Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Paying Attention to the Features Extracted from the Image to Person Re-Identification

S. H. Zahiri* *, R. Iranpoor, N. Mehrshad

Department of Electrical Engineering, Faculty of Engineering, University of Birjand, Birjand, Iran.

Article Info

Article History:
Received 01 July 2024
Reviewed 29 July 2024
Revised 26 September 2024
Accepted 10 October 2024

Keywords:
Person re-identification
Deep learning
Image processing
Convolutional neural network
Computer vision
Image detection

*Corresponding Author's Email Address:
hzahiri@birjand.ac.ir

Abstract

Background and Objectives: Person re-identification is an important application in computer vision, enabling the recognition of individuals across non-overlapping camera views. However, the large number of pedestrians with varying appearances, poses, and environmental conditions makes this task particularly challenging. To address these challenges, various learning approaches have been employed. Achieving a balance between speed and accuracy is a key focus of this research. Recently introduced transformer-based models have made significant strides in machine vision, though they have limitations in terms of time and input data. This research aims to balance these models by reducing the input information, focusing attention solely on features extracted from a convolutional neural network model.

Methods: This research integrates convolutional neural network (CNN) and Transformer architectures. A CNN extracts important features of a person in an image, and these features are then processed by the attention mechanism in a Transformer model. The primary objective of this work is to enhance computational speed and accuracy in Transformer architectures.

Results: The results obtained demonstrate an improvement in the performance of the architectures under consistent conditions. In summary, for the Market-1501 dataset, the mAP metric increased from approximately 30% in the downsized Transformer model to around 74% after applying the desired modifications. Similarly, the Rank-1 metric improved from 48% to approximately 89%.

Conclusion: Indeed, although it still has limitations compared to larger Transformer models, the downsized Transformer architecture has proven to be much more computationally efficient. Applying similar modifications to larger models could also yield positive effects. Balancing computational costs while improving detection accuracy remains a relative goal, dependent on specific domains and priorities. Choosing the appropriate method may emphasize one aspect over another.

Introduction

Recent advancements in deep learning techniques, coupled with increased computational power, have significantly improve the challenges associated with object identification and recognition in images. Object and person detection remain critical issues within the field of computer vision.

While object recognition is intuitive for humans (even a few-month-old child can recognize common objects), teaching computers to achieve the same level of proficiency has been a formidable challenge until the past decade [1]. The resurgence of Convolutional Neural Networks (CNNs) and deep learning for image classification has revolutionized visual perception. The

adoption of CNNs in the large-scale ImageNet Visual Recognition Challenge (ILSVRC) in 2012 by AlexNet [2] inspired further research on its applications in computer vision. Today, object detection finds applications in self-driving cars, identity verification, security, and medical contexts. In recent years, exponential growth in this field has occurred due to rapid advancements in tools and techniques.

Fig. 1: Sample images of datasets used
(a) Market-1501 dataset (b) DukeMTMC dataset (c) MSMT17 dataset.

As the subsequent challenge following general object and person detection in a scene, the problem of person re-identification (ReID) emerges. Due to its diverse applications across various domains, ReID has garnered significant attention. It serves as a fundamental and essential function in intelligent surveillance systems. The task of connecting individuals across different cameras in various locations and time frames is crucial for network-based surveillance systems. ReID is recognized as the problem of identifying individuals and forms the basis for many other important applications [3].

Current research efforts to address the ReID challenge primarily focus on two aspects: feature learning and metric learning.

Feature Learning: Developing feature representations that remain discriminative for identity while being invariant to viewpoint and lighting conditions [4].

Metric Learning: Optimizing the discriminative parameters of a ReID model using machine learning techniques [5].

In some real-world situations or images with certain limitations, the human eye may not be able to identify and distinguish the subject. In such cases, it is possible to proceed by relying on some minor features or states. In these instances, there is no need to focus on the entire image; rather, relying on specific features can be effective in achieving the desired goal more quickly and reliably.

In this work, the goal is to utilize Transformer models. These models have demonstrated significant results across various domains. However, they require large amounts of input data and powerful hardware for training. To address these challenges, we employ a convolutional neural network (CNN) as a feature extractor, using pre-trained models. Subsequently, the extracted features are fed into a smaller Transformer model, enabling attention at the feature level. This approach allows us to increase the input data to the Transformer model and reduce the data volume by leveraging features extracted from an image. Therefore, the combination of CNNs and Transformer-based models forms the foundation of our research in computer vision.

