



## Pedestrian Re-Identification and Tracking Algorithm Based on Cross-Domain Adaptation

Ting Dong<sup>1,2</sup>, Mary Jane C. Samonte<sup>1\*</sup>

<sup>1</sup> School of Information Technology, Mapua University, Manila 1205, Philippines

<sup>2</sup> School of Information Engineering, Yulin University, Yulin 719000, China

Corresponding Author Email: [mjcsamonte@126.com](mailto:mjcsamonte@126.com)

Copyright: ©2024 The authors. This article is published by IETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/ts.410516>

### ABSTRACT

**Received:** 6 May 2024

**Revised:** 6 September 2024

**Accepted:** 12 October 2024

**Available online:** 31 October 2024

#### Keywords:

*pedestrian re-identification, pedestrian tracking, cross-domain adaptation, generative adversarial networks, self-supervised learning*

With the increasing demand for intelligent surveillance and public safety, pedestrian re-identification and tracking technology has become a focal point in the field of computer vision. Traditional algorithms for pedestrian re-identification and tracking exhibit significant performance degradation when applied to cross-domain scenarios, such as those involving different surveillance devices or varying lighting conditions. Although existing studies have made some progress through the use of deep learning techniques, challenges remain in enhancing cross-domain adaptability. To address this issue, this study proposes a pedestrian re-identification image keypoint detection method based on adversarial generative domain adaptation networks, as well as a pedestrian re-identification and tracking algorithm based on deep self-supervised adversarial domain adaptation networks. By combining generative adversarial networks (GANs) with self-supervised learning, the proposed method significantly improves the accuracy and robustness of pedestrian re-identification and tracking in complex cross-domain environments, demonstrating high practical value and applicability.

## 1. INTRODUCTION

In recent years, with the widespread adoption of intelligent surveillance technologies and the growing demand for public safety [1-4], pedestrian re-identification and tracking have garnered increasing attention as core issues within the field of computer vision [5, 6]. Pedestrian re-identification aims to recognize the same individual from images or videos captured from different viewpoints and at different times, while pedestrian tracking involves the continuous localization and tracking of the same target throughout a video sequence. These technologies hold significant potential for applications in smart cities, security surveillance, and public safety. However, the cross-domain challenges present in real-world scenarios, such as the varying distribution of monitoring devices and changes in environmental lighting conditions, substantially increase the complexity and difficulty of pedestrian re-identification and tracking tasks [7-10].

The research on pedestrian re-identification and tracking algorithms is not merely focused on solving technical problems but also serves as a critical approach to enhancing public safety and optimizing resource allocation in the current social context [11, 12]. As technology advances and practical demands continue to grow, existing algorithms for pedestrian re-identification and tracking have become insufficient to meet the needs of complex environments. In particular, traditional algorithms exhibit significant performance degradation in cross-domain scenarios due to their strong dependency on environmental conditions [13-16]. Therefore, developing an

algorithm that can efficiently adapt to cross-domain scenarios is of paramount importance for improving the reliability and efficiency of public safety systems.

Despite the progress made in pedestrian re-identification and tracking algorithms, most existing methods still exhibit notable deficiencies in cross-domain adaptability [17, 18]. Traditional approaches typically rely on large-scale annotated data for training, lacking the ability to effectively extract and adapt to target domain features, resulting in poor performance in cross-domain applications [19-22]. Furthermore, although recent deep learning methods have somewhat improved the accuracy of re-identification and tracking, they still struggle to adequately address the disparities between domains in complex and dynamic cross-domain scenarios.

This study primarily focuses on two key aspects. First, a pedestrian re-identification image keypoint detection method based on adversarial generative domain adaptation networks was proposed, which effectively improved re-identification accuracy in cross-domain scenarios through the adaptive capabilities of GANs. Second, a pedestrian re-identification and tracking algorithm based on deep self-supervised adversarial domain adaptation networks was designed, aiming to enhance model adaptability in the target domain through the combination of self-supervised learning and adversarial training. This study not only introduces innovative algorithm designs but also demonstrates practical effectiveness, offering significant theoretical and practical value for enhancing the cross-domain adaptability of pedestrian re-identification and tracking systems.

## 2. PEDESTRIAN RE-IDENTIFICATION IMAGE KEYPOINT DETECTION BASED ON ADVERSARIAL GENERATIVE DOMAIN ADAPTATION NETWORKS

### 2.1 Problem description

In real-world applications of pedestrian re-identification and tracking, surveillance systems often encounter complex cross-domain challenges. These challenges arise from variations in camera angles, lighting conditions, and background complexity, leading to significant marginal distribution differences between source domain data (query images) and target domain data (pedestrian re-identification candidate images). Specifically, the distributions of the source domain data  $F_T = \{(a'_u, b'_u)\}_{u=1}^v$  and the target domain data  $F_S = \{(a^s_k, b^s_k)\}_{k=1}^{vs}$  differ substantially, which severely limits the performance of traditional pedestrian re-identification algorithms in the target domain.

