Deep Learning with Pytorch: Siamese Network

Dr.T.Deepa[1], Dr.G.Satyavathy[2], Research, Ms.A.Priya3 , Mrs.S.SriSakthi Hamrish[4],

[1]Associate Professor & Head, Department of Computer Science, KPR College of Arts Science and Research

[2]Professor & Head, Department of Computer Science with Data Analytics, KPR College of Arts Science and Research

[3] Lab Technician, Department of Computer Science and Engineering ,United Institute of Technology

[4]Assistant Professor, Department of Computer Science and Engineering, United Institute of Technology, Ms. S. Evangeline Aishwarya[5], Assistant Professor, Department of Artificial Intelligence and Data Science, United Institute of Technology

Abstract - "DEEP LEARNING WITH PYTORCH: SIAMESE NETWORK" is a work that addresses person re-identification (re-ID), a difficult computer vision challenge that entails identifying the same person from several camera angles. Because SNNs may learn similarity instead of straight classification, they are becoming a preferred method for this kind of assignment. Using this method, a ranking loss function is optimized by two concurrent CNNs that learn an embedding, or reduced dimensional representation, of the input images. An overview of the procedures involved in person re-identification using SNNs is given in the study, including training, testing, deployment, network architecture, and data preparation. It makes use of the Triplet Ranking Loss function, a popular loss function for SNNs.For similarity-based learning tasks including face recognition, image matching, and document similarity, Siamese Neural Networks are one kind of neural network design that is utilized. The paper offers a thorough tutorial on training a Siamese neural network for a goal based on similarity, namely using the Siamese Neural Network (SNN) to re-identify images taken by different cameras.

Keywords – *Re-identification, Pytorch, Feature Learning, Multi-camera*

I. Introduction

This study presents a deep learning model for person re-identification, i.e., matching an individual's identity over several non-overlapping camera views. The Siamese network is used to train the model for person re-identification utilizing pairs of photographs of the same person and pairs of images of different persons. Each image is mapped by the network to a high-dimensional feature space where different images are placed farther apart and like photos closer together. PyTorch tensors are created by reading and converting the images inside the dataset. An anchor picture, a positive image, and a negative image are returned by a custom dataset that is constructed. After that, a unique data loader processes these pictures and outputs batches of photographs. A custom model is defined, which uses an Efficient Net backbone to extract embedding from the input images. The model is then trained on the training set using a Triplet Margin Loss function, an Adam optimizer, and a learning rate of 0.001. The best model is saved during training, and the final model is evaluated on the validation.An Efficient Net backbone is used by a custom model that is created to extract embedding from the input images. After that, the model is trained on the training set with a learning rate of 0.001, an Adam optimizer, and a Triplet Margin Loss functions. During training, the best model is saved, and the final model is assessed during validation.

The necessity for precise and dependable person re-identification in surveillance and security systems is the reason behind the beginning of this paper. Occlusion, changes in posture and viewpoint, variations in appearance and lighting make person re-identification a difficult task. Siamese network-based deep learning techniques in particular have shown promise in tackling these problems. The goal of the paper is to investigate the application of PyTorch-based deep Siamese networks with multi-layer similarity constraints for person re-identification. The goal is to use deep learning and the Siamese network architecture to increase the precision and dependability of person re-identification in difficult situations like occlusion. The potential applications of this research are numerous, including surveillance and security systems for public safety, monitoring and tracking individuals in crowded environments, and enhancing the performance of intelligent transportation systems. Improving person re-identification can help prevent crime, assist law enforcement agencies, and enhance the safety and security of communities.

Applications for Siamese networks include image matching, face recognition, signature verification, and re-identification of individuals. The ability of Siamese networks to learn a similarity metric between two input images or sequences is one of their main advantages. This ability makes the networks especially helpful for tasks that require matching or comparing inputs. Siamese networks can be trained to learn a similarity metric between two images of the same person and dissimilarity metric between images of different people. This training can be applied to person re-identification, allowing people to be matched across various camera views. Siamese networks can handle cases where the same person appears differently in different camera views or when the same camera view is occluded, making them a valuable tool for person re-identification.

