sensors

MDPI

Article

Lightweight Multimodal Domain Generic Person Reidentification Metric for Person-Following Robots

Muhammad Adnan Syed [1,2], Yongsheng Ou [1,3,*], Tao Li [2] and Guolai Jiang [1]

[1] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China
[2] Konka R&D Department, Konka Group Co., Ltd., Shenzhen 518053, China
[3] Guangdong Provincial Key Laboratory of Robotics and Intelligent System, Shenzhen 518055, China
* Correspondence: ys.ou@siat.ac.cn

Abstract: Recently, person-following robots have been increasingly used in many real-world applications, and they require robust and accurate person identification for tracking. Recent works proposed to use re-identification metrics for identification of the target person; however, these metrics suffer due to poor generalization, and due to impostors in nonlinear multi-modal world. This work learns a domain generic person re-identification to resolve real-world challenges and to identify the target person undergoing appearance changes when moving across different indoor and outdoor environments or domains. Our generic metric takes advantage of novel attention mechanism to learn deep cross-representations to address pose, viewpoint, and illumination variations, as well as jointly tackling impostors and style variations the target person randomly undergoes in various indoor and outdoor domains; thus, our generic metric attains higher recognition accuracy of target person identification in complex multi-modal open-set world, and attains 80.73% and 64.44% *Rank*-1 identification in multi-modal close-set PRID and VIPeR domains, respectively.

Keywords: person re-identification; impostor resisting metric; multi-modal re-identification metric; lightweight domain generic metric; part-wise attention learning

check for updates

Citation: Syed, M.A.; Ou, Y.; Li, T.; Jiang, G. Lightweight Multimodal Domain Generic Person Reidentification Metric for Person-Following Robots. *Sensors* **2023**, *23*, 813. https://doi.org/10.3390/s23020813

Academic Editor: José María Martínez-Otzeta

Received: 2 November 2022
Revised: 19 December 2022
Accepted: 28 December 2022
Published: 10 January 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

With the advent of deep learning, human–robot interaction (HCI) is increasing rapidly in many applications. A robot following a person is one such application [1,2], where the person-following robots assist humans in elderly assistance and healthcare, work as service robots in industrial use, and also serve as autonomous carts in shopping malls.

Clearly, for all the applications above, it is required to track the person, and for tracking the person, the fundamental step is to first accurately identify the target person *P*1 shown in Figure 1, and then robustly track the target *P*1 in real-world. However, the dynamic real world is highly nonlinear and multimodal where the appearance of target person *P*1 shown in Figure 1 drastically varies from indoor home environment or domain to outdoor domain, such as walking across *Road*1, *Road*2, and *Road*3 in Figure 1, as well as across different outdoor domains, such as shopping mall or airport due to continually varying styles, illumination, poses, and viewpoints. In Figure 1, it can also be noted that the real world is also crowded where the target person *P*1 is also either occluded by other persons, say occluded by distractor *D*1 at time *t*2 on walking across *Road*1, or is also occluded by impostor *I*1 at time *t*4. Thus, in the real world in Figure 1, due to occlusion and nearby impostors, tracking the target person *P*1 is a very difficult task, hence, state-of-the-art trackers [3–6] could lose the tracking of target person *P*1, and either wrongly start following distractor *D*1 at time *t*4, or wrongly start following an impostor person *I*1 at time *t*5, as shown in red rectangles in Figure 1. The real target person *P*1 is walking on *Road*1 at time *t*4, and on *Road*4 at time *t*5, respectively. Therefore, robustly tracking the target person *P*1 in the nonlinear open world is still an unsolved problem, and it requires a robust target person identification in real time to reliably follow the person *P*1 in the real world.

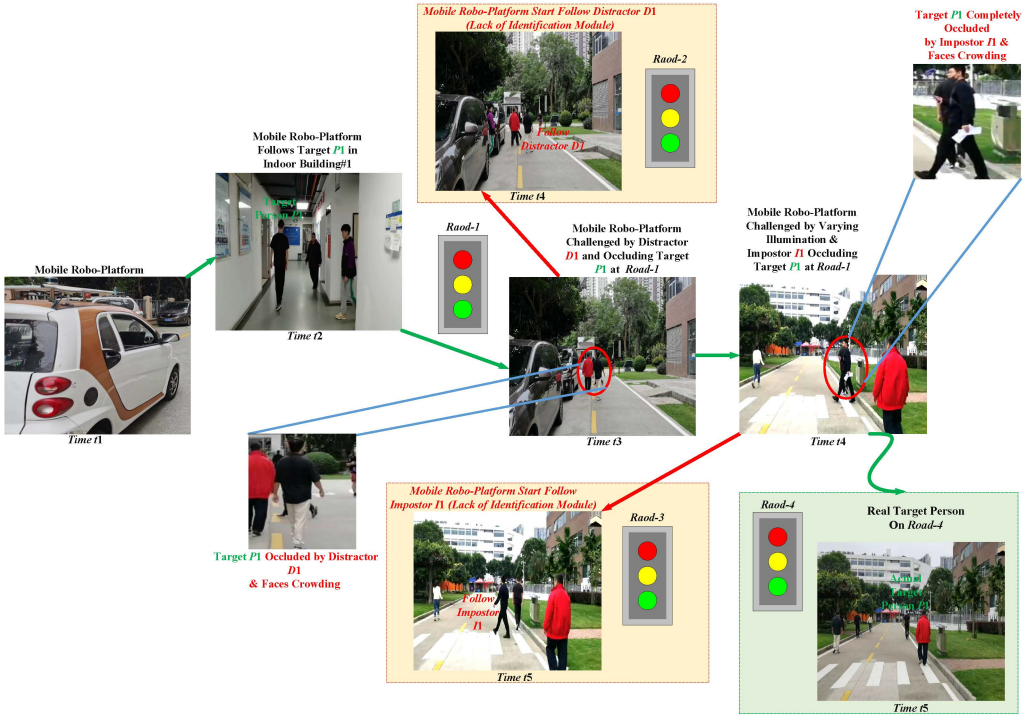


Figure 1. Person-follow robot scenario in outdoor world. The Mobile Roboplatform is Distracted in the Real Multi-Modal Open-Set World During Tracking Target P1.

In past, several works have addressed target person tracking and following. These past works use Laser Range Finder (LRF) [7,8], stereo camera [9], and RGB-D sensors [10] to track and follow the person. These trackers [7–10] have successfully addressed the person following problem, but these methods still get distracted in nonlinear and noisy outdoor environments due to occlusion, whereas these methods also lack the ability to reliably discriminate the target against similar-looking distractors in the outdoor world. Recently, deep visual trackers [11] have also learned to overcome the problem of reliably tracking the person in noisy environments; however, these deep visual trackers have still lacking to specifically address the target person identification problem during tracking and following. Visual trackers [11] are thereby still prone to impostors, and appearance changes the target person $P1$ undergoing in different nonlinear indoor and outdoor environments or domains.

Therefore, to overcome these shortcomings in visual trackers [11], recently, reidentification metrics have been learned and integrated with visual trackers [12] to follow the target person [1,2]. These reidentification metrics are learned by matching color-histograms and gait features [1,13], as well as extracting deep CNN features to learn deep similarity metrics [2,3,14–16]. However, the reidentification metrics in the present works [2,3,14–16] are all learned assuming the naïve world, i.e., it is assumed that the outside world is close-set, unimodal (it is assumed the robot only uses RGB sensor), and the target person $P1$ appearance remains unchanged across moving different domains. Due to these naïve assumptions, not only do the generalization capabilities of the learned reidentification metrics suffer largely in tracking the target in the outside world, but these metrics also get distracted due to the impostors observed across complex scenes and across nonlinear domains during tracking target $P1$, as shown in Figure 1. In Figure 1, at times $t3$ and $t4$ when the actual target $P1$ is completely occluded by distractor $D1$ and impostor $I1$ or if target $P1$ is completely moved out from the perception of mobile roboplatform in Figure 1, then the mobile roboplatform wrongly identifies a distractor $D1$ or impostor $I1$ as target and starts tracking the distractor $D1$ or impostor $I3$ at time $t4$ and $t5$, respectively.

Consequently, it is clear that to robustly track the target person $P1$ across different nonlinear indoor and outdoor environments, improvement of the trackers and integration of robust yet generic reidentification metrics in tracking are needed. The learned robust

and generic reidentification metric, thus, identifies the target person in each consecutive image frame and thereby improves the tracking of the target while largely preventing the tracker to wrongly follow the distractors or impostors. Therefore, in our work, we have learned a lightweight domain generic reidentification metric, referred as M_G, for following the target in outside world. Furthermore, a lightweight novel part-attention learning method is also proposed to accurately identify the target $P1$ across different nonlinear domains, as well as to further improve the tracker to reliably track the target $P1$. The purpose of the lightweight metric is to run the metric on the compact robotic platform, while domain generic metric is learned to tackle (i) the appearance and style changes of target $P1$ across different domains, (ii) to tackle impostors of $P1$ in the scenes, and (iii) to reliably recapture the target person $P1$ again using the novel attention features, if target $P1$ is lost due to occlusion or completely moved out from the robot perception during tracking. The generalization capability of the learned reidentification metric in our work is further improved in a way that our domain generic metric M_G is learned under realistic open-set scenarios, i.e., it is assumed that the appearance of the target person $P1$ varies when $P1$ moves across different indoor and outdoor environments. Then, the novel proposed attention module extracts the attention features from each single body part of the target person $P1$ to learn the deep cross-representations among the different images of the target person $P1$ undergoing appearance changes due to varying styles and modals. Therefore, the learned cross-representations are used to jointly address the pose and occlusion and also used to reject the large number of impostors during identifying the target in outside world, consequently largely improving the tracking of the target in outdoor world. In last, our contributions are:

- A robust vision-based target reidentification metric is proposed for target tracking. Compared to previous reidentification metrics proposed to target tracking, our reidentification metric is cross-modal and can address the style changes across large number of varying environments.
- Our reidentification learns part attention features, and unlike past works, the attention features are more stable to style changes and more robust against impostors. This is because the attention maps in our work are learned locally for each individual part, while during attention learning, it also uses global contextual and semantic information of the individual part. The global contextual provide useful relations among parts, while semantic information provides structural cues.
- Furthermore, the proposed reidentification uses the cross-representation module to jointly address pose and viewpoint changes and learns discriminative cross-view representations to tackle a large number of impostors in the open-set world.
- Finally, the learned metric is learned for the purpose of target tracking; therefore, it is designed with a lightweight backbone, while it is generic to help tracking the target in different nonlinear environments.

2. Related Work

The aim of our work is to learn a reidentification metric to integrate with tracking. Therefore, in this section, we review the recent works learning the reidentification metric for target tracking. Furthermore, in this section, we also review the recent works learning the reidentification metric for target tracking on mobile robots, and in the last, we also review the state-of-the-art related work in person reidentification that learns robust person identification metrics. In the beginning, we first cover the present works learning the reidentification metric for the visual tracking purpose.