Related Works

Research in this field primarily focuses on person re-identification and object recognition, with most methods based on Convolutional Neural Networks (CNNs). A desirable and suitable approach for person re-identification involves designing an appropriate loss function for training a backbone CNN (such as ResNet [6]) used for feature extraction from images. Cross-entropy loss [7] and triplet loss [8] are commonly employed in person re-identification.

The IDE model specified in [9] is a global descriptor. For example in [10], the IDE network fine-tuned on the R-CNN model [11] is proved to be more effective than the one fine-tuned directly on an ImageNet pre-trained model. In many cases the IDE model is a commonly used baseline in deep re-ID systems.

Researchers like Luo et al. [12] proposed BNNeck to better combine cross-entropy and triplet loss functions. The main focus of the study by Sun et al. [13] was to obtain superior features using a Part-based Convolutional Neural Network (PCB). In this approach, a convolutional descriptor of the input image captures features related to different parts of the image. An improved method for part extraction (RPP) is introduced [14]. Additionally, another work presents an integrated perspective on cross-entropy and triplet loss functions [15].

Methods such as PCB [13], MGN [16], AlignedReID [17], SAN [18], and others divide the image into multiple parts and extract local features for each part. These fine-grained features are then used for information aggregation.

The Transformer model [19], originally proposed for sequential data in natural language processing (NLP), has

also shown effectiveness in computer vision tasks. Han et al. [20] and Salman et al. [21] have explored the application of Transformers in computer vision. The Vision Transformer (ViT) [22] directly utilizes pure Transformers on image patches. However, ViT requires large-scale pre-training data. To address this limitation, the DeiT framework introduces a teacher-student strategy specifically for Transformers to accelerate ViT training without the need for large-scale pre-training data [23].

He et al. [24] proposed a pure transformer-based framework for person ReID named TransReID. Specifically, they first encode an image as a sequence of patches and build a transformer-based strong baseline with a few critical improvements, which achieves competitive results on several ReID benchmarks with CNN-based methods.

Methods

Feature extraction and attention to the extracted features are fundamental to this research. To evaluate the implemented models, we must first examine the commonly used datasets in this field. Research in deep learning technology heavily relies on substantial amounts of data for model training. Subsequently, we will investigate the prevalent backbone architectures.

Datasets

To develop robust person re-identification models, it is essential to have datasets with diverse backgrounds, occlusions, and overlapping bodies [25]. While numerous datasets are available for research, some, such as VIPeR [26], GRID [27], and CUHK01 [28], are limited by the number of individuals and the small number of images per person. These datasets often rely on manual labeling methods for person identification. With the advancement of deep learning, smaller datasets are no longer sufficient for training needs. Consequently, large-scale datasets such as CUHK03 [29], Market1501 [30], DukeMTMC [31], and MSMT17 [32] have been proposed and accepted.

The Market-1501 dataset is one of the most well-known datasets for person detection and identification in images. It comprises 1501 different individuals captured in an outdoor environment, with approximately 32,217 images collected from 6 surveillance cameras equipped with various sensors. Each individual has around 6 to 20 full-body images. The dataset is divided into training and evaluation (query and gallery) sets. The training set includes images of the first 751 individuals, with approximately 12 images per person. The test set contains another 750 individuals, each having only one query image and around 4 to 18 gallery images.

The MSMT17 dataset consists of images from a 15-camera network deployed on a university campus. The camera network includes 12 outdoor cameras and 3 indoor cameras. Video collection was conducted over four days with varying weather conditions. For each day, 3 hours of footage were captured, focusing on pedestrian detection and annotation during morning, noon, and afternoon. The final raw video dataset comprises 180 hours of footage, 12 outdoor cameras, 3 indoor cameras, and 12 temporal gaps. Faster RCNN [33] was used for pedestrian bounding box detection, resulting in 126,441 annotated bounding boxes from 4,101 identities. Sample images from all three datasets are shown in Fig. 1.

The DukeMTMC dataset consists of over 2 million frames and more than 2,700 individuals. It includes eight videos, each lasting 85 minutes, recorded at 1080p quality with a frame rate of 60 frames per second. The videos were captured by eight fixed cameras placed around the Duke University campus during periods of heavy pedestrian foot traffic. Calibration data was used to determine homography between images and ground level.