Ii .Literature Review

AUTHOR & YEAR	TITLE OF THE PAPER	DATASET USED	ALGORITHM USED	ACCURACY	CHALLENGES
Meenakshi Choudhary et al.., 2021	Person re-identification using deep siamese network with multi-layer similarity constraints	CUHK03 and Market-1501 dataset	Deep Siamese Network	84.7%	1.Variations in Illumination 2.Small Dataset Size
LiangboWang et al.., 2021	Occluded person re-identification based on differential attention Siamese network	Market1501, DukeMTMC-ReID and Partial-ReID	Differential Attention Siamese Network(DASN)	84.9%	1.Variability in appearance 2.Sensitivity to hyperparameters
DasolJeong et…., 2020	Uniformity Attentive Learning-Based Siamese Network for Person Re-	Market-1501, CUHK03 and DukeMTMC	Siamese network	94.5%	1.Variation in poses and lighting conditions

	Identification	-ReID			
YonghongTian et al.., 2016	Deep Transfer Learning for Person Re-identification	CUHK03, Market1501, and VIPeR	Convoluational Neural Networks (CNNs)	89.4%	1.Domain Shift 2.Label Noise
Ngai-Man Cheung et al.., 2018	Efficient and Deep Person Re-Identification using Multi-Level Similarity	CUHK01 ,VIPeR	Convolution similarity network (CSN)	88.50%	1.Occlusion handling 2.Privacy concerns
Guo-Yuan Fei et al.., 2019	Learning Large Margin Multiple Granularity Features with an Improved Siamese Network for Person Re-Identification	CUHK01, CUHK03, Market-1501 and DukeMTMC -reID	Siamese Multiple Granularity Network (SMGN)	87.1%	1.Data quality 2.Overfitting 3.Human interpretation
Arne Schumann et al.., 2017	Person Re-Identification by Deep Learning Attribute-Complementary Information	CUHK3	Convoluational Neural Networks (CNNs)	84.61%	1.Variability in appearance 2.Privacy concerns
Vivek Tiwarl et al.., 2021	Person re-identification using deep siamese network with multi-layer similarity constraints	CUHK03	Deep Siamese Network	96.1%	1.Small Dataset Size 2.Label Noise
NumanCelebi et al.., 2022	Similarity based person re-identification for multi-object tracking using	MOT17	Deep Siamese neural network	85.61%	1.Less Data Collection 2.Feature Extraction

	deep Siamese network				
Jesus Martinez del Rincon et al.., 2016	Recurrent Convolutional Network for Video-based Person Re-Identification	iLIDS-VID and PRID-2011	Recurrent Neural Network	88.61%	1.Domain Shift 2.Label Noise
Khalid Tahboub et al.., 2017	A Two Stream Siamese Convolutional Neural Network For Person Re-Identification	PRID2011 and iLIDS-VID	Siamese Convolutional Neural Network	92.23%	1.Variation in poses and lighting conditions 2.Label Noise
TuomasEerola et al.., 2020	Siamese Network Based Pelage Pattern Matching for Ringed Seal Re-identification	MOT17	Convolutional Neural Network	82.5%	1.Variability in pelage patterns 2.Quality of images
Yang Wang et al.., 2019	Where-and-When to Look: Deep Siamese Attention Networks for Video-Based Person Re-Identification	iLIDS-VID , PRID 2011 , and MARS	Convolutional Neural Network	82.3%	1.Variation in poses and lighting conditions 2.Quality of images
Shiliang Zhang et al.., 2019	Deep Representation Learning With Part Loss for Person Re-Identification	VIPeR , CUHK03 , and Market1501 .	Convolutional Neural Network	89.87%	1.Less Data Collection 2.Feature Extraction
Alessandro Borgia Et al.., 2018	Cross-View Discriminative Feature Learning for Person Re-Identification	CUHK03 and Market-1501	Convolutional Neural Network	97.54%	1.Variability in appearance 2.Privacy concerns