2.1. Reidentification Metrics in Visual Tracking

Here, we cover the related work that learns reidentification metric for visual tracking purpose. However, the most of the learned models are used for offline tracking purpose. In [17], Neeti et.al. learned a LSTM-based CNN tracker using person reidentification module. They have learned spatiotemporal features of the person for reidentification;

however, during training, it is assumed that the real world is close-set and unimodal, and hence, the model performance could be challenged when deployed in the multimodal open-set world.

In another work in [18], tracking with person reidentification is learned, where the learned tracker tracks the target person in traffic scenes. Similarly, in [19], the authors follow the tracking-by-detection method, where the CNN-based feature matching is used to identify the target person in consecutive frames, and thus, the tracking of the target is done. The frameworks in [18,19] are simple, but these works require fine-tuning the reidentification metric every time for for every unseen domain, and thus have low generalization. Other works including [20,21] solve tracking in the multicamera network for smart city applications. [Edge Video] focuses to learn a lightweight reidentification metric to implement target tracking on edge devices, while Ref. [21] solves the problem of retracking the target after occlusion. Ref. [21] uses reidentification metric to reidentify the target when the target is recaptured again after the occlusion. Even [20,21] have good performance, however, ignored to solve reidentification problem in multimodal open-set world.

Furthermore, Ref. [22] also addressed the problem of target tracking in large scale scenario and proposed to learn an unsupervised reidentification metric for this purpose. The authors believed for the large-scale scenario that unsupervised reidentification requires no label of identities for training, and thus perform better than supervised metrics. Although the above works [17–22] used the learned reidentification metric for visual tracking, the methods are not intended for tracking the target on mobile robots. In order to track the target on mobile robots, lightweight and efficient reidentifications are needed; therefore, in the next subsection, we cover the recent works that specifically learn the reidentification metrics for tracking the target in real time and on mobile robots.

2.2. Reidentification Metrics for Visual Tracking on Mobile Robots

Here, we cover the recent state-of-the-art works learning reidentification metric for person identification and tracking for mobile robots. In [1] height, gait, and appearance features are used to learn an online person classifier to identify the target person to follow, while, Ref. [23] uses human pose estimation to detect the person indoors and then identify the target person using an appearance-features-based reidentification metric to follow. In addition, both [24] and [25] also use the appearance-features-based reidentification metric to detect and track the target person in the indoor environment. Although the method from [1,23–25] tracks well indoors, it lacks the ability to handle nonlinear style and appearance variations that the target person undergoes in outdoor world, and it is largely distracted due to impostors in the outdoor environment.

On the other hand, Ref. [15] also uses convolutional channel features to first identify the target person and then follow the identified target person using the mobile robot. Both [15] and [26] first use laser range finder to track the person position, then Ref. [15] learns the convolutional-channel-features-based classifier to verify the target to follow, whereas Ref. [26] uses monocular camera to perform appearance matching. In another work [2], an online person classifier is also learned to track the target person, but in the robot coordinate space. The authors believed tracking the person in robot coordinate space is more accurate than tracking in the real-world space. The methods [1,2,15] follow the person in both indoor and outdoor environments; however, the learned person classifiers in their works are not generic, while their works also fail to address resisting the impostors in the outdoor world.

Furthermore, some works have used depth sensing to track the position of target. Both [27] and [28] use kinect depth sensor to track the person position. The robot tracks the person, however, in the crowded environment; due to distractors, the accuracy of tracking is largely challenged in the outdoor setting due to the sensitivity of the Kinect sensor.

In other works, authors, e.g., those of [14], propose to track the target person using the Kalman filter. Once the bounding box of target person predicted, the state of Kalman filter is updated by identifying the target bounding box, and the target is then followed. In [3],

another problem of tracking the target in uniform crowd environment is solved. Their method depends on accurate face identification because the target and other persons all have similar appearances in the scene. Even the method performs well, but the method has still not addressed generalization problem across all outdoor environments, as well as Ref. [14] can be distracted due to impostors, whereas Ref. [3] can be distracted in crowded environment due to poor depth sensing.

In addition, there are a few works in reidentification-based person tracking on mobile robots aim to track the person for social and virtual game environments. In [29], both depth and laser range finder sensors are used to track the person position, however, only in the indoor environment, whereas Ref. [30] track and follow the person for the virtual game environment. Even their methods can track the target, but their methods are optimal for close-set scenarios, whereas the real-world scenario is multimodal and open-set. Now, covering the recent works in reidentification-based person tracking on mobile robots, we now further explore the state-of-the-art metrics learned for real-world person reidentification and are covered in detail in the subsection below.

2.3. State-of-the-Art Reidentification Metrics

2.3.1. Deep Metrics

Deep metric learning for Person ReID has been extensively studied in past works, e.g., [31,32]. These metrics, though, aim to address pose, viewpoint, occlusion, and misalignment of parts to attain high similarity; however, they underperform against unseen domains.

2.3.2. Domain Adaptation Metrics

To improve poor generalization, the authors of a few works proposed unsupervised domain adaptation (UDA) [33,34]. UDA adapts the learned ReID metric from labeled source domains to unlabeled target domain, but it is still time-consuming due to collecting data and fine-tuning the metric for each new unseen target domain.

2.3.3. Domain Generalizable Metrics

Therefore, recently, domain generalizable metrics have gained a great amount of attention in ReID [35–38]. Domain generic metrics are learned once and then directly applied for identification on previously unseen domains; however, existing metrics still ignore that the real world is multimodal and open-set, where the same person is seen in several different styles and modals. Therefore, it is desired that the generic metric in the real world (i) matches different Probe and Gallery images of the same person in different modals and styles and (ii) jointly addresses pose, viewpoint, and displacement of parts across views, while (iii) is also robust against impostors in open-world, and (iv) is lightweight, thus it can run on devices in the real world.

3. Methodology

In this work, our aim is to learn person reidentification metric to identify the target person and to integrate the learned reidentification metric with pretrained state of the art visual tracker to track the target; therefore, in this section, we will cover the details of learning the cross-modal domain generic open-set person reidentification metric and will describe its complete framework as shown below in Figure 2. In Figure 2, first realistic training data from open-set multi-modal world is generated, as shown in Figure 2a. In Figure 2a, a large number of images of N different person identities, say person $P1$, $P2$, and P_N, as shown in Figure 2a, are taken from D_T different source/camera domains.

Then, in the next step, a large number of images of each person identity are generated in different random poses. The purpose of generating images in different random poses for each different identity is to augment the training images of each person identity in different poses to learn its pose invariant features for identification. Next, in Figure 2a, taking all the original images of N persons and their generated images in random poses, we

now randomly transfer different images of different persons into different random styles. This is done because in the real world, when tracking the person in the outdoor world, it could move across several different environments, and in each different environment, it could undergo nonlinear style variations; thus, in order to identify the target person in different environments and in different styles, it is needed to learn a style robust person reidentification metric, and therefore, the images of a person are generated in multiple styles to train the reidentification metric in such a way that it could match all the images of the same person in all the different random styles. Finally, in the real world, it is also needed to obtain a cross-modal reidentification metric because it is possible that different mobile robot platforms use different imaging sensors, such as RGB and IR cameras. Therefore, in Figure 2a, to learn a cross-modal reidentification metric, our work takes different original images of N persons and takes their generated images in random poses and in random styles to randomly transform these different training images into different modals, say RGB, Grayscale, and Sketch modals. In our work, RGB images are transformed into Grayscale and Sketch modals due to the reason that in the real world, it is not necessary to always use an RGB sensor on a mobile robot; already, a large number of works have used IR modality [12]. However, a large number of public reidentification datasets have no IR images; therefore, we opted to transform RGB images into grayscale images. Sketch images are also generated to further improve the feature extraction power of the learned metric as well as to help in improving the cross-modality matching. After getting the realistic training data, a lightweight deep CNN backbone is then used to extract the features from the training images, as shown in Figure 2b. This deep CNN feature extraction backbone is designed using efficient residual module as shown in Figure 2b and is described in detail in Section 3.2.1. Furthermore, the feature extraction backbone also uses a novel part-attention module, as shown in Figure 2b, to extract the subtle features of the different individual parts. The details of novel part attention module are covered in Section 3.2.2. Next, using the learned attention features of different individual parts, our work learns cross-representations using the cross-representation learning module as shown in Figure 2b. The purpose of cross-representation learning is to minimize both the style and modality differences across cross-view features, which then feed to fully connected layers to learn the complex feature relationships to finally predict the similarity between the pair of images, as shown in Figure 2c.

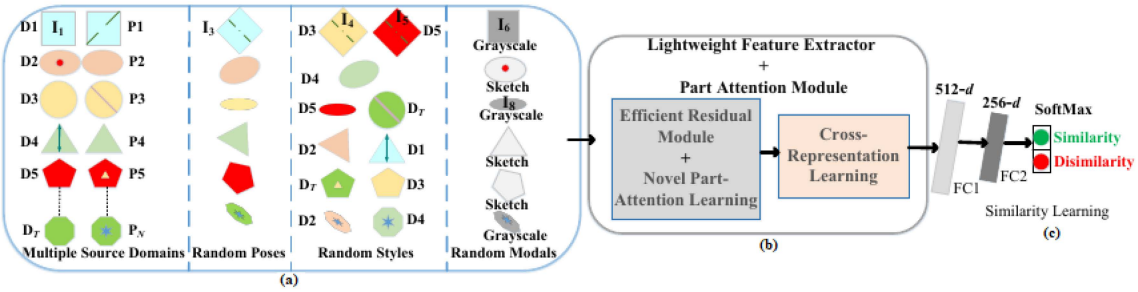


Figure 2. Framework of domain generic metric M_G. (**a**) Real Open-world Training Data. (**b**) Lightweight Features Extractor with Novel Attention and Cross-Representation. (**c**) Similarity Learning.

Now, below in Section 3.1, we will cover the details of generating training data in the complex nonlinear multimodal open-set world.

