Multiscale Attentional Squeeze-And-Excitation Network for Person Re-Identification

Tiancun Guo[1], Qiang Zhou[1]*, Mingliang Gao[1], Gwanggil Jeon[2]*, David Camacho[3]

[1] The Electrical and Electronic Engineering, Shandong University of Technology, Zibo 2 55000 (China)
[2] The Department of Embedded Systems Engineering, Incheon National University, Incheon, 22012 (South Korea)
[3] The Department of Computer Systems Engineering, Technical University of Madrid, Madrid (Spain)

* Corresponding author: zhouqiang@sdut.edu.cn (Q. Zhou), ggjeon@gmail.com (G. Jeon)

Received 14 May 2024 | Accepted 29 October 2024 | Early Access 15 January 2025

unir
LA UNIVERSIDAD
EN INTERNET

Abstract

In recent years, with the advancement of deep learning, person re-identification (Re-ID) has become increasingly significant. The existing person Re-ID methods primarily focus on optimizing network architecture to enhance Re-ID task performance. However, these methods often overlook the importance of valuable features in distinguishing Re-ID tasks, leading to reduced model efficacy in complex scenarios. As a solution, we utilize the attention mechanism to develop the lightweight multiscale Attentional Squeeze-and-Excitation Network (MASENet) that can distinguish between significant and non-significant features. Specifically, we utilize the SEAttention (SE) module to amplify important feature channels and suppress redundant ones. Additionally, the Spatial Group Enhance (SGE) module is introduced to enable networks to enhance semantic learning expression and suppress potential noise autonomously. We conduct comprehensive experiments on Market1501, MSMT17, and VeRi-776 datasets and cross-domain experiments on MSMT17 Ñ Market1501 to validate the model performance. Experimental results prove that the proposed MASENet achieves competitive performance across all experiments.

Keywords

Attention Mechanism, Cross-Domain, Multiscale, Person Re-ID.

DOI: 10.9781/ijimai.2025.01.001

I. Introduction

Person re-identification (Re-ID) is to determine whether pedestrian images extracted from different cameras or different video clips taken from the same camera are the same person. In recent years, person Re-ID has become a pivotal element within intelligent surveillance systems and has received significant attention from the computer vision community. Previous works [1]–[4] have made significant progress in the person Re-ID task. Most approaches still utilize a backbone model initially designed for generic image classification tasks [5]. Recent works [6] illustrate that using different architectures leads to model performance differences. Yet, some works for neural architecture search are still designed based on the traditional neural architecture search (NAS) methods employed for general classification tasks [7], [8]. The traditional NAS is associated with high computational costs and lacks generality. Also, the non-compatibility between the search scheme and actual world training schemes results in suboptimal performance in person Re-ID.

Aiming at the above problems, the MSINet [9] employs a twin comparison mechanism to eliminate the class binding between the training and validation sets. This mechanism offers more suitable supervision for neural architecture search in person Re-ID. It achieves compatibility between the search and real-world training schemes and improves the task's performance. Additionally, a multiscale interaction module is devised to facilitate mutual enhancement among multiscale features. Yet, person Re-ID is a complex and challenging task. The MSINet fails to adequately address scenarios where crucial feature channels have a more pronounced impact on the task. We draw inspiration from the multiscale interaction network (MSINet). Meanwhile, we propose incorporating an attention mechanism to guide the network in prioritizing the more influential feature channels. Concurrently, we empower the architecture to suppress insignificant feature channel information.

In this study, we construct the multiscale Attentional Squeeze-and-Excitation Network (MASENet) by incorporating the Squeeze-and-Excitation (SE) attention module and Spatial Group Enhance (SGE) module. We enhance the ability of the network to capture details in complex scenes, suppress background noise, and adjust features through attention mechanisms to focus on key features. Specifically, the contributions of this work can be summarized as follows:

- The MASENet is proposed to concentrates on the more essential feature information in person Re-ID. The MASENet can acquire more meaningful insights about pedestrians rather than being influenced by noise-disturbing features.

Please cite this article as:
T. Guo, Q. Zhou, M. Gao, G. Jeon, D. Camacho. Multiscale Attentional Squeeze-And-Excitation Network for Person Re-Identification, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 4, pp. 99-106, 2025, http://dx.doi.org/10.9781/ijimai.2025.01.001

- An SE module is adopted to learn each feature channel's significance autonomously. Meanwhile, an SGE module is introduced to generate an attention factor for each spatial position within each semantic group.
- With only 2.5M model parameters, extensive experiments conducted on several public datasets have verified that the proposed model surpasses existing methods in detection accuracy.

II. Related Work

With the advancement of deep learning, person Re-ID tasks have garnered increased attention within the domain of computer vision [10]. Researchers have proposed numerous methodologies in the realm of person Re-ID to enhance performance. Among these, the attention mechanism has gradually emerged as a crucial element in Re-ID.

A. Person Re-Identification

The objective of person Re-ID is to ascertain whether images of a person depict the same individual. The same or different cameras can capture these images at different times. There has been widespread research on person Re-ID based on deep learning [11]–[15]. Methods in deep learning for the person Re-ID task generally fall into two categories: designing more efficient networks and acquiring additional prior knowledge. Luo *et al.* [16] suggested a baseline that relies solely on the global features of ResNet50 to fulfill task performance requirements.

Zhou *et al.* [4] introduced a lightweight omni-scale network (OSNet) to capture various spatial scales and encapsulate multiscale collaborative composite features. Likewise, Li *et al.* [17] explored an efficient network architecture through microarchitecture search. They introduced the Top-k Sample Search strategy to achieve a cost-effective search while avoiding potential local optimal results. Some approaches apply body structure and posture information for site detection or person normalization. For example, Li *et al.* [18] utilized Spatial Transformer Networks (STN) with spatial constraints to learn and locate a person with attitude changes. The FD-GAN [19] is proposed to utilize identity-related and posture-independent representations. The FD-GAN sidesteps the necessity for additional pose information and reduces computational costs. In addition, some approaches [20], [21] concentrate on improving network performance by optimizing the loss function to enhance its relationship with the instance. For example, Gu *et al.* [22] proposed AutoLoss-GMS to search for an improved loss function within the loss function space to aim for efficient and excellent person Re-ID. Chen *et al.* [23] designed a quadruplet loss function and proposed a quadruplet deep network. The network incorporates online hard negative mining to enhance the model's generalization ability. Alternative methods [24], [25] focus on designing part-based models. They aim to emphasize the prominence of the person. Sun *et al.* [26] employed the Multi-Head Self-Attention Module (MHSAM) to address background confusion and occlusion challenges. While performance has been enhanced, the computational burden remains considerable. In this work, we achieve competitive performance with a lightweight architecture at a lower computing cost.

