

Global-local attention with triplet loss and label smoothed cross-entropy for person re-identification

Nha Tran, Toan Nguyen, Minh Nguyen, Khiết Luong, Tai Lam

Faculty of Information Technology, University of Education, Ho Chi Minh City, Viet Nam

Article Info

Article history:

Received Jan 12, 2023

Revised Mar 19, 2023

Accepted Mar 27, 2023

Keywords:

Attention mechanism

Deep features

Multi-level feature

Multi-loss

Person re-identification

ABSTRACT

Person re-identification (Person Re-ID) is a research direction on tracking and identifying people in surveillance camera systems with non-overlapping camera perspectives. Despite much research on this topic, there are still some practical problems that Person Re-ID has not yet solved, in reality, human objects can easily be obscured by obstructions such as other people, trees, luggage, umbrellas, signs, cars, motorbikes. In this paper, we propose a multi-branch deep learning network architecture. In which one branch is for the representation of global features and two branches are for the representation of local features. Dividing the input image into small parts and changing the number of parts between the two branches helps the model to represent the features better. In addition, we add an attention module to the ResNet50 backbone that enhances important human characteristics and eliminates irrelevant information. To improve robustness, the model is trained by combining triplet loss and label smoothing cross-entropy loss (LSCE). Experiments are carried out on datasets Market1501, and duke multi-target multi-camera (DukeMTMC) datasets, our method achieved 96.04% rank-1, 88.11% mean average precision (mAP) on the Market1501 dataset, and 88.78% rank-1, 78.6% mAP on the DukeMTMC dataset. This method achieves performance better than some state-of-the-art methods.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Nha Tran

Faculty of Information Technology, University of Education, Ho Chi Minh City, Viet Nam

280 An Duong Vuong St., Ward 4, District 5, Ho Chi Minh City, Viet Nam

Email: nhatt@hcmue.edu.vn

1. INTRODUCTION

Person re-identification is a research direction on tracking and identifying people in surveillance camera systems with non-overlapping camera perspectives. This is considered an important technology in the field of computer vision and has practical significance in building intelligent monitoring systems. It has many monitoring applications, such as crime detection, personal tracking, and activity analysis [1]–[3]. With the development of deep learning especially convolutional neural networks, person re-ID has made important progress in recent years.

In the past, researchers often used manual methods to solve the problem of a person re-identifying by extracting and selecting features [4]–[7]. Gray and Tao [8] proposed a partition strategy for extracting color and texture features by dividing the image of pedestrians into many horizontal stripes. Other researchers have also utilized more advanced partition methods. The pedestrian image was divided into several triangles to extract features by section in [9]. In [10], hue, saturation, value (HSV) histograms based on body parts, such as the head and trunk, were used to capture spatial information for positioning.

Although much research has been conducted on this topic, there are still some practical problems that have not been fully addressed in the field of person re-identification. One example is the scenario where the

subject to be re-identified is obscured by other objects, which has been studied by [11]. However, with the development of deep learning technology, researchers moved their focus to deep learning techniques, in which models automatically learn pedestrian features. These methods are based on deep learning and are classified into four main groups: Deep metric learning, local feature learning, generative adversarial networks, and Sequence feature learning.

For Deep metric learning method aims to determine the similarity or dissimilarity between two objects (in this case, a pedestrian) by learning distinguishing features from the input image and using a distance function. The distance between two people will have a small value if the same person is similar and vice versa. Deep metric learning is mainly used to constrain the learning process of discriminant features by designing a loss function for the model. Loss functions such as classification loss, verification loss, triplet loss, and contrastive loss are frequently used in research. In [12], triplet loss was used and achieved very good results (state-of-the-art) on both pre-trained and retrained models. The authors emphasized the importance of designing a good triplet loss and encouraged further exploration of the potential of this loss function.