3.1. Multimodal Open-World Training Data

Here, we describe the details of generating training data in nonlinear multimodal open-set world. The real world is nonlinear, multimodal, and open-set where a person can be seen across several different domains in different styles and can also experiences pose, viewpoint, and parts displacement. Therefore, to train a robust metric for the real

world, it is first needed to obtain realistic world data for training. For this purpose, first, N number of different persons are randomly chosen from D_T source domains, as shown in Figure 2a. In order to generate realistic open-world training data, in the first step, as shown in Figure 2a, images in random poses are generated for each person. Taking the image I_1 of person $P1$ seen in domain $D1$, as shown in Figure 2a, a new image instance I_3 for $P1$ in random pose θ_p is generated as:

$$I_3 = \theta(I_1, \theta_p), \tag{1}$$

here θ is pose generation model [39]. Now, Equation (**??**) is used to generate images in random poses for all the N IDs in training data (in our work, images are generated in 8 random poses), and a few generated images for $ID1$ are shown below in Figure 3. Getting images for all the N persons in random poses, our work now generates images in varying styles for all N IDs. In Figure 2a, two instances I_4 and I_5 for ID $P1$ are generated in the styles of domain $D3$ and $D5$, respectively, as:

$$[I_4, I_5] = G\Big(\Big[w_{G_{D_3}}, w_{G_{D_5}}\Big], I_3\Big), \tag{2}$$

here $w_{G_{D_3}}$ and $w_{G_{D_5}}$ are the parameters of translation model G [33] for domains $D3$ and $D5$, respectively, and I_3 is the input image. In Equation (**??**), the purpose of generating images for $P1$ in varying styles [33,34] is to exploit the diverse and varying styles images of $P1$ to learn its style generic representation; thus, the learned generic metric could distinguish $P1$ seen in any random style in the open world. Hence, our work generates images of $P1$ (i) in varying styles in different disjoint views of the same domain $D1$ [34], (ii) in varying styles across random Re-ID domains [33], e.g., in styles of domains $D3$ and $D5$ as shown in Figure 2a, and in varying styles of random detection and recognition datasets, such as Imagenet [40]. Images in varying styles and poses are obtained for N persons; however, the real world is actually multimodal. Therefore, Grayscale image I_6 and Sketch image I_8 in Figure 2a for IDs $P1$ and $P4$ are respectively generated as:

$$\begin{aligned} I_6 &= \varphi(I_1), \\ I_8 &= \phi(I_7), \end{aligned} \tag{3}$$

here, function φ and ϕ convert RGB images I_1 and I_7 of $P1$ and $P4$ into Grayscale and Sketch modals, respectively. Now, in the next Section 3.2, we now describe the details of learning person features using the proposed novel part attention module.



Figure 3. A Few Generated images of ID1 During training. Images are Generated using Pose Model θ [39].

3.2. Novel Part-Attention Feature Learning

Now, using the generated cross-modal open-world training data, here, in this section, we cover the details of features extraction and novel part attention module. In the first, as shown in Figure 2b, person features are extracted with efficient lightweight CNN backbone and then, as shown in Figure 2c, the similarity between extracted pair of part attention features are learned. However, before covering the learning of features extraction, in our methodology, we first cover the details and design process of lightweight CNN backbone shown in Figure 2b below in Section 3.2.1.

3.2.1. Lightweight Backbone Design

Unlike [35–38], efficient residual module [40] is designed in Figure 4 to build a lightweight CNN backbone for learning cross-representations as shown in Figure 2b and learning the similarity between pair of learned features as shown in Figure 2c. Lightweight backbone as shown in Figure 2b is a Siamese network with 10-layers (details of layers are listed in Table 1), but for simplicity, only one stream is shown in Figure 2b. Each Convolution layer in Figure 2b is then realized with efficient residual module, where all convolutions are implemented as mixed depthwise separable convolutions [41] following Wider ResNet [42] strategy, i.e., the number of filters in each successive convolution layer are increased 2×times than previous convolution layer, thereby improving the features representation power with minimal computational cost. In addition, channel shuffle and channel split [43] are also used in efficient residual module in Figure 4 to enable information mixing across different filters and layers, thus further increasing diversity in features. After every convolution layer in Figure 4, Batch Normalization (BN) and ELU activation function are used for faster network convergence. Realizing the lightweight CNN backbone in Figure 2b, then, deep cross-representations are learned for each individual part of a person complimented with novel part-attention learning mechanism, as shown in Figure 2b.

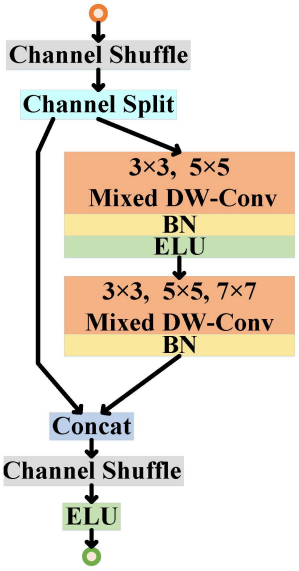


Figure 4. Our Efficient Residual Module Design.

Table 1. Detail of each layer, here, DW: Depthwise and MDW: Mixed-Depthwise. All different convolutions in MDW are 1-layer.

Layer Name	Output Size	10-Layer
conv1	112 × 112	7 × 7 (DW-Conv.), 64, stride 2
conv2_x	56 × 56	3 × 3, Max Pool, stride 2 3 × 3 (DW-Conv.), 128
conv3_x	28 × 28	$\begin{bmatrix} 3\times3,5\times5(\text{MDW}-\text{Conv.}),256 \\ 3\times3,5\times5,7\times7(\text{MDW}-\text{Conv.}),256 \end{bmatrix}\times 1$
conv4_x	14 × 14	$\begin{bmatrix} 3\times3,5\times5(\text{MDW}-\text{Conv.}),512 \\ 3\times3,5\times5,7\times7(\text{MDW}-\text{Conv.}),512 \end{bmatrix}\times 1$
conv5_x	7 × 7	$\begin{bmatrix} 3\times3,5\times5(\text{MDW}-\text{Conv.}),1024 \\ 3\times3,5\times5,7\times7(\text{MDW}-\text{Conv.}),1024 \end{bmatrix}\times 1$
	1 × 1	Global Average Pool, 512-d FC1, 256-*d* FC2, Softmax
FLOPs		1.1×10^9

3.2.2. Novel Part Attention Cross-Representations

Attention learning has been proved promising in Re-ID; however, methods from past works [35,44,45] learn attention globally for the whole body as shown in Figure 5a, and thereby, certain valuable and unique features from different individual parts are loss. Therefore, we argue to learn attention features for each individual part, as shown in Figure 5b, to prevent from loss, as well as, highlight the unique cues of different parts.

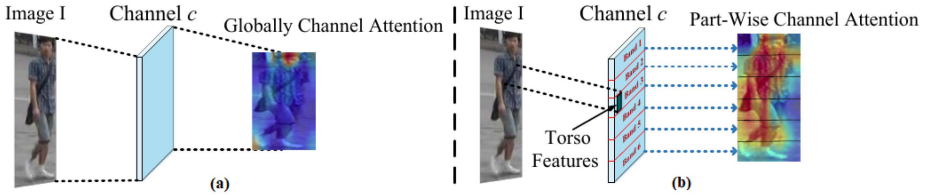


Figure 5. (**a**) Global channel attention [44,45]. (**b**) Our part-wise channel attention.

Novel Channel Attention for Individual Parts: Taking input features $F \in \mathbb{R}^{C \times H \times W}$, where C is the number of filters, while, $H \times W$ are feature spatial dimensions, channel attention for each individual part is then learned by horizontally dividing all the C filters into six spatial bands, as shown in Figure 6a, because each spatial band carries features of different part, as shown in Figure 5b, therefore, learning the attention for different corresponding bands, consequently, learns the attention weights for different corresponding parts. Hence, each single channel is partitioned into six horizontal bands, then the channel attention weight for each single band from each single channel is learned, but before the attention weight is learned it is first needed to capture the spatial, structural, and semantic relations of each band with its C neighboring bands, as shown in Figure 6b. Thus, similar to [44] relationships r and r' between bands x_1 and y_1, and between bands x_1 and z_1 are learned as:

$$r_{(x_1,y_1)} = \vartheta_c(f_{x_1}) \cdot \mu_c(f_{y_1}), \tag{4}$$

here, f_{x_1} and f_{y_1} are features of x_1 and y_1 with dimensions $\mathbb{R}^{1 \times \frac{H}{6} \times \frac{W}{6}}$, and the value $r_{(x_1,y_1)}$ computed as dot product between embedding functions ϑ_c and μ_c [44]. Embedding functions ϑ_c and μ_c are implemented by first flattening features f_{x_1} and f_{y_1}, then, apply 1×1 convolution followed by BN and ELU activation. Now, using Equation (**??**), all the C relations of band x_1 with all the C neighboring bands are then obtained to form the relation vector $\mathbf{r}_1$ for x_1 as: $\mathbf{r}_1 = \left[r_{c_{(x_1,:)}} \right]_{c=1,\ldots,C}$. Relation vector $\mathbf{r}_1$ is then embed with features f_{x_1} as:

$$f'_{x_1} = [pool_{Av}(v_c(f_{x_1})), pool_{Mx}(v_c(f_{x_1})), v_c(r_1)]. \tag{5}$$

here $pool_{Av}$ and $pool_{Mx}$ are global average and max pooling operations. Embedding function v_c first flattens features f_{x_1}, then, both v_c and v_c are implemented as 1×1 Conv followed by BN and ELU activation. Now, features f'_{x_1} are used to learn channel attention weight $a_{c_{x_1}}$, shown in Figure 6c, for band x_1 as:

$$a_{c_{x_1}} = Sigmoid(W_2 ELU(W_1 f'_{x_1})), \tag{6}$$

here, $W1$ and $W2$ are 1×1 Conv followed by BN. Now using Equations (4)–(6), first, all the six weights $a_{c_{x_1}}$, $a_{c_{x_2}}$, $a_{c_{x_3}}$, $a_{c_{x_4}}$, $a_{c_{x_5}}$, and $a_{c_{x_6}}$, shown in Figure 6c, for all the six bands in Channel-1 are computed, then, similarly, all the six weights for all the six bands in all the C channels are computed. Now, computing all the six weights for all the C channels, the six weights of each channel, e.g., Channel-1 weights shown in Figure 6c are taken and then each weight of each corresponding band is broadcasted similar to [45] over the spatial dimensions of each corresponding band to finally obtain channel attention $a'_c \in \mathbb{R}^{H \times W}$ for Channel-1. Following this, the channel attention a'_c for all the C channels are then obtained,

and finally, the attention maps for all C channels are concatenated together to form matrix A_c as: $A_c = [a'_{c'}]_{c'=1....C}$. Channel Attention features are now computed as:

$$F_{a_c} = F \otimes A_c, \tag{7}$$

here $\otimes$ denotes elementwise multiplication [45] between weights A_c and features F.

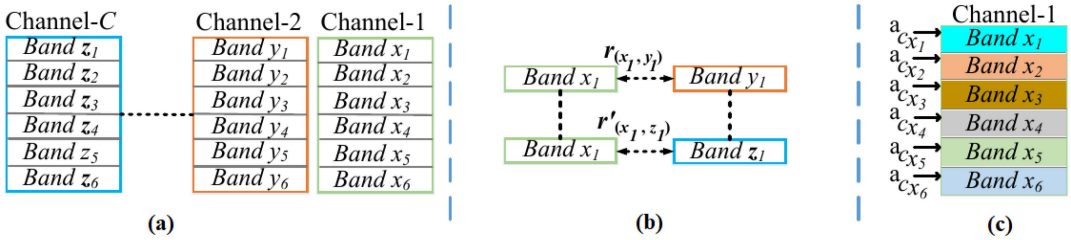


Figure 6. Novel channel attention learning for individual parts. (**a**) Dividing C channels into six bands; (**b**) Learning relations among bands; (**c**) Channel attention weight a'$_c$.

