Specifically, this paper evaluated the following configurations: softmax loss only, triplet loss only, a combination of softmax and triplet losses, and the full FIRCNet model with a global feature vector of varying dimensions (2048, 2048 reduced to 512, and 2048 reduced to 1024). Additionally, this paper incorporated a re-ranking strategy with the FIRCNet (2048) configuration to improve retrieval precision.

Training was conducted over 120 epochs to ensure that each model configuration had ample opportunity to converge. Performance metrics such as mean average precision (mAP) and rank-1, rank-5, and rank-10 accuracies were recorded at specific epoch intervals (10, 20, 30, ..., 120) to monitor progression and stability. These metrics were chosen to provide a comprehensive understanding of the models' identification capabilities at various retrieval depths.

4.2. Result Comparison

Table 1. Comparative Performance of FIRCNet Against State-of-the-Art Models on Person Re-Identification Datasets

Models	Market-1501		Duke-MTMC		MSMT17	
	r1	mAP	r1	mAP	r1	mAP
DL-CNN [18]	85.84	70.33	-	-	-	-
CG+MB [7]	92.6	78.3	82.8	66.7	-	-
Quartet [6]	91.6	75.7	82.4	77.3	-	-
PCB (SoftMax) [19]	93.8	81.6	83.3	69.2	68.2	40.4
CircleLoss(ResNet50) [20]	94.2	84.9	-	-	76.3	50.2
FIRCNet	93.2	81.2	86.3	71.6	77.5	48.3
FIRCNet+Re-ranking	94.2	84.4	85	77.5	79.0	51.1

FIRCNet framework was rigorously evaluated against several state-of-the-art models on the Market-1501, Duke-MTMC, and MSMT17 datasets. The comparison underscores FIRCNet's competitive performance across these challenging benchmarks shown in Table 1.

On the Market-1501 dataset, FIRCNet achieved a rank-1 accuracy of 93.2% and an mAP of 81.2%, which is competitive with other leading methods such as DL-CNN [18] and PCB (softmax)[19]. Notably, FIRCNet outperforms Quartet [6] in mAP, highlighting the efficacy of the integrated approach. Duke-MTMC results were even more promising, with FIRCNet attaining a rank-1 accuracy of 86.3% and mAP of 71.6%. These figures surpass the results of Quartet [6] and CG+MB [7], demonstrating the advantages of the framework in handling the complex variations present in this dataset. The MSMT17 dataset, known for its scale and diversity, presented a more challenging scenario. However, FIRCNet managed to achieve a rank-1 accuracy of 77.5% and a mAP of 48.3%. The implementation of FIRCNet with Re-ranking further improved performance, pushing rank-1 accuracy to 79.0% and mAP to 51.1%, indicating the significant impact of re-ranking on enhancing retrieval precision.

The FIRCNet is compared with several models with a similar approach combining verification modules with identification modules [6,7,18] mentioned above and get a significant improvement on both mAP and Rank 1. Besides these models, two state-of-the-art models are also introduced in the result comparison. Compared to the CircleLoss (ResNet50) [20], which achieved a mAP of 50.2% on the MSMT17 dataset, FIRCNet with Re-ranking showed an improvement in mAP, establishing its potential for real-world application where dataset diversity is high.

Overall, the results from the experiments solidify the position of FIRCNet as a robust and efficient framework for person re-identification. The added benefit of Re-ranking on top of the base FIRCNet model provides a clear pathway for future research to explore further enhancements in retrieval-based person re-identification tasks.

Table 2. Cross-Dataset Evaluation - Training on Market-1501 and Testing on Duke-MTMC

	Testing Dataset			
Training Dataset	Market-1501		Duke-MTMC	
	r-1	mAP	r-1	mAP
Market-1501	94.2	84.4	45.6	31.0
Duke-MTMC	58.3	32.1	85.0	77.5
	Testing Dataset			
Training Dataset	Market-1501		Duke-MTMC	
	r-1	mAP	r-1	mAP
Market-1501	91.6	78.7	37.6	22.6
Duke-MTMC	48.2	21.6	83.4	66.6

Table 2 presents a comparison of cross-dataset generalization capabilities between a baseline model and the FIRCNet model, which incorporates an additional image reconstruction module.