To address this challenge, a pedestrian re-identification image keypoint detection method based on adversarial generative domain adaptation networks was proposed in this study. Initially, a generator  $h(\cdot)$  and a detector  $d_f(\cdot)$  were collaboratively trained to transform the source domain data  $a^t$  into pseudo-target domain data  $a^d$ , such that the distribution  $o(a^d)$  of the pseudo-target domain closely approximates the distribution  $o(a^s)$  of the target domain. Subsequently, the pseudo-target domain data  $F_d$  and the target domain data  $F_s$  were utilized to train the keypoint detector  $d_f(\cdot)$ . By training on the combined dataset  $\{F_d, F_s\}$ , the keypoint detector was enabled to learn features that are more adaptive to the target domain. This ultimately allows for precise keypoint detection on target domain data  $a^t$  during actual pedestrian re-identification and tracking tasks, achieving  $b^s = d_f(a^s)$ .

### 2.2 Overall network architecture

In pedestrian re-identification and tracking tasks, domain discrepancies in cross-domain scenarios significantly impact the accuracy of identification and tracking. To effectively address these challenges, the proposed method leverages GANs to reduce distribution differences between domains, thereby enhancing the model's adaptability and recognition accuracy in the target domain. As a reference network, the CycleGAN architecture serves as an important foundation for this study.

The CycleGAN network architecture comprises two generators,  $H_{S2OR}$  and  $H_{OR2S}$ , and two discriminators,  $F_{OR}$  and  $F_{SIM}$ . The generator  $H_{S2OR}$  is responsible for converting the image style of the source domain into that of the target domain, while the generator  $H_{OR2S}$  performs the reverse task, converting the image style of the target domain into that of the source domain. The discriminators  $F_{OR}$  and  $F_{SIM}$  are tasked with distinguishing whether the input images originate from the target domain or the source domain, respectively. The optimization process of CycleGAN primarily relies on two loss functions: the adversarial loss  $M_{ADV}$  and the cycle consistency loss  $M_{CL}$ . The adversarial loss  $M_{ADV}$  ensures that the generated images closely resemble those of the target domain, making it difficult for the discriminators to differentiate between original and generated images. Meanwhile, the cycle consistency loss  $M_{CL}$  maintains the content consistency of the generated images by ensuring that the images retain their original content features after passing through both generators. The definitions of these loss

functions are as follows:

$$M_{ADV} = R_{a_{OR} \sim A_{OR}} \left[ \log(F_{OR}(a_{OR})) \right] + R_{a_{SIM} \sim A_{SIM}} \left[ 1 - \log(F_{OR}(a'_{OR})) \right] \quad (1)$$

$$M_{CY} = R_{a_{OR} \sim A_{OR}} \left[ \|a''_{OR} - a_{OR}\| \right] + R_{a_{SIM} \sim A_{SIM}} \left[ \|a''_{SIM} - a_{SIM}\| \right] \quad (2)$$

The total loss function of CycleGAN is defined as:

$$M_{CG} = \underset{H_{SIM2OR}}{MIN} \underset{H_{OR2SIM}}{MAX} \underset{H_{OR}}{MAX} \left[ \begin{array}{l} M_{ADV}(H_{SIM2OR}, F_{OR}) \\ + M_{ADV}(H_{OR2SIM}, F_{SIM}) \\ + M_{CY} \end{array} \right] \quad (3)$$

In practical applications of pedestrian re-identification and tracking, cross-domain issues often lead to distributional differences between the source and target domains, thereby affecting the model's detection performance. To effectively address this challenge, an improved network architecture based on CycleGAN, termed RegCycleGAN, was proposed in this study. The architecture of RegCycleGAN is illustrated in Figure 1. RegCycleGAN builds upon the CycleGAN framework by incorporating four additional pedestrian re-identification keypoint detectors:  $DET_{S-S}$ ,  $DET_{S-T}$ ,  $DET_{O-S}$ , and  $DET_{O-T}$ .  $DET_{S-S}$  and  $DET_{S-T}$  are responsible for keypoint detection in source domain images, while  $DET_{O-S}$  and  $DET_{O-T}$  are used for keypoint detection in target domain images. To achieve stable pedestrian re-identification keypoint detection, RegCycleGAN introduces the Exponential Moving Average (EMA) method, which facilitates parameter transfer between the student and teacher networks. Specifically, the parameters of the student networks are progressively updated and transferred to the teacher networks via the EMA method between  $DET_{S-S}$  and  $DET_{S-T}$ , as well as between  $DET_{O-S}$  and  $DET_{O-T}$ . This approach ensures that the teacher networks maintain relatively stable parameters during the training process. By employing this method, the model becomes more robust to data perturbations during training, leading to more stable and accurate keypoint detection results.