For local feature learning method learns the distinguishing features to perform the classification of two objects as similar or different, these distinguishing features come from local features on the input image. Neural networks are used to automatically find local regions containing important information and then extract distinctive features from these regions. The local features in this method must be related to each other to ensure the accuracy of the output. Commonly used local feature learning methods are predefined stripe segmentation, multi-scale fusion, soft attention, pedestrian semantic extraction, and global-local feature learning [13]–[16]. A network called the part-based convolutional baseline (PCB) was proposed in [15]. The input image was divided into p parts evenly ($p=6$) and each classifier predicted the identity of the input image using Cross-Entropy loss as a supervision signal during training. During testing, either p pieces are concatenated to form the final descriptor of the input image. A strategy of learning discriminant features based on many details through the application of a global-local learning method was proposed in [17]. The author has designed the multiple granularity network (MGN) as a multi-branch deep learning network architecture in which one branch is for the representation of global features and two for the representation of local features. In the two branches of learning local features, the author uses the uniform partitioning method to divide the input image into many parts (horizontal division) and the diversity in the number of parts in different branches help the model represent the features in many details. This method has achieved a high performance and surpasses previous methods. In [18], the relationship between local features within an image and the relationship between features in different images was investigated. Local feature histograms were used and the attention mechanism was employed during training to assign varying levels of importance to different features.

For generative adversarial networks (GAN) [19] was introduced by Ian Goodfellow and developed rapidly in recent years. The main application of GAN is to generate new images for the purpose of expanding the dataset [20]. In person re-identification some researchers use GAN to generate new human images with differences in appearance, posture, even lighting contrast and resolution. The CycleGAN technique [21] has been utilized by some researchers to transform the style of images between different datasets. This inspired the proposal of PTGAN [22], a method that transforms the style of images in the domain of one dataset to the domain of another dataset while preserving the identity of the people in the original domain. The aim of this method is to convert the background and brightness of the original domain to the remaining domain in order to increase the diversity of the data. An adversarial network for hard triplet generation (HTG) was proposed in [23] in order to optimize the network's ability.

Sequence Feature Learning is used by some researchers to extract the important information contained in video sequences. This method takes a short video as input and uses both spatial and temporal cues to identify the object. recognizing the lack of spatial and temporal constraints in many person re-identification methods, a method that enables the extraction of both spatial and temporal information was proposed in [24]. The main idea here is that when a person appears in the frame of one camera, for some time t (t is a small value) cannot appear in the frame of another camera. Due to temporal constraints, this method has helped to eliminate many misidentified human images.

In this study, we propose a multi-branch learning network, in order to represent global features and represent local features, the multi-branch model has been increasingly widely used in recent years [25]–[29]. Part-level feature learning is beneficial for learning the discriminatory re-ID model. Global feature learning learns representations over the whole image with no part constraints [30]. It is discriminatory when tracking someone who can pinpoint the exact location of the human body. When the human image is obscured, the learning part-level feature often achieves better performance by exploiting discriminatory body regions [14]. Because of its advantage in occlusion processing, we observe that most of the recently developed state-of-the-art methods adopt the model of feature aggregation, partial-level matching, combining feature-level matching and full human body features [28], [31].

Besides, designing the loss function is very important during training, the common approach is to use a combination of ID loss and metric loss. Our main contributions are:

- Multi-branch with convolutional block attention module (CBAM) [32] is attached to the backbone to help the model learn more detailed features. The features are then automatically grouped into sub-features, each of which helps to narrow the search space of the recognition target.
- Combination triple loss and smooth cross entropy loss to help the network learn more efficiently to improve performance.

2. METHOD

The proposed method is a network named GLML as shown in Figure 1 based on MGN [17] consisting of many branches including one branch for the representation of global features and two for the representation of local features. In the two branches of learning local features, using the homogenous partitioning method to divide the input image into several horizontal parts, and the diversity of the number of parts in different branches helps the model represent are featured in many details. The first part represents the global feature learning branch, which uses the input original image and feature extraction on it. The middle part (Branch 2) and the last part (Branch 3) represent two branches of learning local features, dividing the input image into many small parts and changing the number of parts between the two branches helps the model can represent better features. In addition, a CBAM attention module adds to the ResNet50 [33] backbone to enhance important human features and remove irrelevant information. To improve robustness, the model is trained by combining triplet loss and label smoothing cross-entropy loss (LSCE). The common loss function is constructed based on LSCE loss and triplicate loss.