B. Attention-Based Person Re-ID

Recent years have witnessed considerable success in attention mechanisms in computer vision [27]–[30]. Also, the attention mechanism plays an indispensable role in person Re-ID. A body part detector is utilized to acquire the characteristics of a person's body parts [31], [32]. The connectivity of key points is utilized to generate a mask for human body parts and emphasize the representation of the human body [31]. Nevertheless, these methods heavily depend on the accuracy of analytical models of the human body or pose estimators. For video person Re-ID tasks, multiple methods [33], [34] investigated key time series frames using attention mechanisms. Additionally, there are methods to map 2D images into 3D spaces, facilitating pedestrian matching [35]. [36] introduces a point cloud matching (PCM) strategy to calculate the distance of multi-view convergence and allow for the differentiation of different individuals. Furthermore, Long Short-Term Memory (LSTM) is utilized to construct the motion dynamics of 3D tasks to simplify person matching [37]. A Reinforced Temporal Attention (RTA) based neural network architecture is proposed in [38]. It features a Long Short-Term Memory (CNN-LSTM) face-matching algorithm that utilizes an RGB-Depth conversion method. [39] employs the double attention mechanism to optimize and align features. This approach tackles the challenge of blurred vision in real-world scenarios. Chen et al. [40] proposed a network, named as ABD-Net. Spatial and channel attention are combined in the ABD-Net to directly learn a person's feature information from data and context. In SCSN [41], multiple attention models are cascaded to capture diverse cues. However, the complexity of cascading architectures poses a challenge in avoiding redundant information duplication, which leads to high computational costs. Our focus is on enhancing Re-ID's performance by implementing an attention strategy. Simultaneously, we achieve good performance without incurring undue computing costs.

III. Methodology

In this section, we delve into the details of the methods and modules utilized in the model. We first describe the work accomplished in the baseline (MSINet) [9] to facilitate comprehension. Following that, we elaborate on the details of the SEAttention (SE) module. Subsequently, we describe the Spatial Group Enhance (SGE) module. The structure of MASENet is shown in the Fig 1.

A. Baseline

1. Twins Contrastive Mechanism

The NAS is designed to adaptively search for the optimal network architecture for given data. In [9], defining the common model variable as α and the structure variable as β. In the search space σ, with the network layer i, β_i can manipulate the weighted value of individual operation o. The feature undergoes these operations iteratively. Ultimately, the final output is weighted and generated through the soft maximum of the operational output. Equation (1) describes the output.

$$f(x_i) = \sum_{o\in\sigma} \frac{exp\{\beta_i^o\}}{\sum_{o'\ \in\sigma} e\,xp\left\{\beta_i^{o'}\right\}} \cdot o(x_i) \tag{1}$$

The model parameters are updated based on training results. Subsequently, the schema parameters are updated using validation results. Since the testing and validation datasets share identical categories, Re-ID requires distinct categories to be included in both the training and validation datasets. This discrepancy results in incompatibility between the search scheme and the actual training scheme, potentially leading to suboptimal results. MSINet incorporates the Twin Comparison mechanism (TCM). Two independent auxiliary memories v_{tr} and v_{ver} are employed to reserve training features and validation data. In each iteration, the training loss is initially computed using the training auxiliary memory to provide data for model updates. Given the feature f with the class tag a, the classification loss is expressed as Equation (2).

$$L_{tr}^{cl} = -log\frac{exp(f\cdot v_{tr}^a/\tau)}{\sum_{h=0}^{H_{tr}^n} e\,xp(f\cdot v_{tr}^h/\tau)} \tag{2}$$

where v_{tr}^a represents the memory features associated with the class a, and H_{tr}^n represents the sum number of classes in the training data.