To enhance the performance of cross-domain pedestrian re-identification, the proposed RegCycleGAN network introduces multiple loss functions to optimize various components of the model, thereby enabling efficient pedestrian re-identification keypoint detection. The network first incorporates the CycleGAN loss function  $M_{CG}$ , which ensures that the images generated through the generators retain their original content and structure after being transformed across different domains. Furthermore, the network introduces the query domain detector loss function  $M_{D-O}$  and the re-identification candidate domain detector loss function  $M_{D-S}$ , corresponding to the optimization objectives of the keypoint detectors in the two domains.  $M_{D-O}$  optimizes keypoint detection performance in the query domain, while  $M_{D-S}$  optimizes keypoint detection performance in the candidate domain.

Additionally, the self-consistency loss function  $M_{S-C}$  was proposed within the RegCycleGAN network to enhance the stability of the teacher-student detector networks in keypoint detection tasks. By introducing this loss function between the

teacher and student networks, RegCycleGAN ensures consistency in parameter updates, thereby improving the stability and accuracy of keypoint location detection. Assuming  $\beta_1$  and  $\beta_2$  are balancing factors, the overall loss

function for RegCycleGAN is defined as:

$$M_{RCG} = M_{CG} + \beta_1 M_{D-o} + \beta_2 M_{D-s} \quad (4)$$

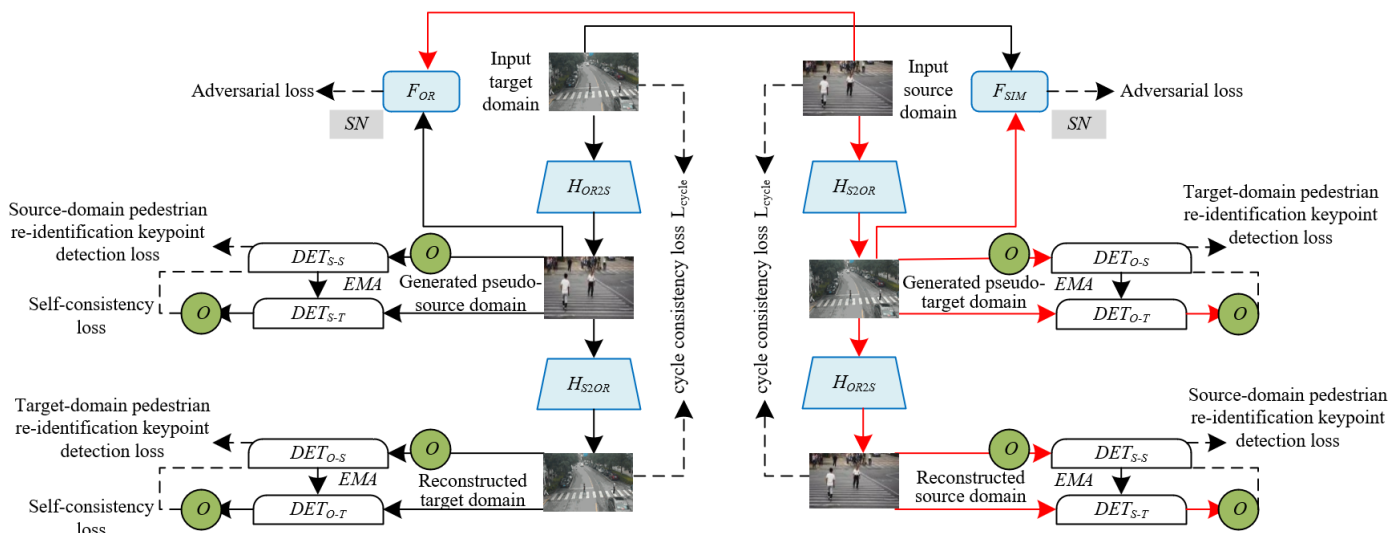


Figure 1. RegCycleGAN network architecture

### 2.3 Equivalent perturbation transformation

The pedestrian re-identification keypoint detector comprises two components: the teacher network  $d(\phi')$  and the student network  $d(\phi)$ . Both networks share the same UNet architecture, which effectively captures keypoint information in pedestrian images for pedestrian re-identification tasks and processes it at different scales. During the training phase, an equivalent perturbation transformation  $po$  was introduced into the input data  $a$  to enhance the model's robustness to data perturbations. This perturbation transformation simulates the style differences and noise between different domains, enabling the model to better adapt to cross-domain data. The perturbed input data was then fed into both the teacher network  $d(\phi')$  and the student network  $d(\phi)$ , each producing keypoint detection results for the pedestrian images. Furthermore, a self-consistency loss function was employed to measure the consistency between the outputs of the teacher and student networks. The design of the self-consistency loss ensures that the outputs of the teacher and student networks remain consistent when faced with the same input data, thereby improving the stability of keypoint detection.