Table 1: Specifications of the datasets used provides information about the datasets used, including the number of training images, the count of individuals in the training set, and the number of evaluation images, which includes query and gallery images.

Table 1: Specifications of the datasets used

Dataset	Train image	Quary image	Gallery image	Num train ID	Num Cam
Market-1501	12936	3368	15913	751	6
DukeMTMC	16522	2228	17661	702	8
MSMT17	32621	11659	82161	1041	15

Backbone Networks

Backbone networks function as initial feature extractors for object recognition and person re-identification tasks. These networks process images as input and generate corresponding feature maps as output. Most of these architectures, originally designed for object detection, typically exclude fully connected layers. Additionally, advanced versions of classification networks are available.

Considering the diverse requirements for accuracy and efficiency, individuals can opt for deeper and more compact architectures such as ResNet [34], ResNeXt [35], or lightweight networks like MobileNet [36], ShuffleNet [37], SqueezeNet [38], Xception [39], MobileNetV2 [40], and MobileNetV3 [41]. When targeting mobile devices, lightweight networks effectively meet the necessary criteria.

MobileNetV1 [36] introduced depthwise separable convolutions as an efficient replacement for traditional

convolution layers.Depthwise separable convolutions are defined by two separate layers: light weight depthwise convolution for spatial filtering and heavier 1x1 pointwise convolutions for feature generation. This method effectively factorize traditional convolution by separating spatial filtering from the feature generation mechanism.

MobileNetV2 [40] introduced the linear bottleneck and inverted residual structure in order to make even more efficient layer structures by leveraging the low rank nature of the problem. MobileNetV3 [41] use a combination of these layers as building blocks in order to build the most effective models. Layers are also upgraded with modified swish nonlinearities. Both squeeze and excitation as well as the swish nonlinearity use the sigmoid which can be inefficient to compute as well challenging to maintain accuracy in fixed point arithmetic so it is replaced by the hard sigmoid.

The nonlinearity is defined as

$$swish\ x = x\ .\sigma(x) \tag{1}$$

A nonlinearity called swish was introduced that when used as a drop-in replacement for ReLU, that significantly improves the accuracy of neural networks [42]-[44].

In MobileNetV3 the sigmoid function has been replaced with its piece-wise linear hard analog. The difference is in use ReLU6 rather than a custom clipping constant. Similarly, the hard version of swish becomes

$$h - swish\ [x] = x\ \frac{ReLU6(x+3)}{6} \tag{2}$$

MobileNetV3 is defined as two models: MobileNetV3-Large and MobileNetV3-Small. These models are targeted at high and low resource use cases respectively. In this work, according to Fig. 2, in backbone section we use both MobileNetV3 versions for feature extraction.

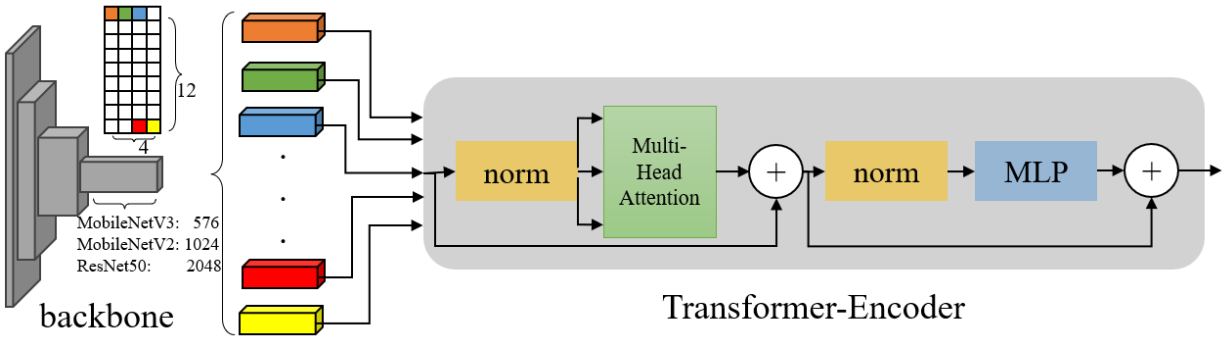

Fig. 2: Attention to the features extracted from the image in the proposed method.