Novel Spatial Part Attention Features: Unlike past [35,44,45], our work learns spatial attention for pixel (i,j) for every 8 filters, i.e., weight a_{s_k} for every 8 filters as shown in Figure 7. This is done to improve spatial attention while preventing the loss of vital patterns that are largely diminish when spatial attention for pixel (i,j) is learned globally over all the C filters, e.g., $a_s(i,j) = -1$ in Figure 7. Now, every time taking 8 filters, e.g., filters $c = 1$ to $c = 8$ in Figure 7, spatial attention a_{s_k} for pixel (i,j) is learned by first learning the relations of pixel (i,j) with all the $(H \times W)$-1 pixels in the corresponding 8 filters as:

$$r_{k,l} = \vartheta_s(f_k) \cdot \mu_s(f_l), \tag{8}$$

here f_k and f_l are 8-dimension feature vectors of pixels (i,j) and (i',j') as shown in Figure 7, while, $r_{k,l}$ is the leaned relation, and embedding functions ϑ_s and μ_s are implemented as 1×1 spatial convolution followed by BN and ELU activation. Now, Equation (8) is used to learn all the $H \times W$ relations of vector f_k of pixel (i,j) to form the relation vector $\mathbf{r}_k$ as: $\mathbf{r}_k = [r_{r'}]_{r'=1,...,H\times W}$, then, vector $\mathbf{r}_k$ embed with vector f_k to form feature f'_k as:

$$f'_k = [pool_{Av}(\nu_s(f_k)), pool_{Mx}(\nu_s(f_k)), v_s(\mathbf{r}_k)]. \tag{9}$$

here, in Equation (9), embedding functions ν_s and v_s are learned as 1×1 Conv followed by BN and ELU activation. Now, the attention a_{s_k} for pixel position (i,j) is learned as:

$$a_{s_k} = Sigmoid(W_2 ELU(W_1 f'_k))), \tag{10}$$

here $W1$ and $W2$ are 1×1 Conv followed by BN. Now, Equations (8)–(10) are first used to learn the spatial weights a_{s_k} for pixel (i,j) for every 8 filters, as shown in Figure 7, and then, similarly, Equations (8)–(10) are also used to learn the spatial weights a_{s_k} for all the $H \times W$ pixels in every 8 filters. Getting the spatial weights a_{s_k} for all the $H \times W$ pixels in all the C filters, first, the learned corresponding $H \times W$ weights for every corresponding 8 filters are broadcasted [45]over the spatial dimensions $H \times W$, and then the attention maps $a_{s_g} \in \mathbb{R}^{8\times H\times W}$ (here $g = 1$ to $C/8$) for every corresponding 8 filters are obtained. These attentions maps are then concatenated together to form spatial attention weights matrix A_s for all C filters as: $A_s \in \mathbb{R}^{C\times H\times W}$. Then, finally, spatial attention features $F_{a_s} \in \mathbb{R}^{C\times H\times W}$ are obtained as:

$$F_{a_s} = F_{a_c} \otimes A_s. \tag{11}$$

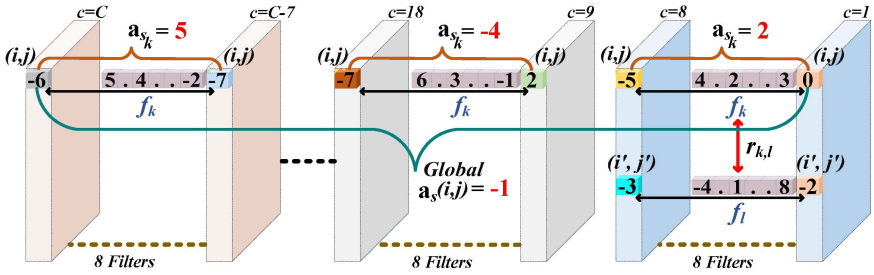


Figure 7. Novel Spatial Attention Learning for Individual Parts.

3.3. Multi-Modal Open-Set Generic Metric

Using the attention features of each single part, cross-representation module [46] shown in Figure 2b now learns the cross-representations for pair of features f_{q_1} and f_{q_2} as:

$$g(q_1, q_2) = CRM(f_{q_1}, f_{q_2}). \tag{12}$$

here, cross-representation module CRM in Equation (12) not only learns the complex nonlinear relationships between features f_{q_1} and f_{q_2} to minimize the existing domain, style, and modality gaps in different environments between positive pair (q_1, q_2), but, at the same time also addresses the pose, viewpoint, and spatial miss-alignment across views. For the given quadruplet q_1, q_2, q_3, q_4, all the cross-representations $g(q1, q2)$, $g(q1, q3)$, and $g(q2, q4)$ are learned using Equation (12), and the representations are then sent to fully-connected layers $FC1$ and $FC2$ shown in Figure 2c with 2-dimension softmax classifier to learn similarity between (q_1, q_2). The learned similarity value is then used to compute the quadruplet loss [47] L_{quad} as:

$$\begin{aligned} L_{quad} = & \sum_{n=1}^{N} \left[g(q_1, q_2)^2 - g(q_1, q_3)^2 + \alpha_1 \right] \\ & + \sum_{n=1}^{N} \left[g(q_1, q_2)^2 - g(q_2, q_4)^2 + \alpha_2 \right], \end{aligned} \tag{13}$$

here, N are total number of quadruplets, $\alpha_1 = 1$ and $\alpha_2 = 0.3$ are the margin values used, while q_3 and q_4 are the impostors of q_1 and q_2, and randomly seen in any domain in any style and modality, respectively.

4. Experiments and Analysis

4.1. Datasets and Data Augmentation

For training metric M_G Market1501, DukeMTMC-reID, CUHK03, and CUHK02 are used as training source domains, while M_G is comprehensively evaluated using domains VIPeR, PRID, GRID, and i-LIDS. For cross-modal evaluation, SYSU-MM01dataset used the following settings in [48]. In addition, random cropping, horizontal flipping, random rotation, color jittering, random contrast, brightness, and label smoothing regularization [49] are used in training to prevent M_G from overfitting.

4.2. Implementation Details

Lightweight CNN backbone in Figure 2b is trained from scratch with randomly initialized weights [50] for 600 epochs on single 16G NVIDIA RTX 2080 Ti GPU. The during-training image resolution is 224 × 224, with the Adam optimizer used with initial learning rate 8 $\times 10^{-5}$, mini-batch size 64, and weight decay 5 $\times 10^{-4}$. All of the code was written in Pytorch.

4.3. Evaluation Metrics and Protocols

Unlike [35–38], M_G is comprehensively evaluated in challenging multimodal, multi-style, close-set, and open-set scenarios. Cumulative Matching Characteristics (CMC) at

Rank-1 and mean average precision (mAP) are the metrics used for all close-set experiments, while for all open-set experiments, true target rate (TTR) is measured against false target rate (FTR) [51] as a performance metric. All results obtained after averaging over 10 trials.

4.3.1. Naïve Close-Set Scenario

Unlike [35–38], in this scenario, M_G is evaluated in two different and difficult settings, in setting#1 and setting#2. In both settings, Probe/Gallery images splits are: VIPeR: 316/316; PRID: 100/649; GRID: 125/900; i-LIDS: 60/60, respectively. Here, VIPeR: 316/316 means there are 316 person identities in test Query view, and similarly, there 316 same person identities in the test Gallery view. In setting#1, matches of Queries observed in a corresponding domain, e.g., observed in CUHK-03 domain, are found from the Gallery view of the corresponding CUHK-03 domain only. Below in Figure 8, the testing scenario of setting#1 is shown.



Figure 8. In Setting#1 it can be seen the matches of all Queries exist in the Gallery View, while the match for each person Seen in VIPeR domain is found from the Gallery of VIPeR domain domain.

On contrary, as shown in Figure 9, the matches of Queries in setting#2 are found from a joint Gallery containing Gallery images from all the test domains, i.e., VIPeR, PRID, GRID, and i-LIDS. Thus, in setting#2, M_G is tested in more realistic and challenging scenario to find the matches of given Query by resisting large number of impostors from different outdoor environments. Identification results of M_G in Setting#1 are summarized in Table 2, where M_G attains 80.73%, 64.44%, and 88.99% Rank-1 identification on PRID, VIPeR, and i-LIDS, respectively.



Figure 9. In Setting#2 the match for given Query image, the match for Query from GRID domain is found from the joint Gallery, where the Gallery contains images from all test domains, i.e., GRID, CUHK-03, and 3DPes.

Table 2. Results on Naïve Close-Set Setting#1.

Methods	Target:PRID		Target:GRID		Target:VIPeR		Target:i-LIDS	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
CBAM [45]	47.52	53.89	33.07	39.17	45.80	51.87	75.08	81.0
HLGAT [52]	48.74	56.84	35.22	43.94	48.96	58.03	79.67	82.64
RAGA [44]	50.69	60.47	37.79	45.51	49.03	58.33	83.91	85.2
SRN [35]	52.1	66.5	40.2	47.7	52.9	61.3	84.1	89.9
D.Norm [36]	60.4	-	41.1	-	53.9	-	74.8	-
MoE [38]	57.7	67.3	46.8	54.2	56.6	64.6	85.0	90.2
Meta [37]	74.2	81.0	48.4	57.9	59.9	68.6	81.3	87.0
Our M_G	80.73	86.91	55.07	61.23	64.44	75.1	88.99	91.0

Clearly, metric M_G attains higher recognition than [35–38,44,52]; it is because M_G learns cross-representations among different indoor and outdoor images of a person complimented with part-attention learning, where learned part-attention pays large focus on unique features from each different part, thus, M_G jointly minimizes pose, viewpoint, and spatial displacement of parts, as well as jointly addresses style, modality, and domain gaps to resist large number of impostors in outdoor world.

Furthermore, in setting#2, though, setting#2 is difficult than setting#1, but M_G still surmounts the challenges in setting#2, and in Table 3, M_G retrieves 77.89%, 52.64%, 61.92%, and 86.23% matches of the Queries at *Rank*-1 from the joint Gallery for test domains PRID, GRID, VIPeR, and i-LIDS, respectively. These results clearly reveal M_G can inherently tackle pose, viewpoint, style, and modality transforms across both nonlinear indoor and outdoor environments, while, complemented with the part-attention mechanism, M_G also learns cross-representations in a way to jointly address occlusion and misalignment of parts; therefore, in Figure 10d, M_G declines the large number of impostors and improves the identification accuracy from *Rank* = 5 in Figure 10a to *Rank* = 1.

Table 3. Results on Naïve Close-Set Setting#2.