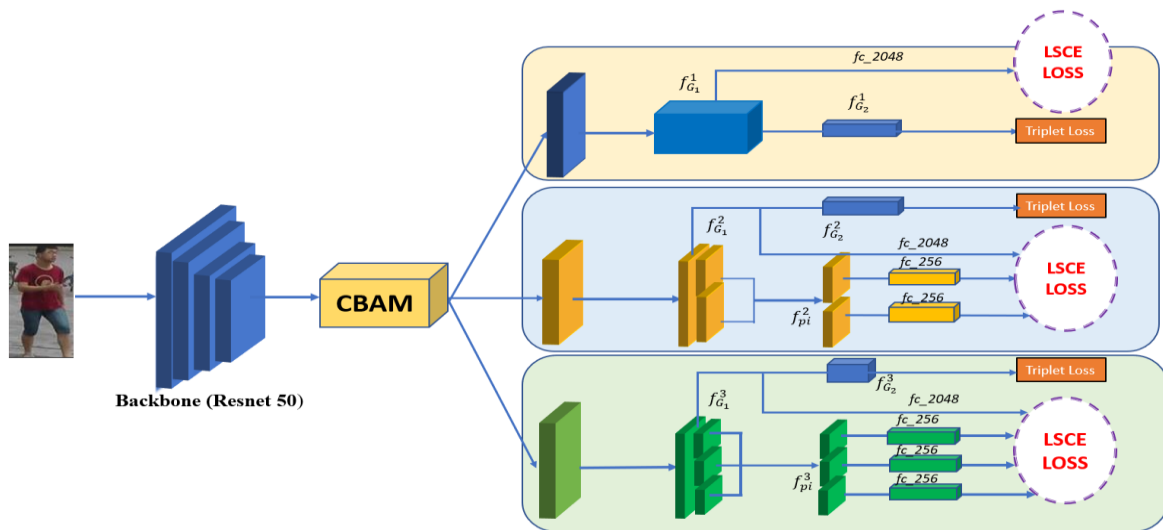


Figure 1. Multi-branch with CBAM and multi loss

Algorithm 1 The MLML Algorithm

Input: A fixed-size mini-batch consisting of $P = 4$ randomly selected identities and $K = 4$ randomly selected images per identity from the training set.

Output: Tensors of feature extraction and reID task.

Initialization: A fixed-size N of epochs.

Repeat

for each epoch $i = 1$ to N do

for each iteration to max iterations do

 Backpropagate CNN and evaluate the loss L according to (1)

end for

if $i \% 50 = 0$:

 Start evaluate;

 Save the model with lowest loss L

end if

end for

Until maximum epochs ($N = 800$) reached.

To make a fast convergence of the training models, the Adam optimizer is adopted. The parameters of Adam have a weight decay of $5e-4$, we set the initial learning rate to $2e-4$, and decay the learning rate to $2e-5$ and $2e-6$ after training for 360 and 720 epochs. The total training process lasts for 800 epochs. To extract the discriminative features for person re-identification, ResNet-50 as the backbone network was used. The batch size of 16 was chosen based on the available hardware resources and the trade-off between training speed and stability. The Pytorch platform with T4 graphical processing unit (GPU) was adopted in our work. The images were resized to 384×128 and subjected to random erasing and horizontal flipping with a probability of 0.5 in order to augment the training data and improve the robustness of the model. We also followed the approach outlined in [34] to improve the performance of the model.

2.1. Attention module

The attention module helps the model learn and focus more on important information than on learning unhelpful background information. We added the CBAM block to the backbone, creating the model shown in Figure 2. the CBAM uses two attention modules: the channel attention module, as shown in Figure 3, and the spatial attention module shown in Figure 4. The envelope channel attention module consists of two feature maps, each consisting of two intermediate layers: average and maximum pooling. Both feature maps are combined by a shared multilayer perceptron layer (MLP), then the output of the feature map is added using the sigmoid activation function. Finally, the multiplicative features between the convolutional layer and the channel attention module were applied to the spatial attention module to determine the positions of the most important features in the given image.