I. Methodology

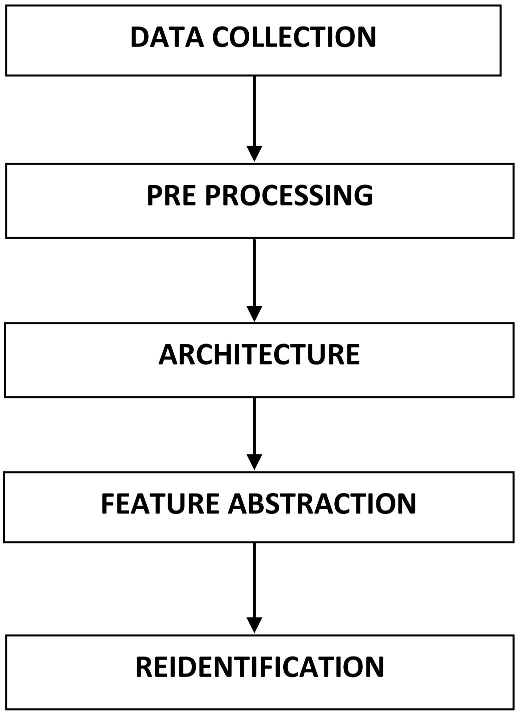


Fig 1 Methodology Overview

A. Data collection

The dataset is collected from a kaggle repository, the dataset comprises 50000 security camera images, which depict various scenes or contexts involving humans. These images are the primary data on which the segmentation task will be performed. Additionally, there are 40000 ground truth images, each associated with a specific camera images. These ground truth images serve as the reference or annotation, providing accurate pixel-level masks or annotations indicating the regions occupied by human objects in the corresponding human images. To establish the relationship between the human images and their corresponding ground truth images, a CSV file is provided. This file contains structured tabular data, typically in a comma-separated values format, with each row representing an entry or sample and the columns providing relevant metadata such as image file names, paths, or other identifiers. The CSV file acts as a mapping or reference, facilitating the association of each human image with its respective ground truth image. Overall, this data collection process involves retrieving human and ground truth images from a kaggle repository, along with a CSV file that provides the necessary linkage between the images. This dataset can then be utilized for training and evaluating machine learning models or algorithms for human segmentation tasks.

B. Pre processing

Firstly, the data is read from a CSV file, dataset containing images of people captured from different camera views. The dataset contains information about the corresponding identities of the people in the images. Once the dataset is collected, the images are preprocessed by resizing them to a consistent size and converting them to grayscale or RGB format. The images are also normalized to ensure consistent color and brightness across the dataset. To introduce more variations and improve the robustness of the model, data augmentation techniques such as flipping, rotation, and color transformations are applied to the images. The augmented dataset is then divided into training and validation sets, with the split depending on the size and composition of the dataset.

After data preparation and augmentation, the images are fed into the Siamese network as pairs. Each pair consists of two images, with one image from the first camera view and the other image from the second camera view. The Siamese network learns to extract features from the two images and compute a similarity score between them. During training, a triplet loss function is used to optimize the network parameters.

Finally, the preprocessed data is organized into batches using data loaders to facilitate efficient and convenient data feeding during training and evaluation. The data loaders handle batching, shuffling, and parallel data loading to optimize the utilization of computational resources. Overall, this data preprocessing pipeline ensures that the input data is properly formatted, augmented, and divided into appropriate sets for training and validation. It prepares the data in a way that promotes effective learning and generalization by the Siamese network.