The design of the equivalent perturbation transformation module aims to simulate environmental changes in cross-domain scenarios by artificially introducing perturbations, thereby enhancing the model's robustness to cross-domain data. In practice, the equivalent perturbation transformation  $o$  is applied to the input data of the student network, which can include various forms of transformations such as image rotation, scaling, colour shift, and noise addition. Through these perturbation operations, the student network learns to extract pedestrian keypoint information consistently under varying conditions. Simultaneously, to ensure consistency between the outputs of the teacher and student networks, the equivalent perturbation transformation  $o$  is also applied to the output results of the teacher network. This consistent perturbation treatment allows the outputs of the teacher and student networks to be compared under the same conditions, effectively facilitating the calculation of the self-consistency

loss. Minimizing the self-consistency loss is the core objective of the equivalent perturbation transformation module. During training, the parameters of the student network were adjusted based on the gradient information from the self-consistency loss, thereby enhancing its ability to adapt to perturbed inputs. On the other hand, the parameters of the teacher network were updated using the EMA method, wherein the parameters are not directly updated through gradient descent but are progressively smoothed based on the parameters of the student network. This update strategy maintains the stability of the teacher network, allowing it to retain high detection accuracy even when confronted with perturbed inputs. Letting  $\beta$  be the smoothing coefficient controlling the parameter updates of the teacher network, the update formula is given by:

$$\phi'_s = \beta\phi'_{s-1} + (1-\beta)\phi_s \quad (5)$$

The definition of the self-consistency loss function is given by:

$$M_{S-C} = (1/|V|) \sum_{u \in V} \|d(o(a), \phi') - d(a, \phi)\|^2 \quad (6)$$

In cross-domain pedestrian re-identification and tracking tasks, accurately locating keypoints in pedestrian images is crucial for improving re-identification performance. However, since keypoint detection is a pixel-level task, local feature information in images is highly sensitive. Traditional coordinate regression methods struggle to cope with complex variations and noise in images. To overcome the convergence challenges in keypoint detection tasks, a heatmap was employed as the training label in the pedestrian re-identification model based on adversarial generative domain adaptation networks. This design effectively addresses the difficulties in model convergence during keypoint detection and enhances overall performance. Assuming that the keypoint coordinates, represented by the center of the heatmap, are denoted as  $(a_u, b_u)$ , and that the standard deviation of the

Gaussian distribution is denoted as  $\delta$ , the generation of the heatmap is defined as follows:

$$g(a_u, b_u) = \exp\left[-(1/2\delta^2)\left((a-a_u)^2 + (b-b_u)^2\right)\right] \quad (7)$$

Assuming the heatmap label is denoted as  $b$  and the student network as  $d(\phi)$ , the loss function  $M_{LA}$  for the keypoint detection network is defined as follows:

$$M_{LA} = (1/|V|) \sum_{u \in V} \|d(a, \phi) - b\|^2 \quad (8)$$

Let  $\eta(s) = e^{-5(1-s/S)^2}$ , where  $S$  represents the acceleration factor. The overall consistency regularization loss function  $M_{DE}$  for the pedestrian re-identification keypoint detection network is defined as follows:

$$M_{DE} = M_{LA} + \eta(s)M_{s-c} \quad (9)$$

In cross-domain pedestrian re-identification and tracking tasks, the application of GANs offers new possibilities for enhancing the robustness and adaptability of models in complex environments. However, GAN models often face the challenge of imbalanced training between the generator and discriminator, particularly when the discriminator becomes too powerful or the generator struggles to keep pace, which leads to vanishing gradients in the generator, making it difficult to improve the quality of the generated images. This issue is especially critical in pedestrian re-identification keypoint detection, where the accuracy and stability of detection heavily rely on the quality of generated images. Poor-quality generated images directly impact the model's detection precision and stability. To address this problem, spectral normalization (SN) from Lipschitz constraints was introduced to constrain the discriminator within the pedestrian re-identification model based on adversarial generative domain adaptation networks. The core idea of SN is to normalize the weights of each convolutional layer in the discriminator, controlling its spectral norm, thereby smoothing the network's derivatives and ensuring 1-Lipschitz continuity. Assuming the dual norm of  $Q$  is denoted by  $\delta(Q)$ , the formula is given by:

$$Q_{TN}(Q) = Q/\delta(Q) \quad (10)$$

### 3. PEDESTRIAN RE-IDENTIFICATION AND TRACKING BASED ON DEEP SELF-SUPERVISED ADVERSARIAL DOMAIN ADAPTATION NETWORKS

#### 3.1 Problem description

In cross-domain pedestrian re-identification and tracking tasks, the primary challenge lies in effectively utilizing data from both the source and target domains, despite the significant differences in their marginal distributions. Specifically, the source domain data is labelled and can be used for supervised learning, while the target domain data is unlabelled. In this scenario, traditional pedestrian re-identification methods are difficult to apply directly, as they rely on the consistency of feature spaces between the target