From the right table, the baseline model exhibits a significant drop in performance when trained on Market-1501 and tested on Duke-MTMC, with rank-1 accuracy falling from 91.6% to 37.6% and mAP from 78.7% to 22.6%. Compared with the left table, FIRCNet, whose result is shown in the right table, enhanced with the image reconstruction module and re-ranking, shows a remarkable improvement in cross-dataset testing. When trained on Market-1501 and tested on Duke-MTMC, rank-1 accuracy and mAP increase to 45.6% and 31.0%, respectively. Similarly, when trained on Duke-MTMC, FIRCNet demonstrates strong performance and improved generalization to Market-1501, with rank-1 accuracy and mAP increasing to 58.3% and 32.1%, respectively.

The improved cross-dataset generalization of the FIRCNet model suggests that the image reconstruction module plays a crucial role in mitigating overfitting. By forcing the network to learn to reconstruct the original images, the model develops a more robust and generalized feature representation. This is evidenced by its superior performance on a dataset it was not trained on, indicating that the learned features are not merely memorized data-specific cues but rather carry transferable knowledge between datasets. The use of image reconstruction, therefore, contributes significantly to the model's ability to generalize across different re-identification scenarios, which is a desirable characteristic in practical applications.

4.3. Ablation Study

Table 3. Ablation Study Results on Market-1501 and Duke-MTMC Datasets Showcasing the Impact of FIRCNet Components

Modules	Market-1501				Duke-MTMC			
	mAP	r-1	r-5	r-10	mAP	r-1	r-5	r-10
softmax	76.1	90.8	96.1	97.4	63.1	80.9	90.4	92.4
triplet	78.2	90.6	95.6	97	70.3	85.3	92.7	94.8
softmax-triplet	81.6	94	97.7	98.5	67.9	84	92.1	94.5
FIRCNet(2048)	81.2	93.2	97.9	98.7	71.6	86.3	93.2	94.9
FIRCNet(2048-512)	79.8	92.7	97.4	98.4	67.9	84.2	92.5	94.6
FIRCNet(2048-1024)	80.4	93.2	97.4	98.2	67.8	83.6	91.7	93.9
FIRCNet(2048 + Re-ranking)	84.4	94.2	98.0	99.5	77.5	85.0	94.7	96.0

An in-depth ablation study was carried out on the Market-1501 and Duke-MTMC datasets to critically assess the contributions of different components in the FIRCNet framework shown in Table 3.

The foundational softmax loss registered a mAP of 76.1% and a rank-1 accuracy of 90.8% on Market-1501, with the Duke-MTMC dataset showing a mAP of 63.1% and a rank-1 accuracy of 80.9%. The standalone triplet loss showed improved robustness, particularly on Duke-MTMC, where it achieved a mAP of 70.3% and rank-1 accuracy of 85.3%. The combination of softmax and triplet losses substantially boosted performance, yielding a mAP of 81.6% and rank-1 accuracy of 94% on Market-1501. The FIRCNet configurations with varying global feature sizes demonstrated the model's adaptability. The FIRCNet (2048) variant achieved a notable mAP of 81.2% and rank-1 accuracy of 93.2% on Market-1501, and mAP of 71.6% with rank-1 accuracy of 86.3% on Duke-MTMC. The reduced feature size variants, FIRCNet (2048-512) and FIRCNet (2048-1024), maintained high performance, indicating that feature compactness does not significantly compromise the model's discriminative power.

Remarkably, the introduction of re-ranking to the FIRCNet (2048) configuration led to a significant leap in accuracy, pushing the mAP to 84.4% and rank-1 accuracy to an impressive 94.2% on Market-1501. On Duke-MTMC, re-ranking improved the mAP to 77.5% and rank-1 accuracy to 85.0%, outperforming all other configurations and underscoring its efficacy in enhancing the retrieval precision.