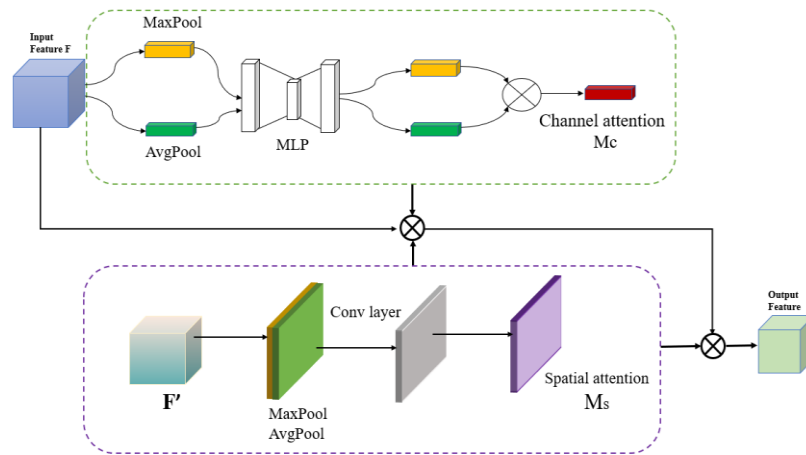


Figure 2. Convolutional block attention module architecture

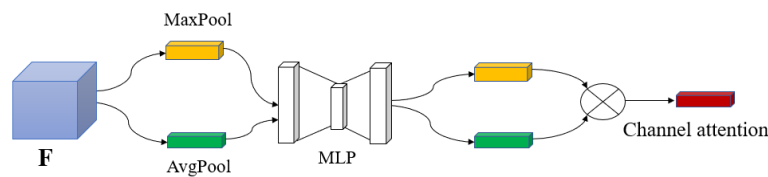


Figure 3. Channel attention module

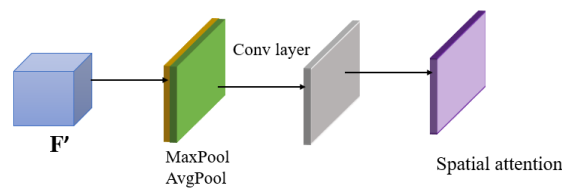


Figure 4. Spatial attention module

2.2. Loss functions

To optimize the distance feature of each class in our model, we employed the triplet loss during training, while softmax cross-entropy loss was utilized to capture the classification difference in each class. The overall loss of our proposed method can be expressed as the sum of triplet loss and softmax cross-entropy loss, as shown in (1). However, most of the studies apply the batch-hard triplet loss [12], [35], an improved version based on the original semi-hard triplet loss, as shown in (2).

$$L_s = L_{triplet} + L_{lsce} \quad (1)$$

$$\mathcal{L}_{triplet} = \sum_{k=1}^{N_K} \sum_{m=1}^{N_M} [\alpha + \max_{n=1 \dots M} \|q_{k,m}^A - q_{k,n}^P\|_2 - \min_{\substack{l=1 \dots K \\ l \neq k}} \|q_{k,m}^A - q_{l,n}^N\|_2]_+ \quad (2)$$

In our model, for the purpose of classifier learning, we have adopted the softmax loss function. This loss function is widely used for multiclass classification problems and is mathematically expressed as the negative logarithm of the softmax function. Specifically, the softmax loss function computes the difference between the predicted probability distribution and the true probability distribution of the classes, as shown in (3).

$$\mathcal{L}_{softmax} = - \sum_{i=1}^N \log \frac{e^{W_{yi}^T + f_i + b_{yi}}}{\sum_{k=1}^C e^{W_k^T + f_i + b_k}} \quad (3)$$

To enhance the robustness of our model and prevent overfitting, we employed the technique of Label Smoothing [36]. This method modifies the target distribution by allocating some probability mass to non-target classes during training, as defined by (4). The soft-margin ε is used to reduce model overconfidence. In our experiments, we set ε to 0.2.