Methods	Target:PRID		Target:GRID		Target:VIPeR		Target:i-LIDS	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
CBAM [45]	39.42	51.33	25.21	37.99	32.61	44.04	69.24	73.0
HLGAT [52]	43.95	50.07	28.77	39.5	37.66	49.72	70.77	76.95
RAGA [44]	44.49	50.1	30.92	40.32	38.55	51.79	72.14	78.63
SRN [35]	47.85	51.72	34.87	43.67	44.99	52.05	72.1	78.41
D.Norm [36]	53.70	60.1	37.22	44.33	49.54	54.41	71.04	80.03
MoE [38]	51.51	63.91	39.35	47.69	48.74	57.0	79.0	83.1
Meta [37]	70.82	78.07	42.61	50.95	51.74	59.97	79.68	84.78
Our M_G	77.89	84.78	52.64	58.85	61.92	73.08	86.23	90.0

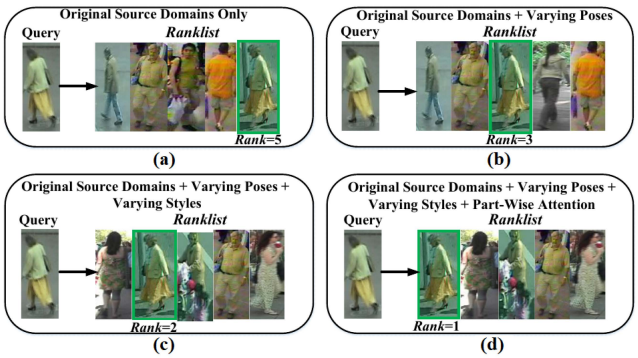


Figure 10. Retrieval results obtained in close-set setting#2 using different training data. (**a**) Training data only contains original image samples with no data augmentation; (**b**) Training data contains original image samples, and image samples are generated in different random poses too; (**c**) Training data contains original image samples, and image samples are generated in different random poses and in different random styles too; (**d**) Training data contains original image samples, and image samples are also generated in different random poses and styles, as well as, part-wise attention is learned to improve features power.

4.3.2. Challenging Close-Set Scenario

The joint Gallery in setting#2 is very challenging; however, the target person being followed by robot when moving across different outdoor environments could undergo style, illumination, pose, and viewpoint changes in the real world. Therefore, to obtain a robust reidentification metric, M_G is tested in the real-world environment where Probe-Gallery pairs could be seen in different modals [53] and in different styles [35]. The real-world testing scenario for Challenging Close-Set is shown below in Figure 11.

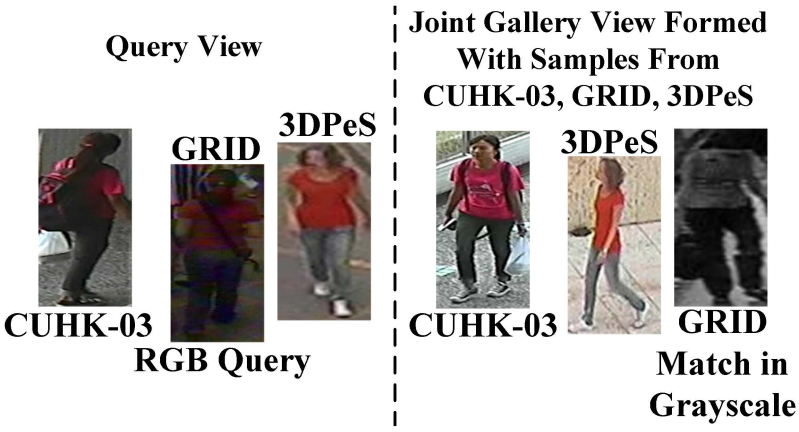


Figure 11. In the Challenging Close-Set Scenario, the Query image, the image of GRID domain could be seen in the RGB modalily (or different style), whereas its match in Gallery view has Grayscale modality.

Therefore, during testing in the Challenging Close-Set Scenario, images in Query and in joint Gallery views are randomly transformed into different modals and styles, then, M_G finds the matches from joint multimodal multistyle Gallery, and the results are given in Table 4.

Table 4. Results on Challenging Close-Set Scenario.

Methods	Target:PRID		Target:GRID		Target:VIPeR		Target:i-LIDS	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
CBAM [45]	30.98	40.88	21.8	28.4	25.74	32.37	58.37	62.66
D.Norm [36]	31.49	42.41	23.54	30.68	27.73	35.20	60.56	69.03
RAGA [44]	34.93	42.99	24.12	32.34	32.64	37.85	64.05	73.08
HLGAT [52]	36.79	43.35	24.31	33.14	35.25	46.39	65.43	76.33
SRN [35]	40.23	49.43	25.06	34.84	38.98	50.27	70.47	76.23
MoE [38]	45.12	53.33	28.49	37.19	44.46	57.96	75.64	80.27
Meta [37]	62.85	71.33	30.94	39.0	44.05	57.37	71.01	79.95
Our M_G	75.04	82.19	49.38	57.83	60.17	71.93	84.04	88.13

In Table 4 M_G attains 3.47% drop at *Rank*-1 accuracy than close-set setting#2. Therefore, the reasons for this drop are analyzed in retrieval results in Figure 12. In Figure 12a,b, it is evident that M_G is robust against style changes and impostors and finds matches at *Rank* = 2; however, in Figure 12c M_G lags in cross-modal matching due to impostors. Clearly, color images dominate intensity images, thus, to improve the multimodal recognition capability of metric M_G in the outdoor environment, it is needed to optimally represent each person in RGB, Grayscale, and Sketch modals during training; hence, M_G can resist a large number of multimodal impostors in outdoor environment.

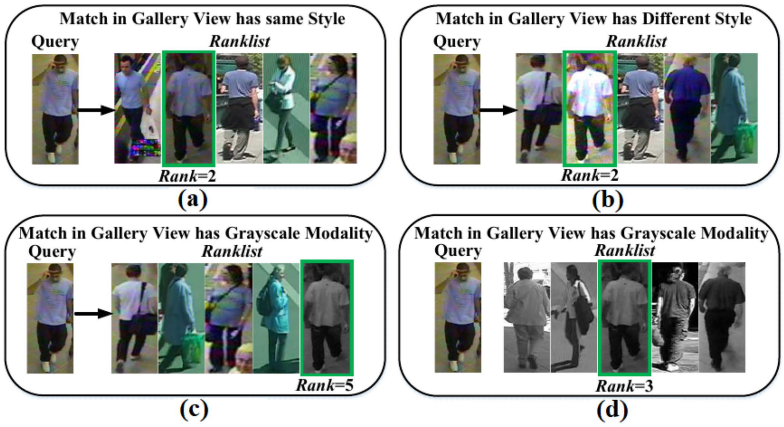


Figure 12. Retrieval results obtained in challenging close-set scenario. (**a**) Finding the match in Gallery view where the correct Gallery image has same style and modality with Query image; (**b**) Finding the match in Gallery view where the correct Gallery image has different style, but, same modality with Query image; (**c**) Finding the match in Gallery view where the correct Gallery image has different style, and different modality with Query image; (**d**) Finding the match in Gallery view where all the images in Gallery view has different style, and different modality with Query image.

Therefore, to find the optimal representation of persons in different modals, our work performed different experiments, and the results are shown in Figure 13. In Figure 13, for domains VIPeR and GRID, it is observed that M_G declines a large number of impostors when a number of images for a large number of training persons have representation ratio 1:1:3 for Grayscale (G.Sc.) vs. Sketch (Sk.) vs. RGB modals. Therefore, M_G is retrained with the representation ratio 1:1:3 to regain the performance, and in Figure 12d, M_G successfully declines impostors to find match at *Rank* = 3.

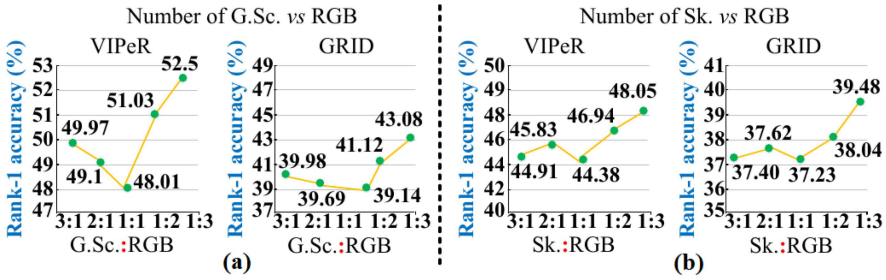


Figure 13. Rank-1 Result Comparison for Different Representation Ratios on VIPeR & GRID for (**a**) G.Sc. vs. RGB (**b**) Sk. vs. RGB.

4.3.3. Open-World Scenario

This is the scenario where the target person $P1$ in the real world moves out from the robot perception, and thus the robot losses the target person. While there is no target person in the robot perception, it is required that the robot following the person has robust reidentification capability to resist both impostors in the open world and at the same time has inherent discriminating ability to reidentify the real target person as soon as the target person is recaptured into the robot perception. Therefore, our work also evaluates metric M_G in the realistic open world. In Figure 14, we have shown the high-level overview of this scenario.

However, unlike [51], for the open-world testing in our work, randomly, 48 person IDs from each testing domain, i.e., from VIPeR, PRID, GRID, and i-LIDS domains are chosen to form the realistic world joint open-set Gallery, then M_G finds the matches for target Query images from the joint open-set Gallery, and the results are given in Table 5. In Table 5, M_G in the open world optimizes the part-attention weights in a way to learn cross-representations to simultaneously decline a large number of impostors in the open

world and also discriminates difficult nontarget Queries to attain 68.02%, 56.09%, and 76.57% *Rank*-1 identification at FTR 0.1% on PRID, VIPER, and i-LIDS, respectively.



Figure 14. Open-World Scenario where it is not necessary that all the Queries have match in Gallery view. Thus, the metric M_G needs to be robust against unknown persons seen in the open world while the target is moved out of perception.

Table 5. Results on Open-World Re-ID.

Methods	Target:PRID		Target:GRID		Target:VIPeR		Target:i-LIDS	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
CBAM [45]	26.95	51.00	18.43	34.26	23.48	44.38	41.07	68.98
APN [51]	28.63	53.29	19.34	36.60	24.80	46.21	43.32	70.87
RAGA [44]	31.05	55.74	19.92	38.34	27.42	49.49	45.82	73.03
HLGAT [52]	32.98	57.06	20.27	40.97	28.14	51.64	47.12	75.55
SRN [35]	37.34	60.49	23.13	44.17	30.43	53.42	50.38	77.00
MoE [38]	43.01	65.09	26.36	47.15	33.01	56.10	52.18	80.67
Meta [37]	45.70	65.39	28.43	49.68	32.11	55.27	51.35	80.29
Our M_G	68.02	80.86	41.3	54.4	56.09	78.1	76.57	86.42

Furthermore, in Figure 15, attention maps and the corresponding rise in *Rank*-1 identification accuracy are analyzed. In Figure 15b, it is revealed that M_G exploits part-attention module and learns cross-representations that declines large number of impostors and nontarget Queries in the open world; thus, *Rank*-1 accuracy at FTR 0.1% rises to 41.3% from 35.87% in Figure 15a.