C. Architecture

Siamese Neural Network is a model architecture which contains at least two parallel, identical, Convolutional Neural Networks. This parallel CNN architecture allows for the model to learn similarity, which can be used instead of a direct classification. SNNs have found uptake primarily for image data, such as in facial recognition, although they do have their uses outside of this domain, each parallel CNN which forms a part of the SNN is designed to produce an embedding, or a reduced dimensional representation, of the input.

Architecture of the Siamese Network These embeddings can be used to optimize a Ranking Loss and at test time used to generate a similarity score. The parallel CNNs can, in theory, take any form. One important point however is that they must be completely identical; they must share the same architecture, shares the same and updated weights, and has the same hyper parameters. This consistency allows the model to compare the inputs it receives, usually one per CNN branch.

D. Feature abstraction

In person re-identification using a Siamese network, segmentation is a critical step for accurate feature extraction from input images. The network is designed with an encoder-decoder architecture that utilizes pre-trained weights for initialization. During feature extraction, the model processes the input images and applies a series of operations, including convolutional and pooling layers, to extract distinctive visual patterns and structures. The encoder component plays a crucial role in feature extraction, enabling the model to efficiently extract meaningful features from the input images.

The output of the feature extraction process is a set of extracted features, which represent the learned representation of the input images. These features are then passed through a similarity metric function, such as the triplet ranking loss function, to calculate the similarity between the input images. This similarity score is then used to identify whether the input images correspond to the same person or not.

By using a Siamese network for person re-identification, the model effectively learn similarity-based representations of the input images, rather than direct classification. This approach saves computational resources and enables efficient transfer learning, as the network has already learned useful features from a large and diverse dataset.

E. Reidentification

Re-identification can be used as a preprocessing step in Siamese networks. In this approach, the input images are first segmented to extract regions of interest, such as the person's body or face, to facilitate subsequent processing and comparison. The segmentation model can be trained using a variety of techniques, such as Triplet Loss Function, to generate binary masks that identify the desired regions within the images. During training, pairs of input images and their corresponding ground truth masks are fed into the segmentation model, which learns to generate predicted masks that closely match the ground truth masks. Once the segmentation model is trained, it can be used to generate masks for new, unseen images. These masks can then be fed into the

Siamese network to extract features and compare images in a more targeted and efficient manner. By leveraging segmentation in this way, the model can improve its accuracy and robustness by focusing on the relevant regions of the input images.

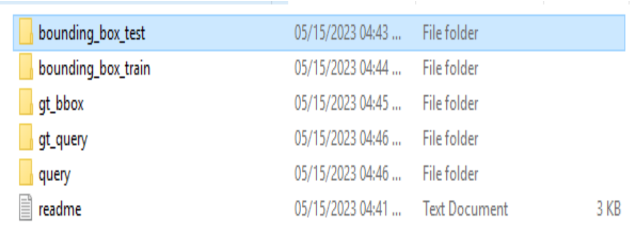
III. Experimental result and analysis

F. Dataset description

The dataset used for training and testing plays a crucial role in the performance of the model. The dataset consists of images of people captured from different viewpoints and in different environments. The images contain variations in lighting conditions, pose, background, occlusion, and clothing appearance. The dataset is divided into three subsets: training, validation, and testing. The training set trains the model's parameters, while the validation set tune hyper parameters and monitor the model's performance during training. The testing sets evaluate the final performance of the trained model on unseen data.

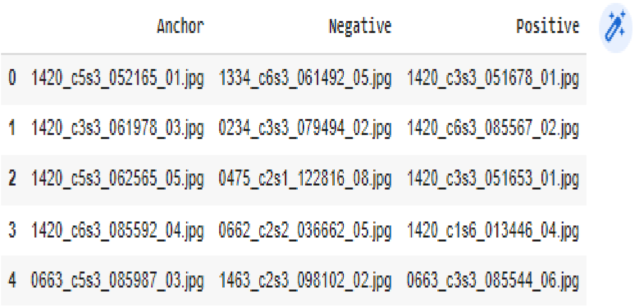
The training dataset contains pairs of images, where each pair consists of two images of the same person captured from different viewpoints or in different environments. The pairs are labeled as either positive or negative, indicating whether the images belong to the same person or different people, respectively. During training, the network is fed pairs of images and their corresponding labels. The goal of the training is to optimize the network's parameters to learn a similarity function that maps two input images to a similarity score.