$$\mathcal{L}_{LSCE} = \sum_{i=1}^N -q_i \log(p_i) \begin{cases} q_i = 0, y \neq i \\ q_i = 1, y = i \end{cases} \quad (4)$$

$$q_i = \begin{cases} 1 - \frac{N-1}{N} \varepsilon & \text{if } i = y \\ \varepsilon / N & \text{otherwise} \end{cases}$$

3. RESULTS AND DISCUSSION

3.1. Datasets

Market-1501 [37]: This large data set was published in 2015, Market-1501 was collected by 5 high-resolution cameras and 1 low-resolution camera in front of a supermarket at Tsinghua University, China. The dataset contains 1501 different people with a total of 32668 images. Compared to CUHK03 [38], Market-1501 has more images and contains many confounding factors (images of people are obscured, photos only show part of the body, ...), so this dataset is evaluated as closer to reality than CUHK03.

DukeMTMC-reID [39]: dataset collected at Duke University, USA through 8 HD still cameras. DukeMTMC-reID includes a training set containing 16522 images of 702 different people, a query set of 2228 images of 702 other people (different from the training set), and a gallery of 17661 images. Both datasets include bounding box annotations and additional metadata such as camera IDs and timestamps and are commonly used to evaluate the performance of person re-identification algorithms using standard evaluation metrics such as rank-1 accuracy and mean average precision (mAP). The Market1501 and DukeMTMC datasets have been widely used in a variety of person re-identification tasks, including cross-camera person search, video-based person re-identification, multi-camera tracking, and pedestrian attribute recognition.

3.2. Result and discussion

Table 1 demonstrates that our model obtained the highest accuracy rates of 87.77% and 88.11% on the Market-1051 dataset. For the DukeMTMC-reID dataset, we achieved a mAP of 76.97% and 78.60%. These results indicate that utilizing an attention mechanism enhances significant human features while filtering out irrelevant information. Furthermore, combined triplets' loss, and smooth lable softmax cross-entropy loss help the network to learn more effectively.

In order to gain a better understanding of the results obtained by our proposed model. We have compared it with several other methods using two datasets: Market – 1051 and DukeMTMC-reID. Table 2 and Table 3 shows that our proposed method has achieved comparable performance to some of the state-of-the-art methods developed in recent years.

Table 1. The results of different methods

Methods	CBAM	Re-rank	Market-150				DukeMTMC			
			R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP
			95.07	98.31	98.49	87.77	88.55	93.63	96.27	76.97
GLML	x		95.04	98.22	99.17	88.11	88.78	94.75	96.41	78.60
		x	96.02	98.10	98.49	94.71	91.34	95.15	96.59	89.41
	x	x	96.29	98.28	98.78	94.82	91.79	95.47	96.72	90.08

Table 2. Comparison of the results of different methods in the market1501 dataset

Methods	R-1	R-5	R-10	mAP	Ref
MGN [17]	95.7	98.3	99.0	86.9	ACMMM2018
PTL + MGN [40]	94.83	-	-	87.34	IJCAI19
IANet [41]	94.40	-	-	83.10	CVPR19
CAR [42]	96.10	-	-	84.70	ICCV19
SAN [43]	96.10	-	-	88.00	AAAI20
DLBC [44]	94.60	98.4	99.1	87.40	ACM MM20
GPS [45]	95.2	98.4	99.1	87.8	CVPR 2021
MSLG [46]	90	97	98	71	Soft Computing 2022
CASN+PCB [47]	94.4	-	-	82.8	CVPR19
MGN + Re-rank [17]	96.6	-	-	94.2	ACM Multimedia 2018
DCDS + Re-rank [48]	95.40	98.3	-	93.30	ICCV19
Auto-ReID + Re-rank [31]	95.40	-	-	94.20	ICCV19
Ours	95.04	98.22	99.17	88.11	
Ours + Re-rank	96.29	98.28	98.78	94.82	

Table 3. Comparison of the results of different methods in the DukeMTMC-reID dataset