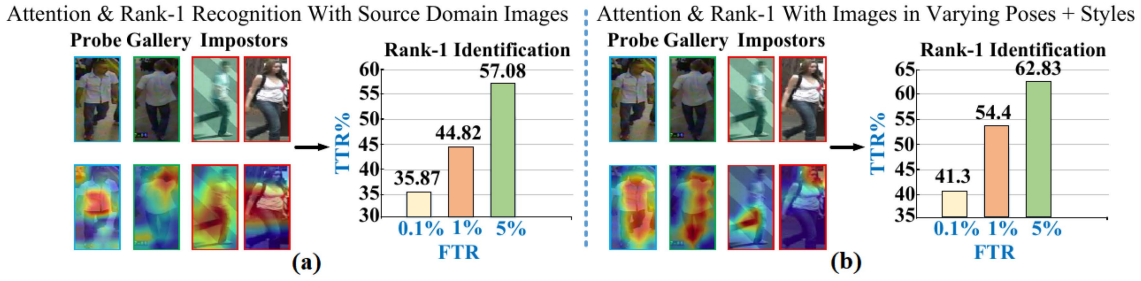


Figure 15. Rank-1 Result Comparison on VIPeR & GRID for (**a**) G.Sc. vs. RGB (**b**) Sk. vs. RGB.

4.3.4. Challenging Open-World Scenario

Though, metric M_G is evaluated in the open world; however, open-set Gallery in real-world is far more challenging where Probe-Gallery pairs can be seen in different modals and styles. In Figure 16, this complex scenario is shown visually.

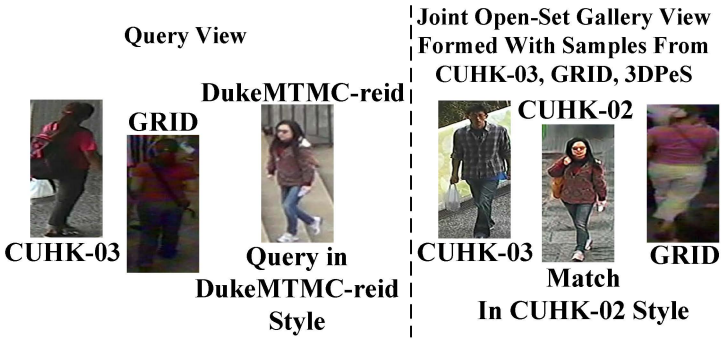


Figure 16. In the Challenging Open-World Scenario, the Query image can be seen in different domains, styles, or modalities, while the match is needed to be found from any other random domain, e.g., the CUHK-02 domain, while the matching image has a style or modality different from that of the Query image.

Therefore, images in Query and in joint Gallery are randomly transformed into different modals and styles during testing, then M_G finds the matches from the joint multimodal multistyle open-set Gallery, and the results are summarized in Table 6.

Table 6. Results on Challenging Open-World Re-ID.

Methods	Target:PRID		Target:GRID		Target:VIPeR		Target:i-LIDS	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
CBAM [45]	19.34	38.27	10.66	20.03	17.58	33.71	28.0	52.14
APN [51]	21.01	47.75	12.12	25.34	18.97	36.40	32.31	59.86
RAGA [44]	22.65	49.38	12.23	25.96	19.72	38.04	34.54	60.43
HLGAT [52]	23.54	51.03	14.18	28.81	21.17	40.47	36.84	64.46
SRN [35]	26.09	54.80	16.83	34.34	23.24	43.22	39.13	68.08
MoE [38]	29.37	59.0	19.04	38.71	25.79	48.35	42.86	71.53
Meta [37]	32.45	60.07	21.22	40.48	25.35	47.98	42.24	71.22
Our M_G	65.11	80.34	37.51	51.77	52.08	76.9	76.04	86.0

M_G, in contrast to [35,37,38,44], is an open-set metric with inherent ability to match cross-modal and cross-style Probe-Gallery pairs in the nonlinear outdoor environment, while the M_G also simultaneously rejects large number of impostors; therefore, in Table 6 M_G attains 65.11%, 37.51%, 52.08%, and 76.04% *Rank*-1 identification at FTR 0.1% on PRID, GRID, VIPeR, and i-LIDS datasets, respectively.

Furthermore, in Figure 17, attention maps and in Figure 18 retrieval results are analyzed in the open-world. Our attention maps in Figure 17 are more focused on individual parts and do not discard unique valuable cues from different filters, therefore, our attention maps are more robust against interenvironment and intraenvironment style and modality transforms than SRN [35]and RAGA [44]. Consequently, M_G successfully identifies cross-modal pair ($Q1$,$G1$), and cross-style pairs ($Q1$,$G2$) and ($Q2$,$G2$) at *Rank*=3, *Rank*=1, and *Rank*=1, respectively, in Figure 18; there exists large number of impostors in Gallery in scenario#1, scenario#2, and scenario#3, whereas [35,37,38] inherently lack matching cross-modal and cross-style images. In scenario#4 M_G, it also matches cross-style pair ($Q2$,$G3$) at *Rank*=2; even the $G3$ is seen in COCO domain style, where underlying nonlinear transforms and impostors in COCO domain affect the retrieval results.

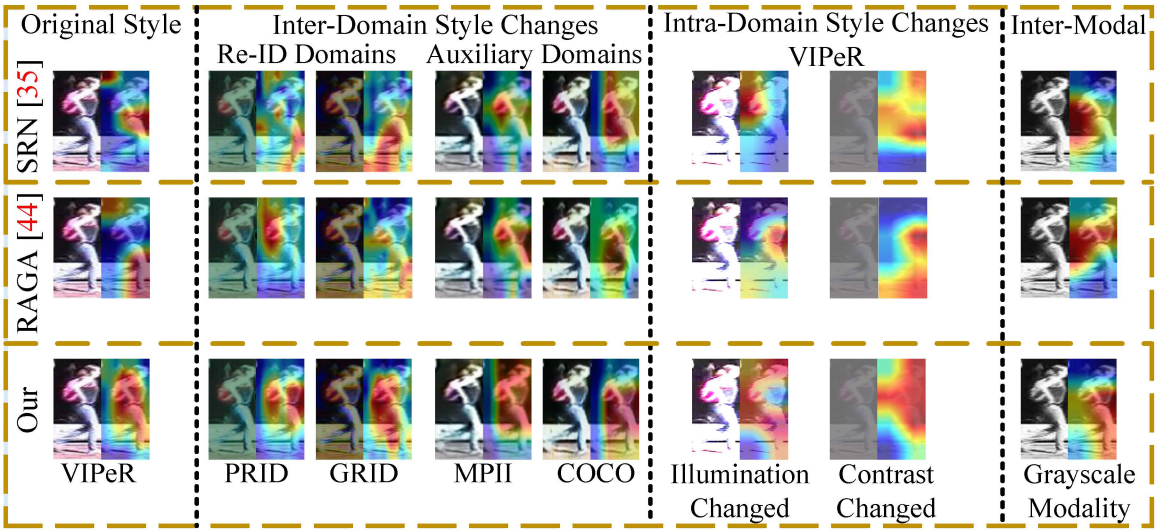


Figure 17. Comparison of Our Part Attention Maps with SRN [35] and RAGA [44].



Figure 18. Retrieval results of Our M_G, SRN [35], Meta [37], and MoE [38] in real-world scenarios. Green Rectangles contain correct matches.

4.3.5. Cross-Modal Scenario

Our M_G is also evaluated against RGB-Infrared matching on SYSU-MM01 dataset. M_G compared to [48,54] is trained in real multi-modal open-world to tackle complex nonlinear transforms and can resist large number of impostors; thus, in Table 7, M_G attains 64.93% and 72.58% *Rank*-1 identification on All-search and Indoor-Search, respectively.

Table 7. Cross-Modal Comparison on SYSU-MM01.

Methods	SYSU-MM01					
	All-Search			Indoor-Search		
	R = 1	R = 10	mAP	R = 1	R = 10	mAP
Hi-CMD [53]	34.94	77.58	35.94	-	-	-
GECNet [48]	53.37	89.86	51.83	60.60	94.29	62.89
cm-SSFT [54]	61.6	89.2	63.2	70.5	94.9	72.6
Our M_G	64.93	92.31	66.04	72.58	94.89	72.98

4.3.6. Computational Complexity

To run M_G in the real world, all convolutions are implemented as mixed depthwise separable convolutions [41]. The computation burden for depthwise separable convolution is $K \times K \times M \times H \times W + M \times N \times H \times W$, whereas the computation burden for standard convolutions are $K \times K \times M \times N \times H \times W$. If kernel size $K \times K$ is 3×3, and feature map dimensions $H \times W$ are 56×56, then, the cost of depthwise separable convolution is $3 \times 3 \times 256 \times 56 \times 56 + 256 \times 256 \times 56 \times 56 = 212{,}746{,}240$ multiplications, while the cost for standard convolutions are $3 \times 3 \times 56 \times 56 \times 256 \times 256 = 1849{,}688{,}064$ multiplications. Clearly, depthwise separable convolution lowers the computation cost by 88%, and thus M_G is realized on smart embedded camera $Hi3516DV300$.

4.3.7. Computation Time

Running time of metric M_G for different image sizes for one forward pass on $Hi3516DV300$ are given in Table 8, where for image size 224×224 M_G takes 29.4ms to process one single pass and obtain similarity.

Table 8. Computation Time for Different Image Sizes.

Image Size	Time/Image (ms)	Device Platform
324×324	52	$Hi3516DV300$
256×256	38.4	$Hi3516DV300$
224×224	29.4	$Hi3516DV300$

4.4. Reidentification-Based Tracking Experiments

To evaluate the learned generic metric M_G in the person tracking applications, our work performed several experiments in outdoor to track target. In Figure 19, the complete framework is shown where the learned reidentification metric M_G is integrated with pretrained CNN tracker [4].

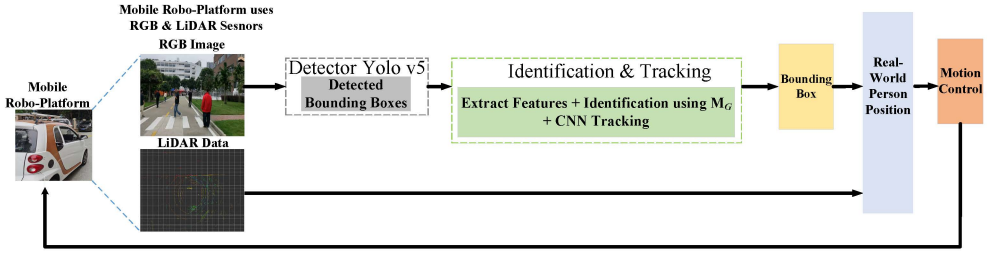


Figure 19. System Overview of Person-Following Robot in Our Work.

The input in Figure 19 is RGB image with LiDAR data, where the RGB image is first sent to Yolov5 detector [55] to obtain the bounding boxes. Reidentification module with the tracker in the next step in Figure 19 then takes the detected bounding boxes and identifies the target using the generic metric M_G and sends the identified box to the pretrained tracker [4] as the input dynamic template of the person to be tracked. To perform the identification of target, Re-ID metric M_G in our work uses prestored features of the target person. Finally, using the tracker prediction box and LiDAR data, the real-world position of the target is then updated. Motion control in last uses the updated position to generate the actuation signals for the robocar.

4.4.1. Experiment Setup and Testing Scenarios

All the experiments in our work are conducted in outdoor, where the person is tracked in different environments, while undergoing illumination and background variations.