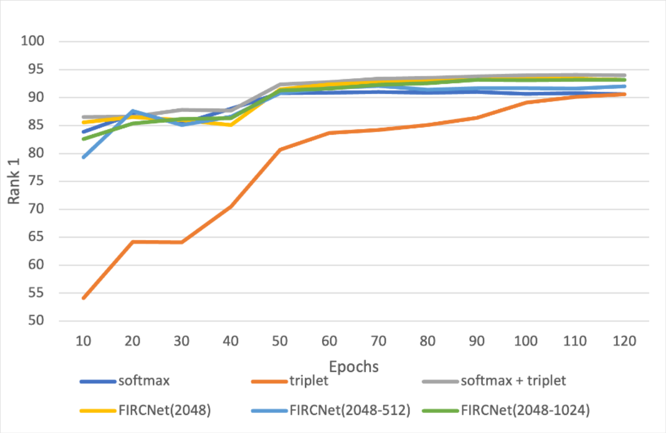

Figure 3. Rank-1 Accuracy Progression Over Training Epochs for FIRCNet Configurations on Market-1501

The incorporation of the classification module within the network, as evidenced in the classification module and the classification with retrieval module configurations shown in Figure 3, increases the convergence speed. The classification alone shows a steady convergence, while the softmax and triplet combination loss demonstrate both faster convergence and higher final accuracy, highlighting the classification module's role in accelerating the network's learning process. This suggests that the classification module provides additional discriminative information early in the training, which allows for quicker learning and more stable convergence, ultimately enhancing the FIRCNet network's ability to learn distinctive features faster and more effectively.

This study demonstrates the significance of each component within the FIRCNet framework. Overall, the ablation study not only confirms the individual and collective contributions of the components in the FIRCNet framework but also highlights the classification module's pivotal role in accelerating the network's learning process. This results in faster convergence to high accuracy, underscoring the value of the design choices in enhancing the performance of person re-identification systems.

5. Conclusion

The research presented in FIRCNet introduces a novel approach to person re-identification by integrating a feature-based image reconstruction network with the combination of classification and retrieval learning. This integration enables the model to leverage the strengths of both identification and verification approaches, resulting in a system that excels in feature discriminability, accurate identity classification, and faithful image reconstruction.

The ablation study and extensive experiments conducted across several major re-id datasets, such as Market-1501, Duke-MTMC, and MSMT17, have demonstrated the superior performance of the FIRCNet framework, especially when enhanced with Re-ranking. The model's innovative loss design, combining triplet loss, cross-entropy loss, and reconstruction loss, has proven to be effective in guiding the network towards a more robust and discriminative feature space that can handle the complexities of person re-identification tasks.

Moreover, FIRCNet's ability to generalize across datasets, as evidenced by the cross-dataset evaluation, suggests that the model's features are not overfitted to specific datasets but are rather generalizable and adaptable to different scenarios. This characteristic is crucial for real-world applications where a model is expected to perform reliably across various environments and conditions.

In conclusion, FIRCNet represents a significant step forward in the realm of person re-identification. The framework's design allows for an effective balance between accuracy and generalizability, making it a robust solution for real-world surveillance and tracking systems. Future work could explore further enhancements to the network's architecture and loss functions, as well as applications in related domains where feature discriminability and image reconstruction are essential.

References

[1] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, “Hierarchical Gaussian Descriptor for Person Re-identification,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1363–1372.

[2] L. Zheng, Y. Yang, and A. G. Hauptmann, “Person re-identification: Past, present and future,” 2016, arXiv:1610.02984.

[3] Zheng, Zhedong & Zheng, Liang & Yang, Yi. (2016). A Discriminatively Learned CNN Embedding for Person Re-identification. ACM Transactions on Multimedia Computing, Communications, and Applications. 14. 10.1145/3159171.