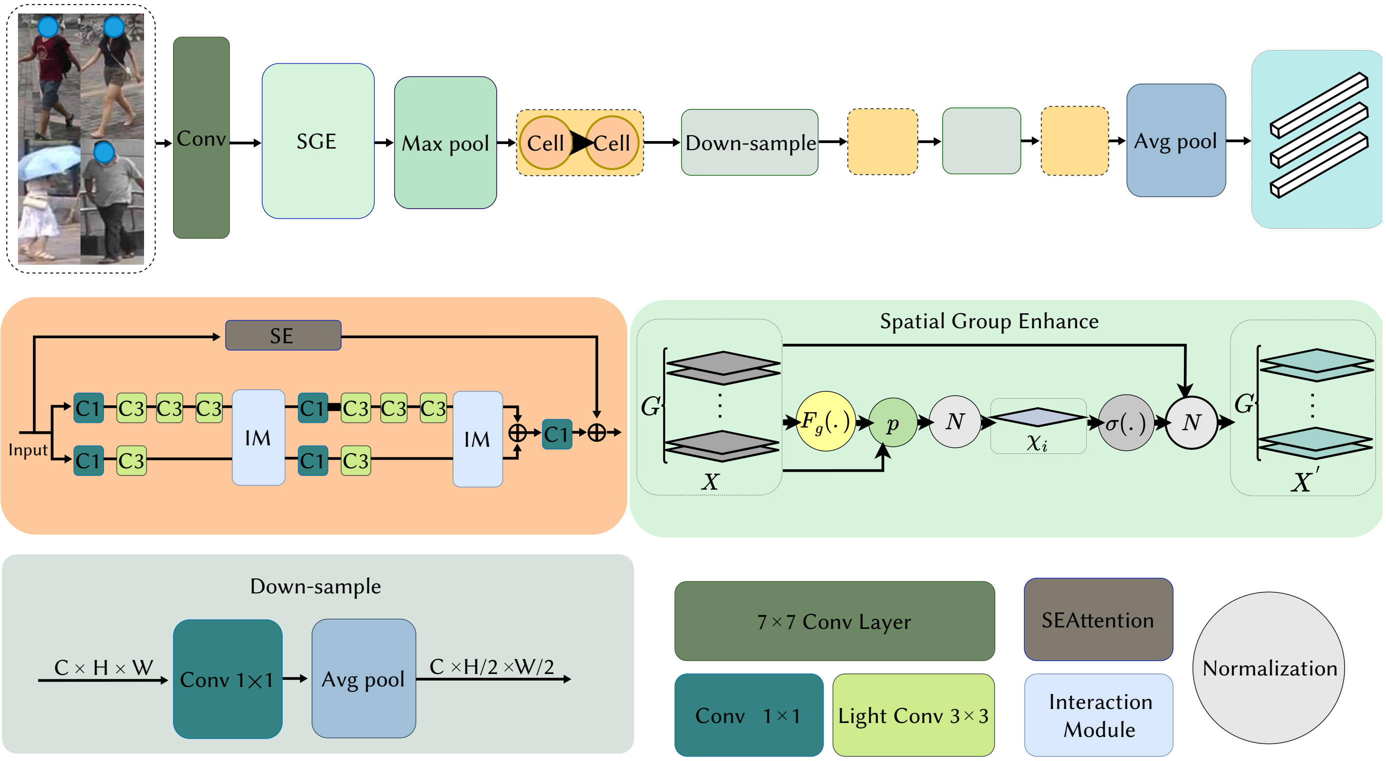

Fig. 1. The design of the proposed MASENet architecture. The MASENet allows for the input of pedestrian or vehicle images. The interaction module facilitates the exchange of information between two branches in each cell. The SE module enables the network to focus on useful feature channels, while the SGE enables each spatial group to enhance the expression of its learning autonomously.

The τ means the temperature argument should be set to 0.05 according to [42]. After the update, feature f is set to the corresponding memory feature by Equation (3).

$$v_{tr}^{a} \leftarrow \gamma v_{tr}^{a} + (1-\gamma) f \tag{3}$$

where γ is set to 0.2 [42]. Substituting v_{ver} with v_{tr} produces validation losses and updates schema parameters. It completes the iteration by validating the loss update pattern parameters.

2. Multiscale Interaction Space

While previous Re-ID research has incorporated multiscale features, it was primarily designed based on experience. MSINet has devised a multiscale interaction space enabling features to interact with one another. As depicted in Fig. 1, features traverse two branches with different receptive field scales within each cell. To achieve a network with low computational complexity, a stack of multiple convolutions is employed to set the scale. The Interaction Module acts as a conduit for the interchange of features and information between the two branches. The IM can execute four operations on an input feature (y_1, y_2): None. It does not involve operations with any parameters, yet accurately outputs (y_1, y_2). Exchange. Considered one of the most powerful interactions, it can be directly interchanged between the two branches and (y_1, y_2). Channel Gate. Channel Gate introduces Channel attention gates by Multi-Layer Perceptron (MLP) [43], [44], as shown in Equation (4):

$$G(y) = \delta(MLP(y)) \tag{4}$$

and returns value $(G(y_1) \cdot y_1, G(y_2) \cdot y_2)$. The MLP consists of two fully connected layers with parameters utilized by both branches. This enables networks to interact with each other by jointly filtering and validating feature channels.

Cross Attention Calculating the correlation between the two branches involves exchanging the keys of the branches. Then, the correlated activation [45] is converted into a mask and appended to the original feature in the right proportion. As depicted in Fig. 1, the two branches are fused through summation operations after the interaction. It's crucial to highlight that the additional parameters introduced by multiple interaction modules are limited. Each unit can be searched in the context of the entire network without being impacted. The interactions o that carry the most weight β_i^o for each layer are saved, thereby shaping the search architecture. After the architecture search, the model undergoes training to incorporate classification ID loss and triple loss, as shown in Equation (5):

$$L_{ID} = \frac{1}{N}\sum_{i=1}^{N} - log(\frac{expM_i^T f_i}{\sum_i e\,xpM_i^T f_i}) \tag{5}$$

where f_i is features array, M_i is the relevant classifier weight. The triple loss is expressed as Equation (6).

$$L_{TRI} = [D(f_a\ ,f_b\ ) - D(f_a\ ,f_n\ ) + \rho]_+ \tag{6}$$

where f_a, f_b, f_n are the inlaid features of the anchor. $D(f_a, f_b)$, $D(f_a, f_n)$ represent the Euclidean distance. ρ is the edge argument. $[.]_+$ means the *max*(., 0) function.

B. Structure of SEAttention Module

Learning extensive feature information solely through convolution kernels and achieving high performance is quite challenging for person Re-ID. Hence, we introduce the SEAttention module. From the perspective of feature channel information, SE specifies channel interdependencies without significantly increasing the network's depth or width. This technique results in only an increase in the number of model parameters. SE does not significantly increase the network's computational complexity. The importance weight of each feature channel can be adjusted based on its varying importance to the network. The network autonomously learns importance weights to enhance crucial feature channels and suppress redundant ones.

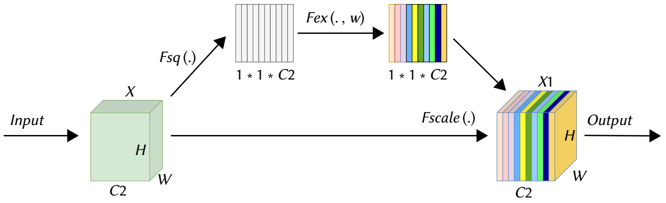

Fig. 2. The model structure of the Squeeze-and-Excitation (SE).

As depicted in Fig. 2, the SE module is integrated into residuals to accentuate the more significant feature channels. The SE module operation is divided into three steps: First, acquire each feature channel's global compression feature through global average pooling. Secondly, the new weight value of each feature channel will be derived from 0 to 1 via two fully connected layers. Lastly, matrix multiplication of the new weight value with the original feature channel will be performed using the SE module's feature channel recalibration function. Then, the output of the two branches is weighted and combined with the output of the SE module after a 1 × 1 conv.