In Figure 20, it can be seen the person is tracked in three challenging and complex outdoor environments, which are referred to as Scenario#1 (Seq#1), Scenario#2 (Seq#2), and

Scenario#3 (Seq#3). In each of these scenarios, the person is tracked for 115 s, 115 s, and 120 s, respectively, while the person in three different scenarios experiences crowding and occlusion with distractors (distractors in our work are the person occluding the target but are the not the impostors of target) and impostors in the scenes, and at several occasions, the person also completely moved out of the perception of mobile roboplatform. The mobile roboplatform visual system consists of RGB camera and LiDAR sensors, both are mounted on top of roboplatform, as shown in Figure 21.



Figure 20. Person-following roboplatform in outdoor world. (**a**) Following the target in black clothing on Road-1; (**b**) Following the target in crowded and occluded scenario on Road-2; (**c**) Continue following the target on Road-3 in varying pose; (**d**) Continue following the target on Road-3 even even there are distractor and impostor.



Figure 21. Robocar Robotic Platform mounted with Senors in Outside World.

Mounted RGB sensor take images with resolution 1080P with 30 frames per second. Furthermore, the depth and RGB data both are processed with on board CPU i7 8700k on the roboplatform, where the memory size is 16G.

4.4.2. Evaluation Metrics and Comparison

To make fair evaluation and comparison with other state-of-the-art works, our work uses four standard metrics [2] for quantitative evaluation. These metrics both evaluate re-identification performance and tracking success and are defined as: Correctly Identified and Tracked (CT): meaning the person is correctly identified and tracked successfully, Correctly Loss (CL): meaning it is correctly identified the person is not in the scene (either moved out or completely occluded) and thus successfully loss, Wrongly Identified and Tracked (WT): meaning the identification metric wrongly identified the impostor or distractor, and the tracker wrongly tracks the distractor or impostor, and the last metric is Wrongly Loss (WL): meaning the identification metric assumes there is no target in the scene; however, there is target present in the scene, and the tracker wrongly loss tracking.

4.4.3. Results and Analysis

Our work compares the performance with three state-of-the-art trackers, which are STARK [4], DiMP [5], and ATOM [6]. Table 9 below summarizes the results for tracking in all the three scenarios and presents the tracking results for STARK [4], DiMP [5], and ATOM [6], as well as results of our model for the evaluation metrics CT, CL, WT, and WL with the total tracking time in seconds and the percentage of identified and tracked frames in the given Seq#1, Seq#2, and Seq#3.

Table 9. Person Reidentification Results for Tracking in the Outdoor World.

		Methods							
		DiMP [5]		ATOM [6]		STARK [4]		Ours	
		Time (Sec)	Frames (%)	Time (Sec)	Frames (%)	Time (Sec)	Frames (%)	Time (Sec)	Frames (%)
Seq. #1	CT	52	45.2	57	49.57	62	53.91	80	69.6
	CL	20	17.39	17	14.78	16	13.92	25	21.74
	WT	13	11.29	15	13.05	13	11.31	6	5.21
	WL	30	26.09	26	22.61	24	20.87	4	3.48
Seq. #2	CT	31	26.96	44	38.26	56	48.69	73	63.48
	CL	17	14.78	10	8.69	12	10.43	25	21.74
	WT	22	19.13	19	16.52	15	13.04	8	6.96
	WL	45	39.13	42	27.83	32	26.08	9	7.82
Seq. #3	CT	28	24.35	38	33.04	40	34.78	69	60.02
	CL	14	12.17	8	6.95	10	8.69	29	25.22
	WT	26	22.61	23	20	20	17.39	12	10.43
	WL	52	45.22	51	44.35	50	43.48	9	7.83

Results in Seq#1: In Table 9 in Seq#1, our work Correctly Tracked (CT) the target person in 69.6% of total frames, whereas original STARK [4] with no identification module Correctly Tracked (CT) target person in 53.91% of total frames. This shows that tracking when complemented with reidentification module can largely improve the target tracking in different nonlinear scenes.

Furthermore, in Figure 22, visual comparison is shown, where it can be seen that when there are no distractors in the frames frame#90 and frame#146, then the identification of target and its tracking is easier; however, when the target is occluded with object in frame#155 in Figure 22, then a few trackers including STARK [4], DiMP [5], and ATOM [6] start tracking distractors. The reason is obvious: STARK [4], DiMP [5], and ATOM [6] all lack the identification ability to distinguish between target and distractor. In addition, when the target is completely moved out of the perception, such as in frames frame#164, frame#165, and frame#222 in Figure 22, then these trackers still continue wrongly tracking either distractors in frame#164 and in frame#165 or impostors in frame#222 since there are no reidentification model to verify if the detected person is a real target or impostor. In last, in Seq#1, it is also evaluated that how the model performs under the scenarios when the person is completely occluded with impostor, such as in frames frame#767 and frame#779. In Figure 22, in frame#767 and frame#779, it can be seen that our model with reidentification metric can successfully identify the person and thereby successfully update the dynamic image template of target to improve tracking, and in last, using both detected bounding box and LiDAR data it can robustly track the person during occlusion. In contrast, in Figure 22, when the person is occluded in frame#767 and frame#779, then DiMP [5] and ATOM [6] wrongly start tracking the impostor in frames frame#767 and frame#779.



Figure 22. Person-Following Tests in Scenario#1 (seq#1), where the Target moves out of Robocar perception and faces crowding and distractions.

Results in Seq#2: In Table 9 in Seq#2, our work Correctly Tracked (CT) the target person in 63.48% of total frames, whereas original STARK [4] with no identification module Correctly Tracked (CT) target person in 48.69% of total frames.

The results are lower than in Seq#1 because Seq#2 is more challenging than Seq#1, where in Seq#2 person moves in a varying illumination environment, while the background noise also affect reidentification, as shown in Figure 23. Furthermore, in Figure 23, the person is also occluded by distractor and impostor in the scene. In Seq#2 in frame#151 and frame#186, our reidentification metric M_G can identify the target successfully in the presence of impostor while other trackers DiMP [5] and ATOM [6] follow impostor person. Furthermore, in frame#191 in Figure 23, when both impostor and target are seen in the scene, the target is occluded by both impostor and distractor; then, in such a scenario, still our identification metric M_G can address both occlusion and illumination variation to identify the target and thereby improve its tracking. Similarly, in other complex scenes, when target is completely moved out of perception in frame#211 in Figure 23 and when the target is occluded by both impostor and distractor in frame#606 in Figure 23, then, in both scenes, our reidentification metric M_G helps tracker to not follow impostor, and the mobile roboplatform stops and wait for the target to reappear and identified.

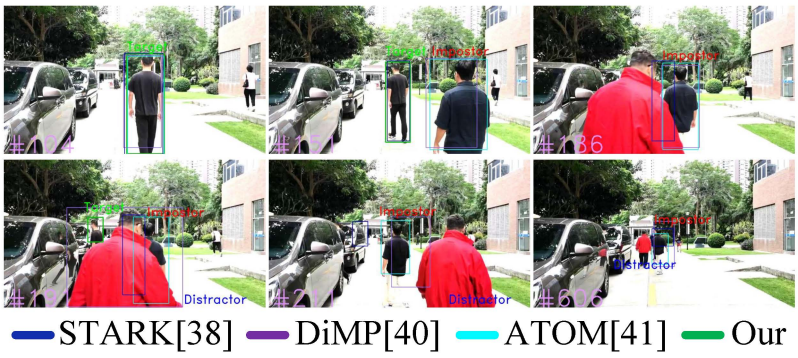


Figure 23. Person-Following Tests in Scenario#2 (seq#2), where the Target moves in varying illumination environment, and is occluded with impostors and distractors, while, Target also moves out of perception.

Results in Seq#3: In Table 9 in Seq#3, our work Correctly Tracked (CT) the target person in 60.02% of total frames, whereas original STARK [4] with no identification module Correctly Tracked (CT) target person in 34.78% of total frames. Seq#3 is far more challenging than Seq#2 and Seq#1, where in Seq#3 the person simultaneously undergoes pose and viewpoint changes, illumination and background variations, as well as, distracted and occluded by distractors and impostors, as shown in Figure 24. However, still, our reidentification metric M_G successfully identified and tracked the person in 60.02% of total frames, compared to the state of art STARK [4], DiMP [5] and ATOM [6]. This is mainly due to the correct identification of the target, which helps the tracking. In Figure 24, in frame#174 in Seq#3, the target person undergoing varying posture, while impostors are in the seen; however, still the learned generic metric M_G discriminates the target well. Furthermore, in frame#282 and frame#286 in Figure 24, when the target is fully occluded by distractor, while an impostor is nearby, still M_G successfully discriminates the impostor, whereas DiMP [5] and ATOM [6] track the impostors. In addition, in frame#293 in Figure 24, when the target reappears after full occlusion, then both our model and STARK [4] track the target well; however, STARK [4] has a little higher localization error in tracking than our model.

Last, in Figure 24 in frames frame#354 and frame#362, the target again occluded by impostor, while, also undergo illumination variations. Though, the situation is challenging, but the metric M_G is trained to address both illumination variations and style variations, and it is resistant against impostors; therefore, M_G continues identifying the person in consecutive frames, i.e., in frame#354, frame#362, and in frame#373.

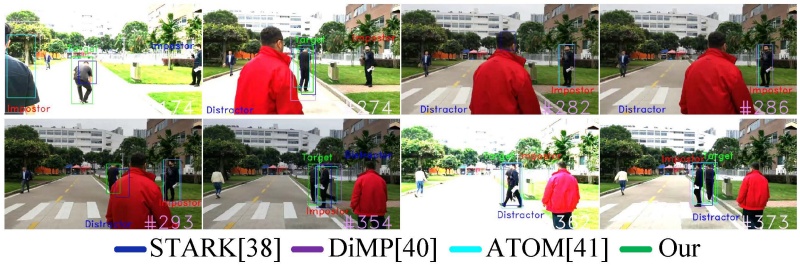


Figure 24. Person-Following Tests in Scenario#3 (seq#3), where the Target moves in varying illumination environment, while the Target is completely occluded with impostors and distractors, and also suffers due to changing backgrounds.

5. Conclusions & Future Directions

This work learns lightweight domain generic metric in the multimodal open world for person-following robots to address the practical world challenges face by person-following robots including nonlinear pose, viewpoint, style, and multimodal transforms, and a novel part-attention module is proposed to learn attention weighted cross-representations to address displacement and occlusion of parts. Thereby, the learned generic metric can resist large number of impostors and nontarget queries in the open world, while the learned metric is also lightweight and can run on robotic platform. Furthermore, future research will focus to improve the learned domain generic reidentification metric to solve multiscale and night reidentification problems for person-following robots.