Fig. 3: illustrates the output heatmap of the backbone network, serving as an example of feature extraction in the backbone architecture.

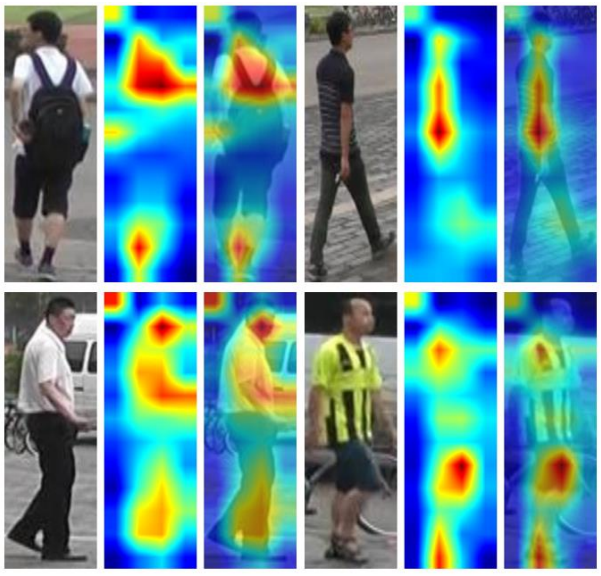

Fig. 3: MobileNetV3 extracted heatmap from the Market-1501 dataset.

Proposed Methode

In this section, the focus is on exploring efforts in person re-identification (ReID) using transformer-based methods. One of the challenges faced by this approach is its high computational cost. Therefore, achieving a balance between computational overhead and task accuracy is the primary goal. In transformer-based architectures, input images are divided into smaller patches, each treated as a patch and fed into the main network. For instance, in person re-identification, if we consider an input image with dimensions of 128x384 and a patch extraction window of 16x16, a total of 8x24 patches will be separated for input to the main network. Subsequently, 192 patches are formed, each with dimensions of 3x16x16, resulting in an array of 768x192.

The objective in this approach is to reduce the input information to the main network. Many details within an image, such as backgrounds, do not require attention mechanisms and are essentially unused information. Since networks like ViT attend to all input information, this increases computational costs. Pure transformer-based models (e.g. ViT, DeiT) split the images into non-overlapping patches, losing local neighboring structures

around the patches.

In this job, the extracted features are used instead of the non-Overlapping patches [24]. As depicted in Fig. 2, the input image is first fed into a column-based network. Depending on the functionality and complexity, various architectures can be used. For the initial step and for speed enhancement, a downsized MobileNetV3 model is employed. The output of this model, after the main layers, is a tensor of size 4x12x576. Thus, ultimately, an array of 48x576 is transferred to the main Transformer network.

In the self-attention layer, the input vector is first transformed into three different vectors: the query vector q, the key vector k. Vectors derived from different inputs are then packed together into three different matrices, namely, Q, K and V [20].

Subsequently, the attention function between different input vectors is computed by calculating scores between them using the formula $Sn = Q.K^T$. These scores are then normalized to ensure gradient stability. Finally, the softmax function is applied to translate the scores into probabilities, resulting in the weighted value matrix.

The process can be unified into a single function:

$$Attention(Q, K, V) = Softmax\left(\frac{Q.K^T}{\sqrt{d_k}}\right).V \quad (3)$$

Multi-head attention is a mechanism that used to boost the performance of the self-attention layer.

There is a residual connection to each sub -layer in the encoder and decoded. A layer-normalization [45] is followed after the residual connection. The output of these operations can be described as:

$$NormLayer(X + Attention(X)) \quad (4)$$

Here, X is used as the input of Multi Head Attention layer [20]. The output is passed through several MLP layers to get the final answer in the form of a feature vector.

During training, cross-entropy loss and the Adam optimizer are used. The execution time for each batch is also reported in the table. For comparison, Fig. 4 presents the visual output. A single query image is selected from the query set, and other images from the same class or similar images are chosen. Finally, it is determined which image belongs to the original image class. This distinction is indicated by green and red colors.

Fig. 4: The example of the final output of the architecture that receives an image from the query set and confirms the match of the person from the set of similar people in the gallery set.