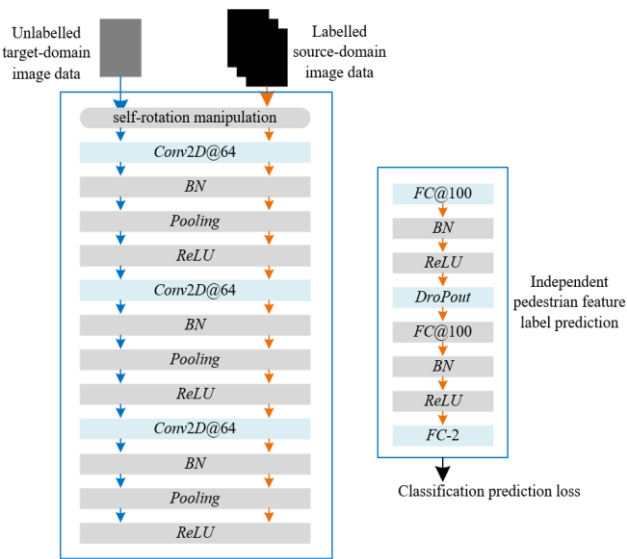
and source domains. However, real-world pedestrian re-identification tasks often involve scenarios that span multiple cameras, different environments, and even varying times, leading to substantial differences in data distribution between the source and target domains.

To address this issue, a pedestrian re-identification model based on deep self-supervised adversarial domain adaptation networks was proposed in this study. The model first generates domain soft labels  $F \in [0,1]$  for each sample through a feature confusion mechanism, where a label value of 1 indicates that the sample features are closer to the source domain, and a value of 0 indicates that the sample features are closer to the target domain. During this process, the model learns how to confuse the features of different domains, bringing the distributions of the source and target domains closer in the feature space. Simultaneously, the model generates self-rotation labels  $E = \{0,1,2,3\}$  for each sample through a self-supervised rotation proxy task, where these labels represent the different states of the sample after being rotated clockwise by  $0^\circ$ ,  $90^\circ$ ,  $180^\circ$ , and  $270^\circ$ , respectively. Through this proxy task, the model is able to learn more robust feature representations without relying on label information. During training, the model optimizes the feature extractor by predicting domain soft labels and self-rotation labels, enabling it to accurately classify the source domain while also developing strong cross-domain generalization capabilities. Finally, the model utilizes the classifier  $d(\cdot)$ , trained on the source domain, to classify samples in the target domain, i.e.,  $b^s = d(a^s)$ , thereby enabling effective application of pedestrian re-identification tasks in the target domain.

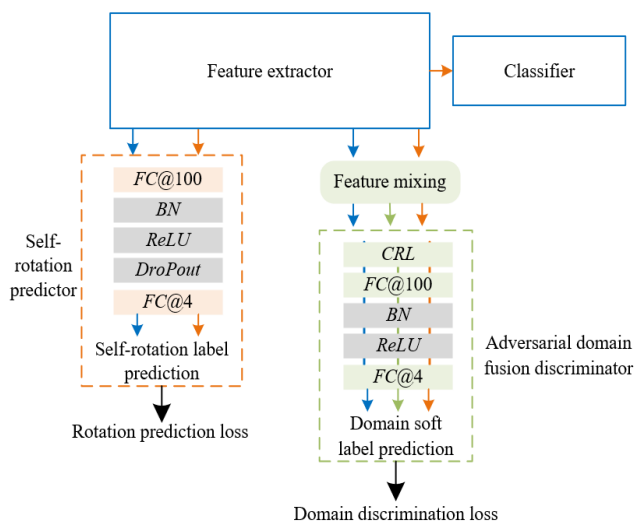
#### 3.2 Overall network architecture

In cross-domain pedestrian re-identification and tracking tasks, it is crucial to effectively address the distribution differences between the source and target domains. To tackle this issue, deep self-supervised adversarial domain adaptation networks were proposed in this study, which enhance the model's re-identification capabilities in the target domain by combining self-supervised learning and adversarial learning. Figure 2 illustrates the network architecture of the feature extractor and classifier within the model. Figure 3 depicts the overall architecture of the deep self-supervised adversarial domain adaptation networks. This network architecture comprises four main components: the feature extractor  $D_d$ , the classifier  $Z_b$ , the adversarial domain fusion discriminator  $Z_f$ , and the self-rotation predictor  $Z_e$ . The feature extractor  $D_d$  is responsible for extracting low-level features from the input images, which serve as the foundation for subsequent processing. In pedestrian re-identification tasks, the input typically includes query images with unique pedestrian feature labels and candidate images without distinct pedestrian feature labels. The feature extractor captures the foundational features of these images, providing a unified feature representation for the entire network. The extracted features are fed into three different sub-networks: the classifier  $Z_b$ , the adversarial domain fusion discriminator  $Z_f$ , and the self-rotation predictor  $Z_e$ . The primary function of the adversarial domain fusion discriminator  $Z_f$  is to perform domain discrimination on the fused features via a gradient reversal layer (GRL), predicting whether these features originate from the source or target domain. Through this mechanism, the network learns feature representations that are consistent across domains, thereby reducing the feature distribution discrepancies between the