Methods	R-1	R-5	R-10	mAP	Ref
MGN [17]	88.70	-	-	78.40	ACMMM2018
IANet [41]	87.10	-	-	73.40	CVPR19
CAR [42]	86.30	-	-	73.10	ICCV19
SAN [43]	87.90	-	-	75.50	AAAI20
DLBC [44]	88.70	94.90	96.60	78.50	ACM MM20
GPS [45]	88.20	95.20	96.70	78.70	CVPR 2021
MSLG [46]	82.00	83.00	91.00	65.00	Soft Computing 2022
CASN+PCB [47]	87.70	-	-	73.70	CVPR19
DCDS + Re-rank [48]	88.50	-	-	86.10	ICCV19
Ours	88.78	94.75	96.41	78.60	
Ours + Re-rank	91.79	95.47	96.72	90.08	

4. CONCLUSION

In this paper, we propose a multi-branch deep learning network architecture, consisting of one branch representing global features and two branches representing local features. By dividing the input image into smaller parts and adjusting the number of parts between the two branches, the model can better capture the features of the image. Furthermore, to enhance the robustness of the model, we combine Triplet Loss and LSCE Loss during training to optimize the feature distance between each class. We also incorporate an attention mechanism into the ResNet50 backbone to enhance important human traits and remove irrelevant information, it can improve performance on mAP measurement compared to the MGN method and achieve performance better than some state-of-the-art methods.

REFERENCES




- [1] Chen Change Loy, T. Xiang, and S. Gong, "Multi-camera activity correlation analysis," pp. 1988–1995, 2010, doi: 10.1109/cvpr.2009.5206827.
- [2] X. Wang, "Intelligent multi-camera video surveillance: A review," *Pattern Recognition Letters*, vol. 34, no. 1, pp. 3–19, 2013, doi: 10.1016/j.patrec.2012.07.005.
- [3] S. I. Yu, Y. Yang, and A. Hauptmann, "Harry potter's marauder's map: Localizing and tracking multiple persons-of-interest by nonnegative discretization," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3714–3720, 2013, doi: 10.1109/CVPR.2013.476.
- [4] L. An, X. Chen, S. Liu, Y. Lei, and S. Yang, "Integrating appearance features and soft biometrics for person re-identification," *Multimedia Tools and Applications*, vol. 76, no. 9, pp. 12117–12131, 2017, doi: 10.1007/s11042-016-4070-2.
- [5] H. M. Hu, W. Fang, G. Zeng, Z. Hu, and B. Li, "A person re-identification algorithm based on pyramid color topology feature," *Multimedia Tools and Applications*, vol. 76, no. 24, pp. 26633–26646, 2017, doi: 10.1007/s11042-016-4188-2.
- [6] R. Zhao, W. Ouyang, and X. Wang, "Person re-identification by saliency matching," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2528–2535, 2013, doi: 10.1109/ICCV.2013.314.