After implementing the specified modifications, the experimental results are summarized in Table 2, reporting the findings related to the Market-1501, DukeMTMC, and MSMT17 datasets. The IDE and PCB-U methods are based on convolutional neural networks. The ViT method is a transformer architecture previously explained, with a patch size and a stride of 16. It has an embedding dimension of 768, a depth and number of heads of 12, and four MLP layers. The reported values for these methods are from reference articles. However, the ViT-Small and DiT-Small methods are smaller transformer models. ViT-Small has an embedding dimension of 768, a depth and number of heads of 8, and three MLP layers. DiT-Small has an embedding dimension of 384, a depth of 12, and a number of heads of 8, with four MLP layers. Additionally, the proposed M-ViT-S model uses a reduced MobileNetV3, and the M-ViT-L model uses a larger version of MobileNetV3 for feature extraction before the reduced transformer model.

The training process included a batch size of 32, and all four mentioned methods were trained under the same conditions for 100 epochs. Results for other methods are also reported, but these models were not trained under the previous fixed conditions, and their results are reported.

The reported values for the IDE, PCB-U, and ViT methods are sourced from relevant references in the table. However, other methods have also been evaluated under identical conditions. Notably, execution time for each image category is a critical consideration, particularly for real-time system performance in hardware-constrained environments. Among the methods, ViT-Small and DiT-Small exhibit the shortest execution time. However, their mAP and subsequent Rank-1 performance significantly decrease. In contrast, CNN-based methods experience increased execution time due to their more complex backbone architectures. The ViT method maintains good accuracy but at the cost of

longer execution time. Remarkably, the M-ViT-Small and M-ViT-Large approaches achieve accuracy comparable to CNN networks while reducing computation time.

Table 2: Comparison of methods for market1501 and cuhk03 datasets

Methods	Dataset	mAP	Rank-1	Time Pre Batch
IDE [9]	Market-1501	68.5	85.3	-
	DukeMTMC	52.8	72.4	
	MSMT17	-	-	
PCB-U [14]	Market-1501	77.4	92.3	0.420
	DukeMTMC	68.8	82.6	
	MSMT17	-	-	
ViT [24]	Market-1501	89.5	95.2	0.434
	DukeMTMC	82.6	90.7	
	MSMT17	69.4	86.2	
ViT-Small	Market-1501	25.6	42.2	0.218
	DukeMTMC	21.4	30.9	
	MSMT17	17.3	24.8	
DiT-Small	Market-1501	30.1	48.6	0.180
	DukeMTMC	22.5	32.5	
	MSMT17	18.6	26.2	
M-ViT-Small	Market-1501	72.1	88.0	0.234
	DukeMTMC	61.2	78.7	
	MSMT17	56.6	72.1	
M-ViT-Large	Market-1501	74.5	89.3	0.331
	DukeMTMC	64.0	79.9	
	MSMT17	58.1	74.7	

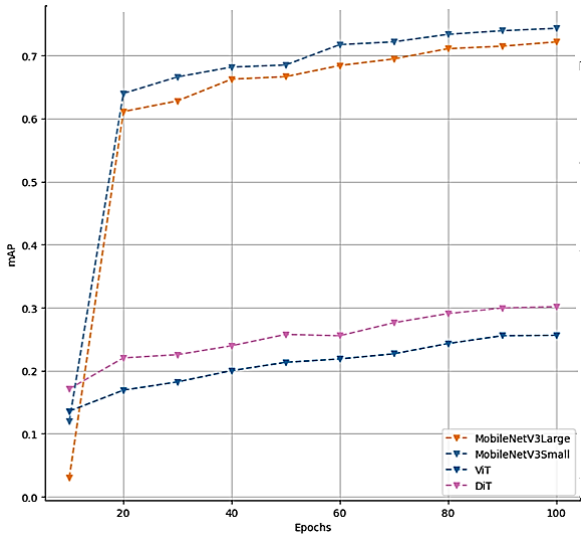

Fig. 5: mAP diagram of evaluation data.

Fig. 5: and Fig. 6: illustrates the network training process, showing the mAP and Rank-1 evaluation data for each stage.