C. Structure of Spatial Group Enhance Module

Feature representations of objects are generated by convolutional neural networks (CNNs) by acquiring semantic sub-features at different levels. Yet, the activation of these sub-features is often influenced by spatial noise. Therefore, we introduce the SGE [46] module to generate attention factors for each spatial position in each semantic group, shown in Fig. 1. It helps the module adjust each sub-feature's importance and suppress potential noise. Specifically, SGE divides feature graphs into groups G along channel dimensions. Each individual group has vectors representing each position in space, as shown in Equation (7):

$$X = \{x_1 \dots x_M\}, x_i \in \mathbb{R}^{\frac{C}{G}}, M = H \times W \tag{7}$$

where C is the number of channels. Within this group space, the network can learn the feature representation of the key region. Unlike CNN, which struggles to obtain uniformly distributed features, SGE utilizes global statistical features through the spatial average function to approximate the semantic vectors learned by the group. The Equation (8) is as follows:

$$F_g(X) = \frac{1}{M}\sum_{i=1}^{M} x_i \tag{8}$$

Then, the global features are utilized to generate an importance weight value for each feature. This weight value is obtained by (9):

$$p_i = F_g(X) \cdot x_i \tag{9}$$

It is worth noting that Eq. (9) can be reformulated as shown in Equation (10):

$$p_i = |F_g(X)||x_i| \cos(\delta_i) \tag{10}$$

where δ_i is the angle between $F_g(X)$ and x_i. We apply spatial normalization to p to avoid bias amplitude discrepancies between samples [47], [48]. It is mathematically expressed Equation (11).

$$p_i' = \frac{p_i - \eta_c}{\phi_c + \varepsilon}, \eta_c = \frac{1}{M}\sum_{j}^{M} p_j, \phi_c^2 = \frac{1}{M}\sum_{j}^{M}(p_j - \eta_c)^2 \tag{11}$$

where ε is a constant added for numerical stability. To ensure that normalization in the network can also represent the identity exchange, a pair of parameters φ, λ is introduced into each coefficient p_i'. The formula for scaling and moving normalized values is shown in Equation (12):

$$\chi_i = p_i'\varphi + \lambda \tag{12}$$

where the quantity of what φ, λ is the same as the number of G. Finally, the original x_i is scaled by the generated importance coefficients χ_i through a sigmoid function gate $\sigma(.)$ over the space, as shown in Equation (13):

$$x_i' = x_i \cdot \sigma(\chi_i) \tag{13}$$

Then, the enhanced feature vectors will be obtained, and the element group will be formed with these enhanced feature vectors. The specific form is given by Equation (14).

$$X' = \{x_1' \dots x_M'\}, x_i \in \mathbb{R}^{\frac{C}{G}}, M = H \times W \tag{14}$$

IV. Experiments

A. Datasets and Evaluation Metrics

The MASENet is tested on two Re-ID datasets about pedestrians: Market1501 [49], MSMT17 [50]. To assess the model's generalization ability, the MASENet is also evaluated on VeRi-776 [51], [52] and MSMT17 ⟶ Market1501 [49]. For simplicity and convenience, the three datasets are named M, MS, and VR. The output evaluation indexes are common performance metrics for person Re-ID, including mean average precision (mAP) and cumulative matching features (CMC).

B. Comparative Experiments With Other Lightweight Network

We initially contrast MASENet with the recently proposed lightweight network by in-domain and cross-domain experiments. The results in the table are pre-trained on ImageNet.

In-Domain Test. The initial learning rate is set at 0.065. During training, the learning rate is adjusted at epochs 150, 225, and 300. We use a Stochastic Gradient Descent (SGD) optimizer with a momentum coefficient of 0.9 and a weight decay of 0.0005. The parameters are updated using triple loss and cross-entropy loss. The value of p in formula (6) is set to 0.3. Adopting the same structure as CDNet [17], and the specific experimental results are shown in Table I.

ResNet50 is the most common backbone network for person Re-ID, but it performs the worst on the three datasets mentioned in this

TABLE I. The Performance on Re-ID Datasets. The Results Are Pre-Trained on ImageNet in Advance

Method	Params	M		MS		VR		MS ⟶ M	
		Rank-1 ↑	mAP ↑	Rank-1↑	mAP ↑	Rank-1 ↑	mAP ↑	Rank-1 ↑	mAP ↑
ResNet50* [16]	~24M	94.5	85.9	75.5	50.4	94.5	73.6	58.8	31.8
OSNet [44]	2.2M	94.8	84.9	78.7	52.9	95.5	76.4	66.6	37.5
CDNet [17]	1.8M	95.1	86.0	78.9	54.7	-	-	-	-
MSINet [9]	2.3M	95.3	89.6	81.0	59.6	**96.8**	78.8	74.9	46.2
MSINet-SAM [9]	2.4M	95.5	**89.9**	80.7	59.5	96.7	79.0	76.3	48.4
MASENet (Ours)	~2.5M	**95.9**	**89.9**	**81.9**	**60.8**	95.9	**79.5**	**77.3**	**50.1**

* represents the results reproduced by the baseline

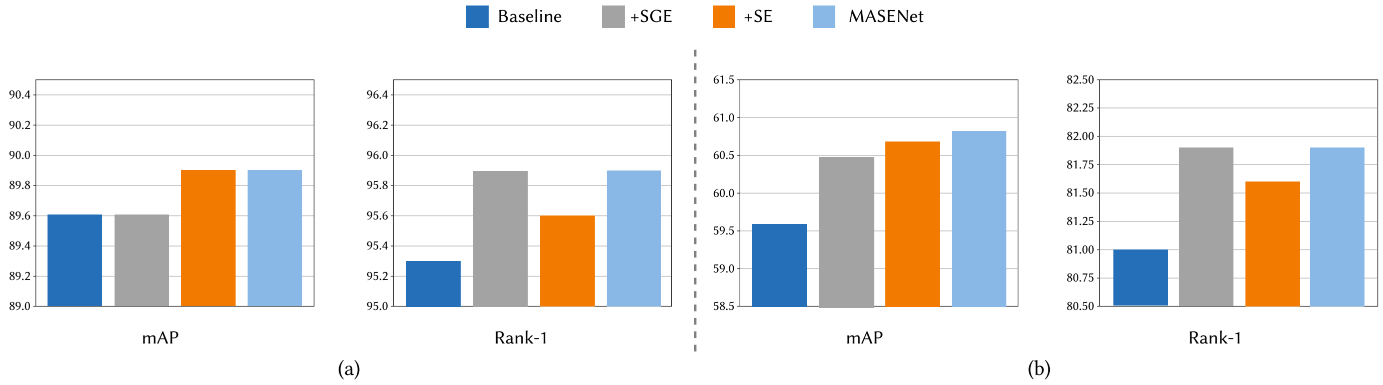

Fig. 3. Ablation experiments on (a) Market1501 dataset; (b) MSMT17 dataset.