source and target domains. Meanwhile, the self-rotation predictor  $Z_e$  further enhances the robustness of the features by predicting the rotation angle ( $0^\circ$ ,  $90^\circ$ ,  $180^\circ$ ,  $270^\circ$ ) of the input images. The self-rotation proxy task involves randomly rotating the input images and generating corresponding rotation labels, allowing the network to learn effectively without explicit labels. This self-supervised mechanism enhances the network's ability to capture image features, especially when target domain data lacks labels, helping the model learn the target domain's characteristics more effectively. Finally, the classifier  $Z_b$  uses the features processed by the adversarial domain fusion discriminator and the self-rotation predictor to predict the unique pedestrian feature labels.  $Z_b$  can more effectively perform classification tasks in the target domain by utilizing the semantically consistent features extracted by the feature extractor  $D_d$ . This approach ensures that even when there are significant distributional differences between images from the source and target domains, the model can still effectively identify pedestrians in the target domain by extracting features consistent across domains.



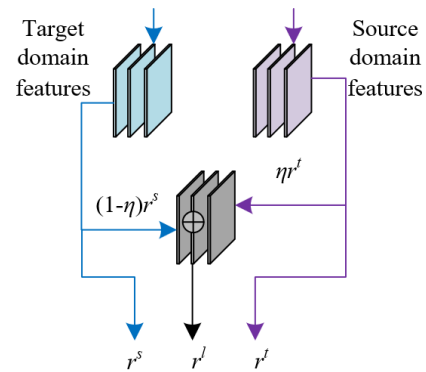
**Figure 2.** Network architecture of the feature extractor and classifier



**Figure 3.** Network architecture of deep self-supervised adversarial domain adaptation networks

### 3.3 Domain feature fusion

In the proposed model, the feature extractor  $D_d$  first extracts feature vectors from the input images of the source and target domains, denoted as the source domain feature vector  $r^s$  and the target domain feature vector  $r^t$ , respectively. However, directly aligning these two feature vectors in the high-dimensional feature space may not sufficiently capture their underlying similarities and differences. To address this, a feature-level mixing module was proposed in this study, which embeds features from different domains into a latent feature space for mixing, as illustrated in Figure 4. This feature fusion is achieved through an adversarial learning approach. In the latent feature space, the source domain and target domain feature vectors  $r^s$  and  $r^t$  are mixed to form a new feature representation. This representation not only retains the individual feature information of both domains but also reinforces their semantic consistency through the fusion process. During this process, the network optimizes the feature alignment by discriminating the domain soft labels of the mixed features, predicting the origin domain based on the feature distribution. The domain labels  $m^s_{DO}$  and  $m^t_{DO}$  represent the data from the source and target domains, with values of 1 and 0, respectively.



**Figure 4.** Schematic diagram of the feature fusion module

To further enhance the alignment of the model, the domain discriminator measures the similarity between the two domains using the features generated by the feature-level mixing module. In this way, the network can not only identify the differences between the source and target domain features but also automatically learn the semantic consistency features shared between the two domains. This alignment method directs the model's focus towards shared features across domains rather than being distracted by domain-specific noise. Assuming the feature fusion ratio is represented by  $\eta \in [0, 1]$ , the generated feature fusion vector  $r^f$  and the corresponding domain soft label  $m^l_{DO}$  are defined as follows:

$$r^f = \eta r^t + (1 - \eta) r^s \quad (11)$$

$$m^l_{DO} = \eta \quad (12)$$

In the model, for each input image  $a_u$ , a feature vector was first extracted using the feature extractor  $D_d$ . This feature vector contains the foundational feature information of the input image. These feature vectors were then sent to the domain discriminator  $Z_f$  for further processing. To enhance the alignment of domain features, the feature vectors pass through

a GRL before entering the discriminator. This layer reverses the gradient during training, aiding the network in learning more general features that are applicable across both the source and target domains. The input features received by the domain discriminator  $Z_f$  include the source domain feature vector  $r^t$ , the target domain feature vector  $r^s$ , and the generated fusion feature vector  $r^l$ . The fusion feature vector  $r^l$ , generated in the latent feature space, was designed to smoothly transition the features between the source and target domains, making the feature alignment process more robust. During this process,  $Z_f$  aligns domain features by distinguishing whether these feature vectors originate from the source domain, the target domain, or a mixture of both. To optimize the discrimination performance of  $Z_f$ , the model employs domain soft labels for standard supervised training. Specifically, by minimizing the cross-entropy loss,  $Z_f$  learns how to differentiate between the source domain, the target domain, and the fused features. This loss function guides  $Z_f$  to more accurately determine the domain origin of the features during training. Simultaneously, the GRL applies gradient reversal to the feature extractor  $D_d$ , gradually enabling  $D_d$  to extract more generalised cross-domain features. Ultimately, the optimized  $D_d$  can extract features that maintain semantic consistency between the source and target domains, thereby improving the re-identification accuracy of the model in the target domain. The loss functions corresponding to the three types of feature vector discrimination are defined as follows:

$$M_{S-D}^t = -R_{a^t-o_t} (1 - m_{DO}^t) \log(1 - Z_f(r^t)) + m_{DO}^t \log Z_f(r^t) \quad (13)$$

$$M_{S-D}^s = -R_{a^s-o_s} (1 - m_{DO}^s) \log(1 - Z_f(r^s)) + m_{DO}^s \log Z_f(r^s) \quad (14)$$

$$M_{S-D}^l = -R_{a^l-o_t, a^s-o_s} (1 - m_{DO}^l) \log(1 - Z_f(r^v)) + m_{DO}^l \log Z_f(r^v) \quad (15)$$

Assuming the balancing factor is denoted by  $\alpha$ , the total loss function  $M_{S-D}$  of the domain discriminator is defined as follows:

$$M_{S-D} = M_{S-D}^t + M_{S-D}^s + \alpha M_{S-D}^l \quad (16)$$

### 3.4 Self-supervised rotation transformation

In cross-domain pedestrian re-identification and tracking tasks, self-supervised learning is a critical method for enhancing the model's generalization ability, particularly when there is a lack of target domain data. Specifically, the self-supervised rotation transformation is achieved through a self-supervised rotation proxy task. In this task, the model automatically generates rotation labels and learns the variations in image features across different rotation angles through the self-rotation predictor  $Z_e$ . The self-rotation predictor  $Z_e$  receives low-level features from the feature extractor  $D_d$  as input and then outputs the probability distribution of all possible high-level feature rotations. Assuming that the network parameters of the feature extractor  $D_d$  and the self-rotation predictor  $Z_e$  are denoted by  $\phi_{Dd}$  and  $\phi_{Ze}$ ,

respectively, the training optimization objective for  $Z_e$  is defined as follows:

$$\text{MIN}_{\phi_{Dd}, \phi_{Ze}} \left( \frac{1}{V} \sum_{u=1}^{V_t} M_e(a_u^t, \phi_{Dd}, \phi_{Ze}) + \frac{1}{V_s} \sum_{u=1}^{V_s} M_e(a_u^s, \phi_{Dd}, \phi_{Ze}) \right) \quad (17)$$

Assuming the rotation label space is represented by  $E$ , the rotation prediction loss function  $M_b$  is defined as follows:

$$M_b = -\frac{1}{V_t} \sum_{a_u \in F_t} \sum_{e=0}^E O_{a_u \rightarrow e} \log Z_e(D_d(a_u)) - \frac{1}{V_s} \sum_{a_k \in F_s} \sum_{e=0}^E O_{a_k \rightarrow e} \log Z_e(D_s(a_k)) \quad (18)$$

### 3.5 Training process

During the training process, the ultimate goal of the model is to simultaneously minimize the loss functions of the various modules. Specifically, the total loss function of the model is composed of three parts: a) The loss  $M_b$  of the classifier  $Z_b$ , which is used to minimize the error in pedestrian identity classification. By optimizing  $L_y$ , the model can effectively recognize and distinguish different pedestrians. b) The loss  $M_e$  of the self-rotation predictor  $Z_e$ , which is aimed at minimizing the error in the self-supervised rotation prediction task. Optimizing  $L_r$  helps enhance the model's ability to capture image features. c) The loss  $M_{S-D}$  of the adversarial domain fusion discriminator  $Z_f$ , which is used to minimize the feature distribution shift between the source domain and the target domain. By optimizing  $L_{soft\_domain}$ , the model can learn more cross-domain consistent features. Assuming the number of image classes is denoted by  $B$ , and the probability that the model predicts  $a_u$  belongs to class  $b$  is denoted by  $O_{au \rightarrow b}$ , the loss  $M_b$  is defined as follows:

$$M_b = -\frac{1}{V_t} \sum_{a_u \in F} \sum_{b=1}^B O_{a_u \rightarrow b} \log Z_b(D_d(a_u)) \quad (19)$$

Assuming the balancing factor is denoted by  $\beta$ , the final total loss function, composed of  $M_b$ ,  $M_e$ , and  $M_{S-D}$ , is given by:

$$M = M_b + \beta M_e + M_{S-D} \quad (20)$$

In this study, pedestrian re-identification and pedestrian tracking are closely related tasks. In practical applications, the results of pedestrian re-identification can directly enhance the accuracy and robustness of pedestrian tracking, especially in multi-camera surveillance systems or extensive public safety scenarios. The primary task of pedestrian re-identification is to confirm and identify the same individual by extracting features and matching pedestrian images across different cameras. Therefore, pedestrian re-identification provides identity information across camera views, helping the tracking system maintain identity consistency as pedestrians move from one camera to another.