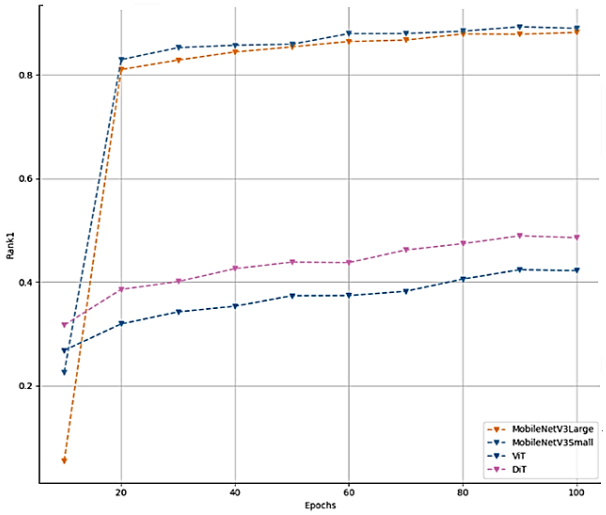

Fig. 6: Rank1 diagram of evaluation data.

Conclusion

The problem of person re-identification presents a multifaceted challenge with practical applications in our daily lives. Striking a balance between speed and accuracy is crucial when implementing a specific model for this purpose. The domain should be flexible enough to allow for the use of different methods based on specific needs.

Among other concerns in this field, data limitations pose significant challenges compared to other computer vision domains. Given the importance of input data in deep learning, tasks such as classification and identification involve extensive datasets. However, the person re-identification (ReID) problem is relatively limited in terms of available data. Another significant challenge is data quality. Since these data are often collected by cameras that lack optimal quality, methods capable of establishing meaningful connections between image components are of special importance.

By modifying existing models and leveraging essential features of each approach, the results indicate improved accuracy when using transformer models. Processing speed remains relatively unchanged due to the reduction in input information. The proposed method serves as an initial step in combining convolutional neural network (CNN) models with transformers, enhancing computational efficiency. Subsequent steps will focus on further improving models to achieve optimal results.

The adoption of newer transformer models with better computational speed, alongside more precise feature extraction, paves the way for future advancements. Data augmentation and related techniques can also contribute to enhancing performance in upcoming research areas. Furthermore, changes in models and the use of techniques to achieve faster and more appropriate responses are among important aspects to consider. Given the hardware limitations in real-time applications, reaching an optimally accurate model is crucial. This can

be achieved by reducing computational complexity in architectures, provided it does not compromise accuracy. Additionally, considering that in larger datasets similar to MSMT17, the accuracy of all models is lower, one could argue the possibility of overfitting in the models. Models may perform well only on a specific dataset and not exhibit suitable performance elsewhere. Given the limited dataset availability, this approach can be used for evaluating models in other computer vision applications, such as semantic segmentation, identification, and more, to ensure the introduced model has more precise and efficient performance.

In summary, this work aims to establish an interactive relationship between hardware constraints and sufficient accuracy. While many existing methods may achieve higher accuracy, the trade-off with increased computational demands is inevitable. This study seeks to explore specific adjustments and implementations to improve results while considering this balance between accuracy and computational cost.

Author Contributions

Dr. Zahiri has drawn the general road map. R .Iranpoor has searched for important articles in this field. Then, by checking the results and collecting the necessary data, the implementation of the proposed method has been done. Dr. Mehrshad reviewed the results and made changes in the way of implementation and final editing of the work.

Acknowledgment

The authors would like to thank the editor and anonymous reviewers.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

ReID	Person re-Identification
CNN	Convolutional Neural Network
ViT	Vision Transformer
DeiT	Data-efficient image transformer
PCB	Part-based Convolutional Baseline
RPP	Random Partitioning Pooling
mAP	mean Average Precision
CMC	Cumulative Matching Characteristics

References

[1] S. S. A. Zaidi, M. S. Ansari, A. Aslam, N. Kanwal, M. Asghar, B. Lee, "A survey of modern deep learning based object detection models," Digital Signal Process., 126: 103514, 2022.

[2] A. Krizhevsky, I. Sutskever, G. E. Hinton, "Imagenet classification with deep convolutional neural networks," Advances in neural information processing systems, 25(2), 2012.

[3] W. Wei, W. Yang, E. Zuo, Y. Qian, L. Wang, "Person re-identification based on deep learning—An overview," J. Visual Commun. Image Represent., 82: 103418, 2022.

[4] M. Farenzena, L. Bazzani, A. Perina, V. Murino, M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in Proc. 2010 IEEE computer society conference on computer vision and pattern recognition: 2360-2367, 2010.