- [7] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-January, pp. 3652–3661, 2017, doi: 10.1109/CVPR.2017.389.
- [8] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5302 LNCS, no. PART 1, pp. 262–275, 2008, doi: 10.1007/978-3-540-88682-2_21.
- [9] N. Gheissari, T. B. Sebastian, P. H. Tu, J. Rittscher, and R. Hartley, "Person reidentification using spatiotemporal appearance," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 1528–1535, 2006, doi: 10.1109/CVPR.2006.223.
- [10] A. Das, A. Chakraborty, and A. K. Roy-Chowdhury, "Consistent re-identification in a camera network," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8690 LNCS, no. PART 2, pp. 330–345, 2014, doi: 10.1007/978-3-319-10605-2_22.
- [11] J. Zhuo, Z. Chen, J. Lai, and G. Wang, "Occluded person re-identification," *Proceedings - IEEE International Conference on Multimedia and Expo*, vol. 2018-July, 2018, doi: 10.1109/ICME.2018.8486568.
- [12] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," Mar. 2017, [Online]. Available: <http://arxiv.org/abs/1703.07737>.
- [13] Y. Fu *et al.*, "Horizontal pyramid matching for person re-identification," *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, pp. 8295–8302, 2019, doi: 10.1609/aaai.v33i01.33018295.
- [14] Y. Sun *et al.*, "Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2019-June, pp. 393–402, 2019, doi: 10.1109/CVPR.2019.00048.
- [15] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11208 LNCS, pp. 501–518, 2018, doi: 10.1007/978-3-030-01225-0_30.
- [16] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang, "A siamese long short-term memory architecture for human re-identification," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9911 LNCS, pp. 135–153, 2016, doi: 10.1007/978-3-319-46478-7_9.
- [17] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," *MM 2018 - Proceedings of the 2018 ACM Multimedia Conference*, pp. 274–282, 2018, doi: 10.1145/3240508.3240552.
- [18] Z. Zhang, H. Zhang, and S. Liu, "Person re-identification using heterogeneous local graph attention networks," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 12131–12140, 2021, doi: 10.1109/CVPR46437.2021.01196.
- [19] I. Goodfellow *et al.*, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020, doi: 10.1145/3422622.
- [20] N. T. and H. N. Tai Lam, Tinh Nguyen, "A system for design of handbag using generative adversarial networks," *International Journal of Advanced Engineering*, vol. 4, no. 2, pp. 93–100, 2021.
- [21] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2017-October, pp. 2242–2251, 2017, doi: 10.1109/ICCV.2017.244.
- [22] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer GAN to bridge domain gap for person re-identification," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 79–88, 2018, doi: 10.1109/CVPR.2018.00016.
- [23] Y. Zhao, Z. Jin, G. jun Qi, H. Lu, and X. sheng Hua, "An adversarial approach to hard triplet generation," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11213 LNCS, pp. 508–524, 2018, doi: 10.1007/978-3-030-01240-3_31.
- [24] G. Wang, J. Lai, P. Huang, and X. Xie, "Spatial-temporal person re-identification," *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, pp. 8933–8940, 2019, doi: 10.1609/aaai.v33i01.33018933.
- [25] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2017-October, pp. 3980–3989, 2017, doi: 10.1109/ICCV.2017.427.
- [26] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian, "GLAD: Global-local-alignment descriptor for scalable person re-identification," *IEEE Transactions on Multimedia*, vol. 21, no. 4, pp. 986–999, 2019, doi: 10.1109/TMM.2018.2870522.
- [27] F. Zheng *et al.*, "Pyramidal person re-identification via multi-loss dynamic training," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2019-June, pp. 8506–8514, 2019, doi: 10.1109/CVPR.2019.000871.
- [28] X. Chen *et al.*, "Saliency-guided cascaded suppression network for person re-identification," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3297–3307, 2020, doi: 10.1109/CVPR42600.2020.00336.
- [29] H. Yao, S. Zhang, R. Hong, Y. Zhang, C. Xu, and Q. Tian, "Deep representation learning with part loss for person re-identification," *IEEE Transactions on Image Processing*, vol. 28, no. 6, pp. 2860–2871, 2019, doi: 10.1109/TIP.2019.2891888.
- [30] H. Luo *et al.*, "A strong baseline and batch normalization neck for deep person re-identification," *IEEE Transactions on Multimedia*, vol. 22, no. 10, pp. 2597–2609, 2020, doi: 10.1109/TMM.2019.2958756.
- [31] R. Quan, X. Dong, Y. Wu, L. Zhu, and Y. Yang, "Auto-reID: Searching for a part-aware convnet for person re-identification," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2019-October, pp. 3749–3758, 2019, doi: 10.1109/ICCV.2019.00385.
- [32] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11211 LNCS, pp. 3–19, 2018, doi: 10.1007/978-3-030-01234-2_1.
- [33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-December, pp. 770–778, 2016, doi: 10.1109/CVPR.2016.90.
- [34] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, vol. 2019-June, pp. 1487–1495, 2019, doi: 10.1109/CVPRW.2019.00190.