The validation set tune hyper parameters, such as the learning rate, number of layers, and batch size, to improve the model's performance. The testing set is used to evaluate the final performance of the model on unseen data. The performance of the model is typically measured in terms of metrics such as accuracy, precision, recall, and F1-score.



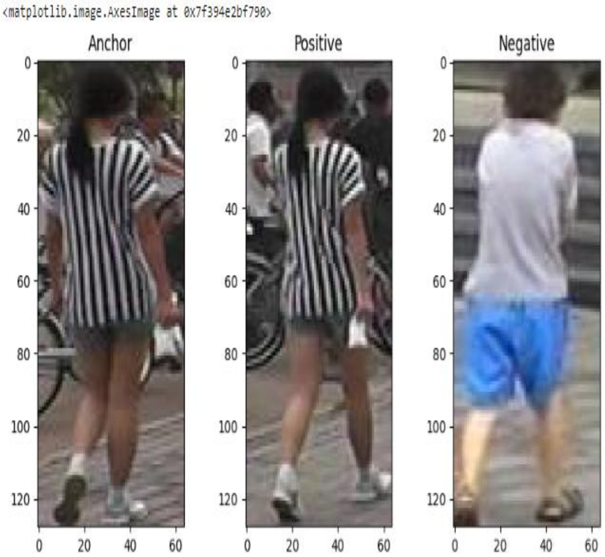
bounding_box_test	05/15/2023 04:43 ...	File folder	
bounding_box_train	05/15/2023 04:44 ...	File folder	
gt_bbox	05/15/2023 04:45 ...	File folder	
gt_query	05/15/2023 04:46 ...	File folder	
query	05/15/2023 04:46 ...	File folder	
readme	05/15/2023 04:41 ...	Text Document	3 KB

Fig 2 Directories in Dataset



	Anchor	Negative	Positive
0	1420_c5s3_052165_01.jpg	1334_c6s3_061492_05.jpg	1420_c3s3_051678_01.jpg
1	1420_c3s3_061978_03.jpg	0234_c3s3_079494_02.jpg	1420_c6s3_085567_02.jpg
2	1420_c5s3_062565_05.jpg	0475_c2s1_122816_08.jpg	1420_c3s3_051653_01.jpg
3	1420_c6s3_085592_04.jpg	0662_c2s2_036662_05.jpg	1420_c1s6_013446_04.jpg
4	0663_c5s3_085987_03.jpg	1463_c2s3_098102_02.jpg	0663_c3s3_085544_06.jpg

Fig 3 Visualization of CSV file



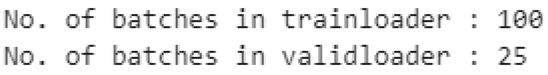
```
<matplotlib.image.AxesImage at 0x7f394e2bf790>
```

Fig 4 Images of Anchor, Positive and Negative

A. *Training and evaluation*

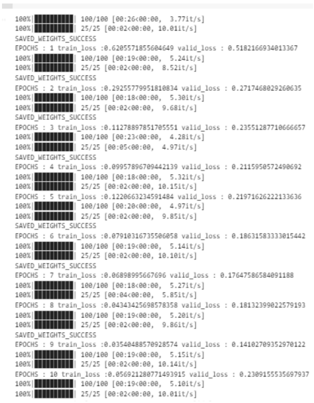
The training phase of the person re-identification using Siamese network, the model accurately recognize and match images of people by optimizing its parameters based on a dataset of input image pairs and corresponding labels indicating whether the images depict the same person or not. The process involves iterating over the dataset, generating pairs of input images, calculating a loss that quantifies the difference between the predicted and ground truth labels, and updating the model's parameters to minimize the loss. This iterative process enables the model to improve its ability to recognize and match images of people over time. A separate dataset is used for testing the performance of the model. This dataset contains input image pairs and corresponding labels, but the model has not seen these images during training.