Author Contributions: Conceptualization, M.A.S., Y.O., T.L., G.J.; methodology, M.A.S.; software, M.A.S.; validation, M.A.S., G.J.; writing—original draft preparation, M.A.S.; writing—review and editing, Y.O., T.L., G.J.; supervision, Y.O., T.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research is funded by the National Key Research and Development Program of China (No. 2018AAA0103001), National Natural Science Foundation of China (No. U1813208, 62173319, 62063006), Guangdong Basic and Applied Basic Research Foundation (No. 2020B1515120054), Shenzhen Fundamental Research Program (No. JCYJ20200109115610172).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Koide, K.; Miura, J. Identification of a specific person using color, height, and gait features for a person following robot. *Robot. Auton. Syst.* **2016**, *84*, 76–87. [CrossRef]
2. Koide, K.; Miura, J.; Menegatti, E. Monocular person tracking and identification with on-line deep feature selection for person following robots. *Robot. Auton. Syst.* **2020**, *124*, 103348. [CrossRef]
3. Ghimire, A.; Zhang, X.; Javed, S.; Dias, J.; Werghi, N. Robot Person Following in Uniform Crowd Environment. *arXiv* **2022**, arXiv:2205.10553.
4. Yan, B.; Peng, H.; Fu, J.; Wang, D.; Lu, H. Learning Spatio-Temporal Transformer for Visual Tracking. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 10448–10457.
5. Bhat, G.; Danelljan, M.; Gool, L.V.; Timofte, R. Learning Discriminative Model Prediction for Tracking. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6181–6190.
6. Danelljan, M.; Bhat, G.; Khan, F.S.; Felsberg, M. ATOM: Accurate Tracking by Overlap Maximization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 4655–4664.
7. Arras, K.O.; Mozos, O.M.; Burgard, W. Using Boosted Features for the Detection of People in 2D Range Data. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Rome, Italy, 10–14 April 2007; pp. 3402–3407.
8. Leigh, A.; Pineau, J.; Olmedo, N.; Zhang, H. Person tracking and following with 2D laser scanners. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Seattle, WA, USA, 26–30 May 2015; pp. 726–733.

9. Chen, B.X.; Sahdev, R.; Tsotsos, J.K. Person Following Robot Using Selected Online Ada-Boosting with Stereo Camera. In Proceedings of the Conference on Computer and Robot Vision (CRV), Edmonton, AB, Canada, 16–19 May 2017; pp. 48–55.
10. Munaro, M.; Menegatti, E. Fast RGB-D people tracking for service robots. *Auton. Robot.* **2014**, *37*, 227–242. [CrossRef]
11. Chen, Z.; Zhong, B.; Li, G.; Zhang, S.; Ji, R. Siamese Box Adaptive Network for Visual Tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 6668–6677.
12. Flores, S.; Jost, J. Person Re-Identification on a Mobile Robot Using a Depth Camera. In Proceedings of the IEEE International Symposium on Industrial Electronics (ISIE), Anchorage, AK, USA, 1–3 June 2022; pp. 115–122.
13. Tsai, T.H.; Yao, C.H. A robust tracking algorithm for a human-following mobile robot. *IET Image Process.* **2020**, *15*, 786–796. [CrossRef]
14. Pang, L.; Cao, Z.; Yu, J.; Guan, P.; Chen, X.; Zhang, W. A Robust Visual Person-Following Approach for Mobile Robots in Disturbing Environments. *IEEE Syst. J.* **2020**, *14*, 2965–2968. [CrossRef]
15. Koide, K.; Miura, J. Convolutional Channel Features-Based Person Identification for Person Following Robots. In *International Conference on Intelligent Autonomous Systems*; Springer: Berlin/Heidelberg, Germany, 2018; Volume 867, pp. 186–198.
16. Liu, J.; Chen, X.; Wang, C.; Zhang, G.; Song, R. A person-following method based on monocular camera for quadruped robots. *Biomim. Intell. Robot.* **2022**, *2*, 100058. [CrossRef]
17. Narayan, N.; Sankaran, N.; Setlur, S.; Govindaraju, V. Re-identification for Online Person Tracking by Modeling Space-Time Continuum. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 18–22 June 2018; pp. 1519–1527.
18. Nissimagoudar, P.C.; Iyer, N.C.; Gireesha, H.M.; Pillai, P.; Mallapur, S. Multi-pedestrian Tracking and Person Re-identification. In *International Conference on Soft Computing and Pattern Recognition*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 178–185.
19. Babaee, M.; Li, Z.; Rigoll, G. A dual CNN–RNN for multiple people tracking. *Neurocomputing* **2019**, *368*, 69–83. [CrossRef]
20. Neff, C.; Mendieta, M.; Mohan, S.; Baharani, M.; Rogers, S.; Tabkhi, H. REVAMP2T: Real-Time Edge Video Analytics for Multicamera Privacy-Aware Pedestrian Tracking. *IEEE Internet Things J.* **2020**, *7*, 2591–2602. [CrossRef]
21. Fleuret, F.; Shitrit, H.B.; Fua, P. Re-identification for Improved People Tracking. In *Person Re-Identification*; Springer: London, UK, 2014; pp. 309–330.
22. Liu, Q.; Chen, D.; Chu, Q.; Yuan, L.; Liu, B.; Zhang, L.; Yu, N. Online multi-object tracking with unsupervised re-identification learning and occlusion estimation. *Neurocomputing* **2022**, *483*, 333–347. [CrossRef]
23. Welsh, J.B. Real-Time Pose Based Human Detection and Re-identification with a Single Camera for Robot Person Following. Ph.D. Thesis, University of Maryland, College Park, MD, USA, 2017.
24. Thakran, A.; Agarwal, A.; Mahajan, P.; Kumar, S. Vision-Based Human-Following Robot. In Advances in Data Computing, Communication and Security. Springer: Singapore, 2022; pp. 443–449.
25. Algabri, R.; Choi, M.T. Deep-Learning-Based Indoor Human Following of Mobile Robot Using Color Feature. *Sensors* **2020**, *20*, 2699. [CrossRef] [PubMed]
26. Chebotareva, E.; Hsia, K.H.; Yakovlev, K.; Magid, E. Laser Rangefinder and Monocular Camera Data Fusion for Human-Following Algorithm by PMB-2 Mobile Robot in Simulated Gazebo Environment. In Proceedings of the 15th International Conference on Electromechanics and Robotics "Zavalishin's Readings", Ufa, Russia, 15–18 April 2020; pp. 357–369.
27. Condés, I.; Cañas, J.M. Person Following Robot Behavior Using Deep Learning. In *Advances in Physical Agents*; Springer: Cham, Switzerland, 2019; pp. 147–161.
28. Anuradha, U.A.D.N.; Kumari, K.W.S.N.; Chathuranga, K.W.S. Human detection and following robot. *Int. J. Sci. Technol. Res.* **2020**, *9*, 6359–6363.
29. Montesdeoca, J.; Toibero, J.M.; Jordan, J.; Zell, A.; Carelli, R. Person-Following Controller with Socially Acceptable Robot Motion. *Robot. Auton. Syst.* **2022**, *153*, 104075. [CrossRef]
30. Gemerek, J.; Ferrari, S.; Wang, B.H.; Campbell, M.E. Video-guided Camera Control for Target Tracking and Following. *IFAC-PapersOnLine* **2019**, *51*, 176–183. [CrossRef]
31. Yan, S.; Zhang, Y.; Xie, M.; Zhang, D.; Yu, Z. Cross-domain person re-identification with pose-invariant feature decomposition and hypergraph structure alignment. *Neurocomputing* **2022**, *467*, 229–241. [CrossRef]
32. Li, Y.; He, J.; Zhang, T.; Liu, X.; Zhang, Y.; Wu, F. Diverse part discovery: Occluded person re-identification with part-aware transformer. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 2898–2907.
33. Deng, W.; Zheng, L.; Ye, Q.; Guoliang, D.W.; Zheng, L.; Ye, Q.; Kang, G.; Yang, Y.; Jiao, J. Image-Image Domain Adaptation with Preserved Self-Similarity and Domain-Dissimilarity for Person Re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 994–1003.
34. Zhong, Z.; Zheng, L.; Zheng, Z.; Li, S.; Yang, Y. Camera Style Adaptation for Person Re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 5157–5166.
35. Jin, X.; Lan, C.; Zeng, W.; Chen, Z.; Zhang, L. Style Normalization and Restitution for Generalizable Person Re-Identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 3140–3149.
36. Jia, J.; Ruan, Q.; Hospedales, T.M. Frustratingly Easy Person Re-Identification: Generalizing Person Re-ID in Practice. In Proceedings of the British Machine Vision Conference (BMVC), Cardiff, UK, 9–12 September 2019; pp. 1–14.

37. Choi, S.; Kim, T.; Jeong, M.; Park, H.; Kim, C. Meta Batch-Instance Normalization for Generalizable Person Re-Identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 3424–3434.
38. Dai, Y.; Li, X.; Liu, J.; Tong, Z.; Duan, L.Y. Generalizable Person Re-identification with Relevance-aware Mixture of Experts. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 16140–16149.
39. Zhu, Z.; Huang, T.; Shi, B.; Yu, M.; Wang, B.; Bai, X. Progressive Pose Attention Transfer for Person Image Generation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 2347–2356.
40. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
41. Tan, M.; Le, Q.V. MixConv: Mixed Depthwise Convolutional Kernels. In Proceedings of the British Machine Vision Conference (BMVC), Cardiff, UK, 9–12 September 2019; pp. 1–13.
42. Zagoruyko, S.; Komodakis, N. Wide Residual Networks. In Proceedings of the British Machine Vision Conference (BMVC), York, UK, 19–22 September 2016; pp. 1–12.
43. Ma, N.; Zhang, X.; Zheng, H.T.; Sun, J. ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 122–138.
44. Zhang, Z.; Lan, C.; Zeng, W.; Jin, X.; Chen, Z. Relation-Aware Global Attention for Person Re-Identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 3183–3192.
45. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
46. Wang, F.; Zuo, W.; Lin, L.; Zhang, D.; Zhang, L. Joint Learning of Single-Image and Cross-Image Representations for Person Re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1288–1296.
47. Chen, W.; Chen, X.; Zhang, J.; Huang, K. Beyond Triplet Loss: A Deep Quadruplet Network for Person Re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1320–1329.
48. Zhong, X.; Lu, T.; Huang, W.; Ye, M.; Jia, X.; Lin, C.W. Grayscale Enhancement Colorization Network for Visible-infrared Person Re-identification. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 1418–1430. [CrossRef]
49. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
50. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1026–1034.
51. Li, X.; Wu, A.; Zheng, W.S. Adversarial Open-World Person Re-Identification. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 287–303.
52. Zhang, Z.; Zhang, H.; Liu, S. Person Re-identification using Heterogeneous Local Graph Attention Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 12131–12140.
53. Choi, S.; Lee, S.; Kim, Y.; Kim, T.; Kim, C. Hi-CMD: Hierarchical Cross-Modality Disentanglement for Visible-Infrared Person Re-Identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10254–10263.
54. Lu, Y.; Wu, Y.; Liu, B.; Zhang, T.; Li, B.; Chu, Q.; Yu, N. Cross-Modality Person Re-Identification With Shared-Specific Feature Transfer. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 13376–13386.
55. Jocher, G. YOLOv5 by Ultralytics. Released date: 2020-5-29. Available online: https://github.com/ultralytics/yolov5 (accessed on 1 November 2022).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.