article. Additionally, ResNet50 heavily relies on ImageNet pre-training operations. Unlike other datasets, the MS dataset presents more complex situations, such as background noise and attitude changes. The style of the MS dataset is more in line with real-world application scenarios. To overcome the challenges of complex scenarios, MASENet integrates SE and SGE modules. The SE module adjusts the importance of each feature channel adaptively to allow the network to learn and emphasize key features. The SGE module produces attention factors for each spatial location to adjust subfeature importance and mitigate background noise. These introduced modules enhance the network's feature representation ability and capture the detailed elements of complex scenes more effectively. Compared to MSINet, the MASENet improved mAP and Rank-1 by 1.3% and 1.2%. The results on MS validate that MASENet is more effective at handling complex scenarios and focusing on more important feature channels than the baseline. OSNet [44] and CDNet [17] are recent architectures designed for Re-ID, both addressing the issue of multiscale feature fusion. CDNet utilizes the traditional NAS scheme for searching. Table II shows the optimal interaction within each cell. It shows that the MASENet outperforms most lightweight networks.

TABLE II. The Detail About Interaction Operation. N: None; E: Exchange; G: Channel Gate; C: Cross Attention

Cell.1		Cell.2		Cell.3		Cell.4		Cell.5		Cell.6	
1	2	3	4	5	6	7	8	9	10	11	12
G	G	E	G	C	G	G	N	G	C	E	C

The model employed for person gender Re-ID is applied to the VR dataset for experiments. Table I indicates that in the VR experiment, mAP has increased by 1.2%.This case signifies an enhancement in the model's processing capability for generally complex scenes.

Cross-Domain Test. Cross-domain experiments are commonly employed to assess the generalization ability of models. MASENet is pre-trained with 250 epochs and fine-tuned to prevent overfitting. Table 1 demonstrates that ResNet50 is susceptible to image styles. The efficient interaction of MSINet can be effectively extended to other image domains. To enhance the generalization ability of MSINet, [9] introduced the spatial alignment module (SAM) module to align spatial correlation between person images. Yet, the performance of the proposed network on MS ⟶ M shows a substantial improvement compared to the MSINet-SAM. The results that mAP and Rank-1 are respectively up 1.7% and 1% than MSINet-SAM, further demonstrating the significant enhancement the module brings to the model.

C. Comparative Results With State-of-Art Methods

Table III offers additional insight into the supervised performance contrast between the proposed MASENet and SOTA methods on M and MS datasets. MASENet successfully achieves the objective of high precision with reduced computational requirements. The proposed method achieved an mAP of 89.9% and a Rank-1 accuracy of 95.9% on the Market1501 dataset.

TABLE III. The Performance Contrast Between MASENet and SOTA Methods on Market1501 and MSMT17 Datasets

Method	M		MS	
	Rank-1 ↑	mAP ↑	Rank-1↑	mAP ↑
Auto-ReID+ [55]	95.8	88.2	80.8	59.5
RGA-SC [54]	**96.1**	88.4	80.3	57.5
BAT-Net [56]	95.1	87.4	79.5	56.8
SFT [57]	94.1	87.5	79.0	58.3
CARL [53]	95.8	89.2	-	-
DRL-Net [58]	94.7	86.9	78.4	55.3
GCN [59]	95.3	85.7	-	-
PAT [60]	95.4	88.0	-	-
C2F [61]	94.8	87.7	-	-
BoT [62]	94.5	85.9	-	-
MGN* [63]	95.7	86.9	76.9	52.1
ISP [64]	95.3	88.6	-	-
OSNet [44]	93.6	81.0	71.0	43.3
CDNet [17]	95.1	86.0	78.9	54.7
MSINet [9]	95.3	89.6	81.0	59.6
MASENet (Ours)	95.9	**89.9**	**81.9**	**60.8**

Similarly, on the MS dataset, the proposed method achieves an mAP of 60.8% and a Rank-1 accuracy of 81.9%. CARL [53] introduces a measure of camera pairing loss for learning. Compared with CARL, the proposed method improves mAP and Rank-1 by 0.1% and 0.7% on the M dataset. It is worth noting that compared with MS dataset, M dataset has a simple style and certain limitations. During the training process of MASENet, the advantages brought by further feature enhancement may be difficult to fully exert. This situation may result in the limited performance improvement of the proposed method in the M dataset. Additionally, RGA-SC [54] incorporates a relation-aware global attention module. On the MS dataset, MASENet outperforms with a 1.6% boost in Rank-1 accuracy and a 4%enhancement in mAP. Although MASENet's Rank-1 performance on the M dataset is slightly lower than that of the RGA-SC method, it still demonstrates near-optimal performance. It verifies the effectiveness of matching top-ranked predictions.

Evaluation of the challenging MSMT17 dataset reveals that the proposed network also possesses the ability to handle challenging scenarios.

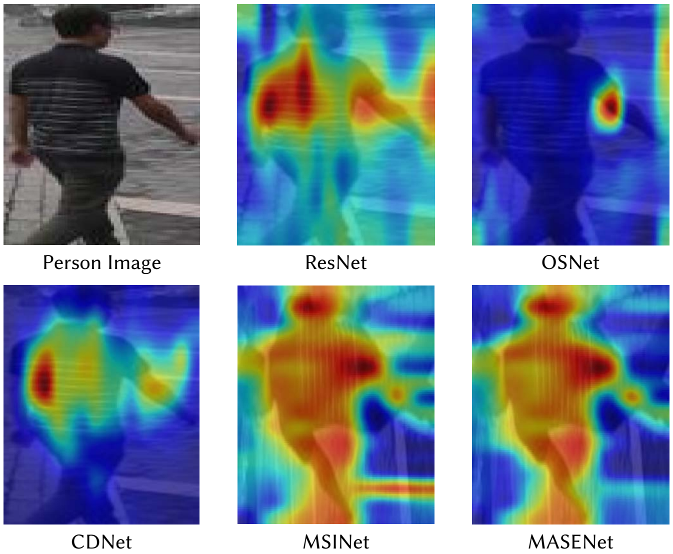

Fig. 4. Attention maps are generated from baseline (MSINet), OSNet, CDNet, and the proposed MASENet.