Fig 5 Training loop images



```
No. of batches in trainloader : 100
No. of batches in validloader : 25
```

Fig 6 Batches in training and validation



```
100%|██████████| 100/100 [00:26<00:00,  3.77it/s]
100%|██████████| 25/25 [00:02<00:00, 10.01it/s]
SAVED_WEIGHTS_SUCCESS
EPOCHS : 1 train_loss :0.6205571855604649 valid_loss : 0.5182166934013367
100%|██████████| 100/100 [00:19<00:00,  5.24it/s]
100%|██████████| 25/25 [00:02<00:00,  8.52it/s]
SAVED_WEIGHTS_SUCCESS
EPOCHS : 2 train_loss :0.29255779951810834 valid_loss : 0.2717468029260635
100%|██████████| 100/100 [00:18<00:00,  5.30it/s]
100%|██████████| 25/25 [00:02<00:00,  9.68it/s]
SAVED_WEIGHTS_SUCCESS
EPOCHS : 3 train_loss :0.1127889785170555 valid_loss : 0.23551287710666657
100%|██████████| 100/100 [00:23<00:00,  4.28it/s]
100%|██████████| 25/25 [00:05<00:00,  4.97it/s]
SAVED_WEIGHTS_SUCCESS
EPOCHS : 4 train_loss :0.09957896709442139 valid_loss : 0.2115950572490692
100%|██████████| 100/100 [00:18<00:00,  5.32it/s]
100%|██████████| 25/25 [00:02<00:00, 10.15it/s]
EPOCHS : 5 train_loss :0.1220663234591484 valid_loss : 0.2197162622133636
100%|██████████| 100/100 [00:20<00:00,  4.97it/s]
100%|██████████| 25/25 [00:02<00:00,  9.85it/s]
SAVED_WEIGHTS_SUCCESS
EPOCHS : 6 train_loss :0.0791031673550605 valid_loss : 0.18631583333015442
100%|██████████| 100/100 [00:19<00:00,  5.14it/s]
100%|██████████| 25/25 [00:02<00:00, 10.10it/s]
SAVED_WEIGHTS_SUCCESS
EPOCHS : 7 train_loss :0.06898995667696 valid_loss : 0.17647586584091188
100%|██████████| 100/100 [00:18<00:00,  5.27it/s]
100%|██████████| 25/25 [00:04<00:00,  5.85it/s]
EPOCHS : 8 train_loss :0.04343425698578358 valid_loss : 0.18132399022579193
100%|██████████| 100/100 [00:19<00:00,  5.20it/s]
100%|██████████| 25/25 [00:02<00:00,  9.86it/s]
SAVED_WEIGHTS_SUCCESS
EPOCHS : 9 train_loss :0.03540488570928574 valid_loss : 0.14102709352970122
100%|██████████| 100/100 [00:19<00:00,  5.15it/s]
100%|██████████| 25/25 [00:02<00:00, 10.14it/s]
EPOCHS : 10 train_loss :0.05692128077149391 valid_loss : 0.2309155535697937
100%|██████████| 100/100 [00:19<00:00,  5.10it/s]
100%|██████████| 25/25 [00:02<00:00, 10.01it/s]
```

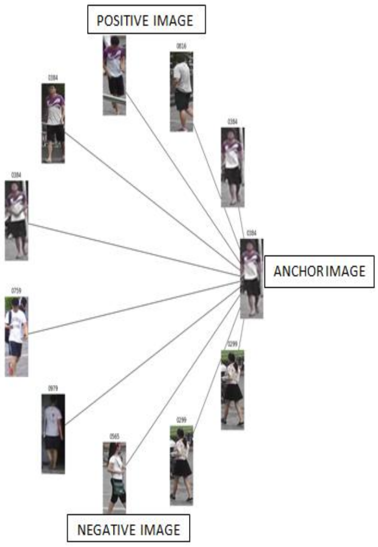


Fig 7 Anchor Images into Positive and Negative images

IV.Conclusion And Future Enhancement

The Siamese Neural Network is used in person re-identification, which involves matching individuals across different camera views. The model is trained using triplet loss, which optimizes the network to learn feature representations that can accurately differentiate between individuals. By fine-tuning the model on large datasets, state-of-the-art performance can be achieved.