[5] W. S. Zheng, S. Gong, T. Xiang, "Person re-identification by probabilistic relative distance comparison," in Proc. CVPR 2011: 649-656, 2011.

[6] K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition," in Proc. the IEEE conference on computer vision and pattern recognition: 770-778, 2016.

[7] Z. Zheng, L. Zheng, Y. Yang, "A discriminatively learned cnn embedding for person reidentification," ACM Trans. Multimedia Comput. Commun. Appl. (TOMM), 14(1): 1-20, 2017.

[8] H. Liu, J. Feng, M. Qi, J. Jiang, S. Yan, "End-to-end comparative attention networks for person re-identification," IEEE Trans. Image Process., 26(7): 3492-3506, 2017.

[9] L. Zheng, Y. Yang, A. G. Hauptmann, "Person re-identification: Past, present and future," arXiv preprint arXiv:1610.02984, 2016.

[10] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, Q. Tian, "Person re-identification in the wild," in Proc. the IEEE conference on computer vision and pattern recognition: 1367-1376, 2017.

[11] R. Girshick, J. Donahue, T. Darrell, J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in Proc. the IEEE conference on computer vision and pattern recognition: 580-587, 2014.

[12] H. Luo et al., "A strong baseline and batch normalization neck for deep person re-identification," IEEE Trans. Multimedia, 22(10): 2597-2609, 2019.

[13] Y. Sun, L. Zheng, Y. Yang, Q. Tian, S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in Proc. the European conference on computer vision (ECCV): 480-496, 2018.

[14] Y. Sun, L. Zheng, Y. Li, Y. Yang, Q. Tian, S. Wang, "Learning part-based convolutional features for person re-identification," IEEE Trans. Pattern Anal. Mach. Intell., 43(3): 902-917, 2019.

[15] Y. Sun et al., "Circle loss: A unified perspective of pair similarity optimization," in Proc. the IEEE/CVF conference on computer vision and pattern recognition: 6398-6407, 2020.

[16] G. Wang, Y. Yuan, X. Chen, J. Li, X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in Proc. the 26th ACM international conference on Multimedia : 274-282, 2018.

[17] H. Luo, W. Jiang, X. Zhang, X. Fan, J. Qian, C. Zhang, "Alignedreid++: Dynamically matching local information for person re-identification," Pattern Recognit., 94: 53-61, 2019.

[18] J. Qian, W. Jiang, H. Luo, H. Yu, "Stripe-based and attribute-aware network: A two-branch deep model for vehicle re-identification," Meas. Sci. Technol., 31(9): 095401, 2020.

[19] A. Vaswani et al., "Attention is all you need," Adv. Neural Inf. Process. Syst., 30, 2017.

[20] K. Han et al., "A survey on visual transformer," arXiv preprint arXiv:2012.12556, 2020.

[21] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, M. Shah, "Transformers in vision: A survey," ACM Comput. Surv. (CSUR), 54(10s): 1-41, 2022.

[22] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.

[23] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, "Training data-efficient image transformers & distillation through attention," in Proc. International Conference on Machine Learning: 10347-10357, 2021.

[24] S. He, H. Luo, P. Wang, F. Wang, H. Li, W. Jiang, "Transreid: Transformer-based object re-identification," in Proc. the IEEE/CVF International Conference on Computer Vision: 15013-15022, 2021.

[25] D. Wu et al., "Deep learning-based methods for person re-identification: A comprehensive review," Neurocomputing, 337: 354-371, 2019.

[26] D. Gray, H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in Proc. 10th European Conference on Computer Vision, Part I 10: 262-275, 2008.

[27] C. C. Loy, T. Xiang, S. Gong, "Multi-camera activity correlation analysis," in Proc. 2009 IEEE Conference on Computer Vision and Pattern Recognition: 1988-1995, 2009.

[28] W. Li, R. Zhao, X. Wang, "Human reidentification with transferred metric learning," in Proc. 11th Asian Conference on Computer Vision, Part I 11: 31-44, 2013.

[29] W. Li, R. Zhao, T. Xiao, X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in Proc. IEEE Conf. Computer Vision and Pattern Recognition: 152-159, 2014.

[30] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, "Scalable person re-identification: A benchmark," in Proc. IEEE International Conference on Computer Vision: 1116-1124, 2015.