The attention maps are shown in Fig. 4. It proves that the introduced modules aid in filtering out inconsequential background noise, thereby enhancing the model's focus on the critical features of the pedestrian.

D. Ablation Studies

The performance improvement of MASENet primarily stems from the inclusion of SE and SGE modules. In this section, we conduct ablation experiments to validate the effectiveness of each module in enhancing network performance. The detailed results can be found in Table III. Additionally, we visualize the output from the baseline with independently introduced SE and SGE modules. As illustrated in Fig 5, the incorporation of SE and SGE modules effectively accentuates personal features while suppressing background noise interference. **Baseline**. Compared with other methods, baseline conducts neural architecture searches through twin comparison mechanisms. An effective interactive module also enables information exchange between two branches. These results, from the Market1501 dataset, with a map of 89.6% and Rank-1 with 95.3%, and from the MSMT17 dataset, with a map of 59.6% and Rank-1 with 81.0%, illustrate the improvement of model performance. However, real-world situations are intricate, and background noise can affect the model's performance in person Re-ID tasks. The performance of the baseline on MS suggests that it has not experienced significant improvement compared to other methods.

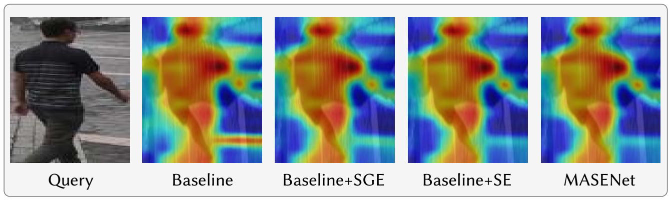

Fig. 5. Visualization of network output. The same sample is selected as in Fig 4. The baseline refers to MSINet.

SE Module. SE autonomously learns the importance of each feature channel in the feature map and assigns a weight value to enhance important feature channels. Fig 3 illustrates that SE improves the network's performance on both M and MS datasets. In particular, MASENet improves mAP and Rank-1 by 1.1% and 0.6% on the MS dataset. This demonstrates that the SE's attention to important feature channels effectively enhances the model's accuracy for person retrieval.

SGE Module. SGE generates attention factors for each spatial position in each semantic group. This capability empowers the network to independently enhance the expression of spatial semantic learning and suppress potential noise. Notably, the feature space enhancement mechanism proves especially advantageous for CMC. However, improvements in mAP are influenced by multiple factors, including dataset characteristics and optimization strategies. Moreover, the enhancement mechanism of the SGE module might alter feature distribution, potentially introducing deviations that impact the mAP performance. Overall, the SGE module is considerably more effective than the SE module in improving the network's cumulative matching feature.

V. Conclusion

In this work, we proposed a baseline approach for architectural search and incorporated the attention mechanism to create MASENet. Specifically, we introduce the SEAttention module to improve the network's attention to valuable feature channels. The Spatial Group Enhance module is introduced to enhance the expression of spatial semantic learning and suppress noise. This equips the network to address person Re-ID tasks with more complex backgrounds and poses. Experimental results demonstrate that MASENet exhibits outstanding performance and generalization ability on both person Re-ID and vehicle datasets. In the future, further optimization based on the CNN network architecture and SGE module will be explored. Additionally, the application of lightweight architecture and the enhancement of generalization performance will be pursued to adapt to complex Re-ID tasks.

Acknowledgment

This work has been funded by Grants: PLEC2021-007681 (XAI-DisInfodemics), PID2020-117263GB-100 (FightDIS), and PCI2022-134990-2 (MARTINI) of the CHISTERA IV Cofund 2021 program, funded by MCIN/AEI/10.13039/ 501100011033 and by the "European Union NextGenerationEU/PRTR"; by Calouste Gulbenkian Foundation, under the project MuseAI - Detecting and matching suspicious claims with AI, and by "Convenio Plurianual with the Universidad Polit'ecnica de Madrid in the actuation line of Programa de Excelencia para el Profesorado Universitario".

References

[1] Y. Dai, J. Liu, Y. Bai, Z. Tong, L.-Y. Duan, "Dual- refinement: Joint label and feature refinement for unsupervised domain adaptive person re-identification," *IEEE Transactions on Image Processing*, vol. 30, pp. 7815–7829, 2021.

[2] B. Yang, J. Chen, M. Ye, "Towards grand unified representation learning for unsupervised visible- infrared person re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 11069–11079.

[3] C. Zhao, X. Lv, Z. Zhang, W. Zuo, J. Wu, D. Miao, "Deep fusion feature representation learning with hard mining center-triplet loss for person re-identification," *IEEE Transactions on Multimedia*, vol. 22, no. 12, pp. 3180–3195, 2020.

[4] K. Zhou, Y. Yang, A. Cavallaro, T. Xiang, "Learning generalisable omni-scale representations for person re- identification," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 9, pp. 5056–5069, 2021.

[5] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[6] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, S. C. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE transactions on*

pattern analysis and machine intelligence, vol. 44, no. 6, pp. 2872–2893, 2021.

[7] H. Liu, K. Simonyan, Y. Yang, "Darts: Differentiable architecture search," *arXiv preprint arXiv:1806.09055*, 2018. Available online: https://arxiv.org/abs/1806.09055.

[8] X. Dong, Y. Yang, "Searching for a robust neural architecture in four gpu hours," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1761–1770.

[9] J. Gu, K. Wang, H. Luo, C. Chen, W. Jiang, Y. Fang, S. Zhang, Y. You, J. Zhao, "Msinet: Twins contrastive search of multi-scale interaction for object reid," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19243–19253.

[10] G. Zou, G. Fu, X. Peng, Y. Liu, M. Gao, Z. Liu, "Person re-identification based on metric learning: a survey," *multimedia tools and applications*, vol. 80, no. 17, pp. 26855–26888, 2021.