- [35] S. Liao and S. Z. Li, "Efficient PSD constrained asymmetric metric learning for person re-identification," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2015 International Conference on Computer Vision, ICCV 2015, pp. 3685–3693, 2015, doi: 10.1109/ICCV.2015.420.
- [36] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-December, pp. 2818–2826, 2016, doi: 10.1109/CVPR.2016.308.
- [37] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2015 International Conference on Computer Vision, ICCV 2015, pp. 1116–1124, 2015, doi: 10.1109/ICCV.2015.133.
- [38] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 152–159, 2014, doi: 10.1109/CVPR.2014.27.
- [39] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9914 LNCS, pp. 17–35, 2016, doi: 10.1007/978-3-319-48881-3_2.
- [40] Z. Yu *et al.*, "Progressive transfer learning for person re-identification," *IJCAI International Joint Conference on Artificial Intelligence*, vol. 2019-August, pp. 4220–4226, 2019, doi: 10.24963/ijcai.2019/586.
- [41] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "Interaction-and-aggregation network for person RE-identification," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2019-June, pp. 9309–9318, 2019, doi: 10.1109/CVPR.2019.00954.
- [42] S. Zhou, F. Wang, Z. Huang, and J. Wang, "Discriminative feature learning with consistent attention regularization for person re-identification," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2019-October, pp. 8039–8048, 2019, doi: 10.1109/ICCV.2019.00813.
- [43] X. Jin, C. Lan, W. Zeng, G. Wei, and Z. Chen, "Semantics-aligned representation learning for person re-identification," *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*, pp. 11173–11180, 2020, doi: 10.1609/aaai.v34i07.6775.
- [44] J. Chen *et al.*, "Deep local binary coding for person re-identification by delving into the details," *MM 2020 - Proceedings of the 28th ACM International Conference on Multimedia*, pp. 3034–3043, 2020, doi: 10.1145/3394171.3413979.
- [45] B. X. Nguyen, B. D. Nguyen, T. Do, E. Tjiputra, Q. D. Tran, and A. Nguyen, "Graph-based person signature for person re-identifications," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 3487–3496, 2021, doi: 10.1109/CVPRW53098.2021.00388.
- [46] J. Liu, P. Tiwari, T. G. Nguyen, D. Gupta, and S. S. Band, "Multi-scale local-global architecture for person re-identification," *Soft Computing*, vol. 26, no. 16, pp. 7967–7977, 2022, doi: 10.1007/s00500-022-06859-6.
- [47] M. Zheng, S. Karanam, Z. Wu, and R. J. Radke, "Re-identification with consistent attentive siamese networks," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2019-June, pp. 5728–5737, 2019, doi: 10.1109/CVPR.2019.00588.
- [48] L. T. Alemu, M. Shah, and M. Pelillo, "Deep constrained dominant sets for person re-identification," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2019-October, pp. 9854–9863, 2019, doi: 10.1109/ICCV.2019.00995.

BIOGRAPHIES OF AUTHORS






Nha Tran    received BS. Computer Science (2014), MS. Computer Science (2019) from HCM University of Education. He is currently a lecturer of Information Technology Faculty, Ho Chi Minh city University of Education. His research interest includes computer vision, information retrieval, affective computing, machine learning. You can contact him via email: nhatt@hcmue.edu.vn.






Toan Nguyen    received his B.Sc. (Computer Science) from Ho Chi Minh City University of Education in 2022. His research includes machine learning, deep learning, computer vision, image processing. He can be contacted at email: toannn20@gmail.com.






Minh Nguyen    He is currently a 3rd-year student at the Ho Chi Minh City University of Education, Ho Chi Minh City, Vietnam. His research includes machine learning, deep learning, and computer vision. He can be contacted at email: nguyendatminhvn@gmail.com.



Khiet Luong    received BS. Information Technology (2016), MS. Computer Science (2018). He is currently a lecturer at the University of Education, Ho Chi Minh City, Vietnam. Concurrently, his research interests include computer vision, Software engineering, and machine learning. You can contact him via email: khietltn@hcmue.edu.vn.



Tai Lam    received his B.Sc. (Computer Science) from Ho Chi Minh City University of Education in 2022. He is currently an AI engineer at Emage Development company. His research includes machine learning, deep learning, computer vision. You can contact him via email: lamtai2105@gmail.com.