[31] E. Ristani, F. Solera, R. Zou, R. Cucchiara, C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in Proc. European Conference on Computer Vision: 17-35, 2016.

[32] L. Wei, S. Zhang, W. Gao, Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in Proc. IEEE Conference on Computer Vision and Pattern Recognition: 79-88, 2018.

[33] S. Ren, K. He, R. Girshick, J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," IEEE Trans. Pattern Anal. Mach. Intell., 39(6): 1137-1149, 2017.

[34] S. Targ, D. Almeida, K. Lyman, "Resnet in resnet: Generalizing residual architectures," arXiv preprint arXiv:1603.08029, 2016.

[35] S. Xie, R. Girshick et al., "Aggregated residual transformations for deep neural networks," in Proc. the IEEE Conference on Computer Vision and Pattern Recognition: 1492-1500, 2017.

[36] A. G. Howard et al., "Mobilenets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017.

[37] J. Zang, L. Wang, Z. Liu, Q. Zhang, G. Hua, N. Zheng, "Attention-based temporal weighted convolutional neural network for action recognition," in Proc. Artificial Intelligence Applications and Innovations: 97-108, 2018.

[38] F. N. Iandola, S. Han, M. W. Moskewicz et al. "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size," arXiv preprint arXiv:1602.07360, 2016.

[39] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in Proc. IEEE Conference on Computer Vision and Pattern Recognition: 1251-1258, 2017.

[40] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L. C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in Proc. IEEE Conference on Computer Vision and Pattern Recognition: 4510-4520, 2018.

[41] A. Howard et al., "Searching for mobilenetv3," in Proc. IEEE/CVF International Conference on Computer Vision: 1314-1324, 2019.

[42] S. Elfwing, E. Uchibe, K. Doya, "Sigmoid-weighted linear units for neural network function approximation in reinforcement learning," Neural networks, 107: 3-11, 2018.

[43] Y. Guo, D. Zhou, W. Li, J. Cao, "Deep multi-scale Gaussian residual networks for contextual-aware translation initiation site recognition," Expert Syst. Appl., 207: 118004, 2022.

[44] P. Ramachandran, B. Zoph, Q. V. Le, "Searching for activation functions," arXiv preprint arXiv:1710.05941, 2017.

[45] J. L. Ba, J. R. Kiros, G. E. Hinton, "Layer normalization," arXiv preprint arXiv:1607.06450, 2016.

Biographies

Seyed Hamid Zahiri received the B.Sc., M.Sc. and Ph.D. degrees in Electronics Engineering from Sharif University of Technology, Tehran, Tarbiat Modarres University, Tehran, and Mashhad Ferdowsi University, Mashhad, Iran, in 1993, 1995, and 2005, respectively. Currently, he is a Professor with the Department of Electronics Engineering, University of Birjand, Birjand, Iran. His research interests include pattern recognition, evolutionary algorithms, swarm intelligence algorithms, and soft computing.

- Email: hzahiri@birjand.ac.ir
- ORCID: 0000-0002-1280-8133
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Rasool Iranpoor was born on July 19, 1991. He received M.Sc. degree in Electronic Engineering from Birjand University, Birjand, Iran, in 2018. He is currently a Ph.D. student at Birjand University to receive a Ph.D. degree in Electronics Engineering. His research interests include Machine Learning, Image Processing, Computer Vision, and Deep Learning Algorithms.

- Email: rasool.iranpoor@birjand.ac.ir
- ORCID: 0000-0002-7769-259X
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Nasser Mehrshad received the B.Sc. degree from Ferdowsi University of Mashhad in 1994. He completed his M.Sc. degree in Biomedical Electronics Engineering at Tarbiat Modares University in 1998 and received his Ph.D. in the same field in 2005. Currently, he serves as a full-time faculty member in the Department of Electrical and Electronic Engineering at the University of Birjand. His research interests include machine vision, digital signal processing, and biomedical engineering.

- Email: nmehrshad@birjand.ac.ir
- ORCID: 0000-0001-8678-3402
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

How to cite this paper:
S. H. Zahiri, R. Iranpoor, N. Meheshad, "Paying attention to the features extracted from the image to person re-identification," J. Electr. Comput. Eng. Innovations, 13(2): 267-274, 2025.

DOI: 10.22061/jecei.2024.10968.752

URL: https://jecei.sru.ac.ir/article_2206.html