[11] S. Liao, S. Z. Li, "Efficient psd constrained asymmetric metric learning for person re-identification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3685–3693.

[12] Y. Liu, G. Zou, G. Chen, M. Gao, L. Yin, "Unsupervised person re-identification based on distribution regularization constrained asymmetric metric learning," *Applied Intelligence*, vol. 53, no. 23, pp. 28879–28894, 2023.

[13] B. Chen, W. Deng, J. Hu, "Mixed high-order attention network for person re-identification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 371–381.

[14] H. Liu, J. Feng, M. Qi, J. Jiang, S. Yan, "End-to- end comparative attention networks for person re-identification," *IEEE transactions on image processing*, vol. 26, no. 7, pp. 3492–3506, 2017.

[15] G. Chen, T. Gu, J. Lu, J.-A. Bao, J. Zhou, "Person re- identification via attention pyramid," *IEEE Transactions on Image Processing*, vol. 30, pp. 7663–7676, 2021.

[16] H. Luo, W. Jiang, Y. Gu, F. Liu, X. Liao, S. Lai, J. Gu, "A strong baseline and batch normalization neck for deep person re-identification," *IEEE Transactions on Multimedia*, vol. 22, no. 10, pp. 2597–2609, 2019.

[17] H. Li, G. Wu, W.-S. Zheng, "Combined depth space based architecture search for person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6729–6738.

[18] D. Li, X. Chen, Z. Zhang, K. Huang, "Learning deep context-aware features over body and latent parts for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 384–393.

[19] Y. Ge, Z. Li, H. Zhao, G. Yin, S. Yi, X. Wang, *et al.*, "Fd-gan: Pose-guided feature distilling gan for robust person re-identification," *Advances in neural information processing systems*, vol. 31, 2018.

[20] D. Cheng, Y. Gong, S. Zhou, J. Wang, N. Zheng, "Person re-identification by multi-channel parts-based cnn with improved triplet loss function," in *Proceedings of the iEEE conference on computer vision and pattern recognition*, 2016, pp. 1335–1344.

[21] Y. Yuan, J. Zhang, Q. Wang, "Deep gabor convolution network for person re-identification," *Neurocomputing*, vol. 378, pp. 387–398, 2020.

[22] H. Gu, J. Li, G. Fu, C. Wong, X. Chen, J. Zhu, "Autoloss- gms: Searching generalized margin-based softmax loss function for person re-identification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 4744–4753.

[23] W. Chen, X. Chen, J. Zhang, K. Huang, "Beyond triplet loss: a deep quadruplet network for person re- identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 403–412.

[24] G. Chen, J. Lu, M. Yang, J. Zhou, "Spatial-temporal attention-aware learning for video-based person re- identification," *IEEE Transactions on Image Processing*, vol. 28, no. 9, pp. 4192–4205, 2019.

[25] Y. Sun, Q. Xu, Y. Li, C. Zhang, Y. Li, S. Wang, J. Sun, "Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 393–402.

[26] H. Tan, X. Liu, B. Yin, X. Li, "MHSA-Net: Multihead self-attention network for occluded person re-identification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 11, pp. 8210-8224, 2023.

[27] X. Guo, M. Gao, W. Zhai, J. Shang, Q. Li, "Spatial-frequency attention network for crowd counting," *Big data*, vol. 10, no. 5, pp. 453–465, 2022.

[28] S. Jiao, J. Wang, G. Hu, Z. Pan, L. Du, J. Zhang, "Joint attention mechanism for person re-identification," *IEEE Access*, vol. 7, pp. 90497–90506, 2019.

[29] R. Quan, X. Dong, Y. Wu, L. Zhu, Y. Yang, "Auto- reid: Searching for a part-aware convnet for person re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3750–3759.

[30] W. Zhai, M. Gao, Q. Li, G. Jeon, M. Anisetti, "Fpanet: feature pyramid attention network for crowd counting," *Applied Intelligence*, vol. 53, no. 16, pp. 19199–19216, 2023.

[31] J. Xu, R. Zhao, F. Zhu, H. Wang, W. Ouyang, "Attention-aware compositional network for person re- identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2119– 2128.

[32] C. Song, Y. Huang, W. Ouyang, L. Wang, "Mask- guided contrastive attention model for person re- identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1179–1188.

[33] G. Chen, J. Lu, M. Yang, J. Zhou, "Learning recurrent 3d attention for video-based person re-identification," *IEEE Transactions on Image Processing*, vol. 29, pp. 6963– 6976, 2020.

[34] G. Chen, Y. Rao, J. Lu, J. Zhou, "Temporal coherence or temporal motion: Which is more critical for video- based person re-identification?," in *Computer Vision– ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, 2020, pp. 660–676, Springer.

[35] Z. Zheng, X. Wang, N. Zheng, Y. Yang, "Parameter-efficient person re-identification in the 3d space," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 6, pp. 7534-7547, 2024.

[36] M. Munaro, A. Fossati, A. Basso, E. Menegatti, L. Van Gool, "One-shot person re-identification with a consumer depth camera," *Person Re-Identification*, pp. 161–181, 2014.

[37] A. Haque, A. Alahi, L. Fei-Fei, "Recurrent attention models for depth-based person identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1229–1238.

[38] N. Karianakis, Z. Liu, Y. Chen, S. Soatto, "Person depth reid: Robust person re-identification with commodity depth sensors," *arXiv preprint arXiv:1705.09882*, 2017.

[39] J. Si, H. Zhang, C.-G. Li, J. Kuen, X. Kong, A. C. Kot, G. Wang, "Dual attention matching network for context-aware feature sequence based person re- identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5363– 5372.

[40] T. Chen, S. Ding, J. Xie, Y. Yuan, W. Chen, Y. Yang, Z. Ren, Z. Wang, "Abd-net: Attentive but diverse person re-identification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 8351–8361.

[41] X. Chen, C. Fu, Y. Zhao, F. Zheng, J. Song, R. Ji, Y. Yang, "Salience-guided cascaded suppression network for person re-identification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3300–3310.

[42] Y. Ge, F. Zhu, D. Chen, R. Zhao, *et al.*, "Self-paced contrastive learning with hybrid memory for domain adaptive object re-id," *Advances in neural information processing systems*, vol. 33, pp. 11309–11321, 2020.