[4] Zhai, Y., Guo, X., Lu, Y., & Li, H. (2018, November 29). In defense of the classification loss for person re-identification. arXiv.org. https://arxiv.org/abs/1809.05864

[5] Zheng, Z., Zheng, L., & Yang, Y. (2017). A discriminatively learned CNN embedding for person reidentification. ACM Transactions on Multimedia Computing, Communications, and Applications, 14(1), 1–20. https://doi.org/10.1145/3159171

[6] Khatun, A., Denman, S., Sridharan, S., & Fookes, C. (2020). Joint identification–verification for person re-identification: A four stream deep learning approach with improved quartet loss function. Computer Vision and Image Understanding, 197–198, 102989. https://doi.org/10.1016/j.cviu.2020.102989

[7] Zhai, Y., Guo, X., Lu, Y., & Li, H. (2019). In defense of the classification loss for person re-identification. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). https://doi.org/10.1109/cvprw.2019.00194

[8] Xie, Z., Li, L., Zhong, X., Zhong, L., & Xiang, J. (2020). Image-to-video person re-identification with cross-modal embeddings. Pattern Recognition Letters, 133, 70–76. https://doi.org/10.1016/j.patrec.2019.03.003

[9] Ma, L., Guan, Z., Dai, X., Gao, H., & Lu, Y. (2023). A cross-modality person re-identification method based on joint middle modality and representation learning. Electronics, 12(12), 2687. https://doi.org/10.3390/electronics12122687

[10] Zahra, A., Perwaiz, N., Shahzad, M., & Fraz, M. M. (2023). Person re-identification: A retrospective on domain-specific open challenges and future trends. Pattern Recognition, 142, 109669. https://doi.org/10.1016/j.patcog.2023.109669

[11] Wang, Z., Yuan, X., Yamasaki, T., Lin, Y., Xu, X., & Zeng, W. (2020, November 23). Re-identification = retrieval + verification: Back to essence and forward with a new metric. arXiv.org. https://arxiv.org/abs/2011.11506

[12] Kansal, K., & Subramanyam, A. V. (2019). Autoencoder Ensemble for person re-identification. 2019 IEEE Fifth International Conference on Multimedia Big Data. https://doi.org/10.1109/bigmm.2019.00-15

[13] Khatun, A., Denman, S., Sridharan, S., &amp; Fookes, C. (2020). Joint identification–verification for person re-identification: A four stream deep learning approach with improved quartet loss function. Computer Vision and Image Understanding, 197–198, 102989. https://doi.org/10.1016/j.cviu.2020.102989

[14] Lin, S., Li, C.-T., & Kot, A. C. (2020, November 25). Multi-domain adversarial feature generalization for person re-identification. arXiv.org. https://arxiv.org/abs/2011.12563

[15] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, & Q. Tian, “Scalable Person Re-identification: A Bench- mark,” in Proc. IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1116–1124.

[16] Zhedong Z., Liang Z., &Yi Y. 2017. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In Proceedings of the IEEE International Conference on Computer Vision, 754–3762.

[17] Wei, L., Zhang, S., Gao, W., & Tian, Q. 2018. Person transfer gan to bridge domain gap for person re-identification. In Proceedings of the IEEE CVPR, 79–88.

[18] Zheng, Z., Zheng, L., & Yang, Y. (2017). A discriminatively learned CNN embedding for person reidentification. ACM Transactions on Multimedia Computing, Communications, and Applications, 14(1), 1–20. https://doi.org/10.1145/3159171

[19] Sun, Y., Zheng, L., Yang, Y., Tian, Q., & Wang, S. (2018). Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). Computer Vision – ECCV 2018, 501–518. https://doi.org/10.1007/978-3-030-01225-0_30

[20] Sun, Y., Cheng, C., Zhang, Y., Zhang, C., Zheng, L., Wang, Z., & Wei, Y. (2020). Circle loss: A unified perspective of pair similarity optimization. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr42600.2020.00643