However, there are still challenges in this field, such as handling occlusions, pose variations, and changes in appearance over time. The need for labeled data and the computational cost of training these models can be limiting. To address these challenges, future research could focus on developing more robust data preparation methods, selecting alternative loss functions, exploring novel network architectures, and extending the application of Siamese Networks to other domains.

Despite these challenges, person re-identification using Siamese Networks is a rapidly evolving field with many opportunities for further research and development. As new ideas are explored and the technology continues to mature, we can expect even more powerful and accurate models to be developed for this important task.

References

[1] Zhang, J., Zheng, W., Zhang, Z., & Huang, Y. (2019). Person Re-identification: Past, Present and Future. arXiv preprint arXiv:1905.02122.

[2] Dey, S., Kairouz, P., Ramaswamy, S., Sahu, S., Viswanath, P., &Xu, H. (2019, May). Differentially private federated learning: A client level perspective. In 2019 IEEE International Conference on Communications (ICC) (pp. 1-7).IEEE.

[3] Koch, G., Zemel, R., &Salakhutdinov, R. (2015).Siamese neural networks for one-shot image recognition.In ICML deep learning workshop (Vol. 2).

[4] Hoffer, E., Ailon, N., &Lindenbaum, M. (2015). Deep metric learning using Triplet network.arXiv preprint arXiv:1412.6622.

[5] Schroff, F., Kalenichenko, D., &Philbin, J. (2015).FaceNet: A unified embedding for face recognition and clustering.In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 815-823).

[6] S. Paisitkriangkrai, C. Shen, and A. van den Hengel.Learning to rank in person re-identification with metric ensembles.In Proceedings of the IEEE International Conference on Computer Vision, pages 1846–1855, 2015.

[7] R. Arandjelovic and A. Zisserman.All about vlad.In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1578–1585, 2013.

[8] L. Sun, K. Jia, D. Yeung, and B. E. Shi. Human re-identification by matching compositional template with cluster sampling.In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1426–1433, 2014.

[9] Q. Leng, J. Hu, Y. Zhang, and J. Xie.Person re-identification by deep learning attribute-complementary information. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 59–66, 2015.

[10] S. Wu, Y. Wang, Q. Hu, J. Zhang, and J. Luo. Person re-identification using deep feature fusion based on multi-layer similarity aggregation. IEEE Access, 7:131090–131101, 2019.

[11] Yan, Y., Ni, B., Song, Y., Ma, C., Yan, Y., & Yang, X. (2021). A Siamese Network for Person Re-identification with Multi-scale Features. IEEE Access, 9, 56400-56409.

[12] Wu, Z., Zhang, Y., Zhang, K., Zhang, Z., & Li, Y. (2019).Joint discriminative and generative learning for person re-identification. IEEE Transactions on Image Processing, 28(3), 1054-1068.

[13] Zheng, Z., Zheng, L., & Yang, Y. (2016).A discriminatively learned CNN embedding for person reidentification. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 14(1), 1-13.

[14] Hoffer, E., Ailon, N., &Kupnik, M. (2015).Deep metric learning using Triplet network.In International Workshop on Similarity-Based Pattern Recognition (pp. 84-92).Springer, Cham.

[15] Schroff, F., Kalenichenko, D., &Philbin, J. (2015).FaceNet: A unified embedding for face recognition and clustering.In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 815-823).