[43] S. Woo, J. Park, J.-Y. Lee, I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[44] K. Zhou, Y. Yang, A. Cavallaro, T. Xiang, "Omni- scale feature learning for person re-identification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3702–3712.

[45] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3146–3154.

[46] X. Li, X. Hu, J. Yang, "Spatial group-wise enhance: Improving semantic feature learning in convolutional networks," *arXiv preprint arXiv:1905.09646*, 2019. Available online: https://arxiv.org/abs/1905.09646.

[47] Y. Wu, K. He, "Group normalization," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[48] S. Qiao, H. Wang, C. Liu, W. Shen, A. Yuille, "Micro-batch training with batch-channel normalization and weight standardization," *arXiv preprint arXiv:1903.10520*, 2019. Available online: https://arxiv.org/abs/1903.10520.

[49] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1116–1124.

[50] L. Wei, S. Zhang, W. Gao, Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 79–88.

[51] X. Liu, W. Liu, T. Mei, H. Ma, "A deep learning-based approach to progressive vehicle re-identification for urban surveillance," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, 2016, pp. 869–884.

[52] X. Liu, W. Liu, H. Ma, H. Fu, "Large-scale vehicle re-identification in urban surveillance videos," in *2016 IEEE international conference on multimedia and expo (ICME)*, 2016, pp. 1–6.

[53] J. Wu, Y. Yang, Z. Lei, Y. Yang, S. Chen, S. Z. Li, "Camera-aware representation learning for person re-identification," *Neurocomputing*, vol. 518, pp. 155–164, 2023.

[54] Z. Zhang, C. Lan, W. Zeng, X. Jin, Z. Chen, "Relation-aware global attention for person re-identification," in *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 2020, pp. 3186–3195.

[55] H. Gu, G. Fu, J. Li, J. Zhu, "Auto-reid+: Searching for a multi-branch convnet for person re-identification," *Neurocomputing*, vol. 435, pp. 53–66, 2021.

[56] P. Fang, J. Zhou, S. K. Roy, L. Petersson, M. Harandi, "Bilinear attention networks for person retrieval," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 8030–8039.

[57] C. Luo, Y. Chen, N. Wang, Z. Zhang, "Spectral feature transformation for person re-identification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4976–4985.

[58] M. Jia, X. Cheng, S. Lu, J. Zhang, "Learning disentangled representation implicitly via transformer for occluded person re-identification," *IEEE Transactions on Multimedia*, vol. 25, pp. 1294–1305, 2022.

[59] G. Xie, X. Wen, L. Yuan, H. Xu, Z. Liu, "Global correlative network for person re-identification," *Neurocomputing*, vol. 469, pp. 298–309, 2022.

[60] Y. Li, J. He, T. Zhang, X. Liu, Y. Zhang, F. Wu, "Diverse part discovery: Occluded person re-identification with part-aware transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2898–2907.

[61] A. Zhang, Y. Gao, Y. Niu, W. Liu, Y. Zhou, "Coarse- to-fine person re-identification with auxiliary-domain classification and second-order information bottleneck," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 598–607.

[62] H. Luo, Y. Gu, X. Liao, S. Lai, W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2019, pp. 1487-1495.

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

[64] K. Zhu, H. Guo, Z. Liu, M. Tang, J. Wang, "Identity- guided human semantic parsing for person re-identification," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, 2020, pp. 346–363, Springer.

Tiancun Guo

Tiancun Guo is pursuing an M.S. degree at the School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, China. His research interests include person re-identification, deep learning, and computer vision.

Qiang Zhou

Qiang Zhou (corresponding author) received his Ph.D. degree from Chongqing University, Chongqing, China, in 2016. He is an associate professor in the Department of Smart Grid Information Engineering, School of Electrical and Electronic Engineering, Shan Dong University of Technology, Shandong, China. His current research interests include artificial intelligence, image recognition, and smart power grid.

Mingliang Gao

Mingliang Gao received his Ph.D degree in Communication and Information Systems from Sichuan University. He is now an associate professor and vice dean at the Shandong University of Technology. He was a visiting lecturer at the University of British Columbia during 2018-2019. His research interests include computer vision, machine learning, and intelligent optimal control. He has published over 150 journal/conference papers in IEEE, Springer, Elsevier, and Wiley.

Gwanggil Jeon

Gwanggil Jeon(corresponding author) received a Ph.D. degree from the Department of Electronics and Computer Engineering, Hanyang University, Seoul, Korea, in 2008. From 2009.09 to 2011.08, he was with the School of Information Technology and Engineering, University of Ottawa, Ottawa, ON, Canada, as a Post-Doctoral Fellow. Dr. Jeon is an IEEE Senior Member, a recipient of the IEEE Chester Sall Award in 2007, the ETRI Journal Paper Award in 2008, and the Industry-Academic Merit Award by the Ministry of SMEs and Startups of Korea Minister in 2020. He serves as a full professor at Shandong University of Technology, Zibo, China, and Incheon National University, Incheon, Korea. His research interests include computer vision, machine learning, and the Internet of Things.

David Camacho

David Camacho is full professor at Computer Systems Engineering Department of Universidad Politécnica de Madrid (UPM), and the head of the Applied Intelligence and Data Analysis research group (AIDA: https://aida.etsisi.uam.es) at UPM. He holds a Ph.D. in Computer Science from Universidad Carlos III de Madrid in 2001 with honors (best thesis award in Computer Science). He has published more than 300 journals, books, and conference papers. His research interests include Machine Learning (Clustering/Deep Learning), Computational Intelligence (Evolutionary Computation, Swarm Intelligence), Social Network Analysis, Fake News and Disinformation Analysis. He has participated/led more than 50 research projects (Spanish and European: H2020, DG Justice, ISFP, and Erasmus+), related to the design and application of artificial intelligence methods for data mining and optimization for problems emerging in industrial scenarios, aeronautics, aerospace engineering, cybercrime/cyber intelligence, social networks applications, or video games among others. He has served as Editor in Chief of Wiley's Expert Systems since 2023 and sits on the Editorial Board of several journals, including Information Fusion, IEEE Transactions on Emerging Topics in Computational Intelligence (IEEE TETCI), Human-centric Computing and Information Sciences (HCIS), and Cognitive Computation among others.