Pedestrian tracking based on re-identification results begins by capturing images of pedestrians using cameras deployed at multiple locations. The pedestrian re-identification model then



extracts features from these images. This model has undergone adaptive training under multi-domain conditions, enabling it to handle variations in images caused by different cameras, lighting conditions, and angles, thereby achieving accurate re-identification results. Once a pedestrian's identity is recognized, this identity information is transmitted to the pedestrian tracking module. The core task of pedestrian tracking is to continuously label and locate pedestrians in each frame of the video based on the identified identity information and position. Using the features and identity labels extracted by the re-identification model, the tracking module can establish pedestrian trajectories across consecutive frames. Even in cases of brief occlusion or movement, the system can continue tracking the target by leveraging prior re-identification information. For instance, in a surveillance system within a large shopping mall or airport, when a pedestrian exits the view of one camera, another camera may capture their image, at which point the re-identification results will assist the tracking module in re-locating and reconnecting the pedestrian's trajectory.

The integration of pedestrian re-identification and tracking offers significant advantages in practical applications. For example, in public security surveillance, the system can identify a target suspect through the pedestrian re-identification model and continuously track the suspect's movement using the tracking module, regardless of whether they move into the view of different cameras or experience occlusion. This cross-camera pedestrian tracking method greatly enhances the efficiency and accuracy of the surveillance system, mitigating the issues of tracking interruptions that often arise in traditional tracking methods due to changes in viewpoint or lighting conditions.

#### 4. EXPERIMENTAL RESULTS AND ANALYSIS

As shown in the experimental results in Table 1, there are significant performance differences under different domain conditions. When using only source domain training data, the F1 score was 4.851%, PPV (precision) was 14.265%, and TPR (recall) was 2.879%, indicating very poor performance. This result suggests that relying solely on source domain data results in insufficient generalization capability in the target domain. In contrast, when using only target domain training data, the F1 score significantly increased to 41.265%, with PPV reaching 74.268% and TPR at 28.152%. This demonstrates that, even in the absence of source domain data, the model can learn effectively within the target domain, though there remains room for further improvement. Finally, when combining source and target domain data, the F1 score was 37.589%, PPV was 60.125%, and TPR was 26.354%. Although mixed training performed slightly worse than using target domain data alone, its advantage lies in balancing the characteristics of both domains, offering a more robust solution. The results suggest that direct transfer of source domain data to the target domain is ineffective, highlighting significant domain differences and the challenge of model generalization in cross-domain scenarios. Therefore, training with only source domain data is insufficient to address the complexity of the target domain. While using only target domain data achieves better performance, it does not fully leverage the knowledge contained in the source domain, indicating that the scarcity of target domain data limits further model improvement. Although mixing source and target

domain data shows slightly lower evaluation metrics compared to using only target domain data, this strategy demonstrates better robustness and adaptability, making it suitable for pedestrian re-identification tasks in cross-domain scenarios.

**Table 1.** Experimental results of domain difference comparison

Training Dataset	Evaluation Metric (%)		
	F1	PPV	TPR
Source domain only	4.851	14.265	2.879
Target domain only	41.265	74.268	28.152
Mixed source and target domains	37.589	60.125	26.354

**Table 2.** Evaluation metrics for independent pedestrian feature detection results when using candidate image target domain for training

Evaluation Metric	Model	D1	D2	D3	D4	$\mu \pm \sigma$
F1 (%)	CycleGAN	16.785	18.524	18.254	17.214	117.265 $\pm$ 0.665
	DiscoGAN	37.124	35.231	38.265	40.215	38.254 $\pm$ 1.785
	Proposed model	45.225	51.298	45.124	46.265	47.256 $\pm$ 2.568
PPV (%)	CycleGAN	33.658	32.685	33.568	31.582	32.785 $\pm$ 0.889
	DiscoGAN	67.582	75.124	66.254	72.16	70.151 $\pm$ 3.854
	Proposed model	79.125	79.235	75.321	77.265	77.624 $\pm$ 1.856
TPR (%)	CycleGAN	11.125	12.897	12.258	12.125	12.235 $\pm$ 0.658
	DiscoGAN	25.362	23.568	27.235	28.254	25.362 $\pm$ 1.854
	Proposed model	31.598	33.325	32.236	33.265	32.154 $\pm$ 0.721

The data presented in Table 2 reveals significant differences in independent pedestrian feature detection performance across different detection points (D1–D4) among the three models. The CycleGAN model's F1 score ranges from 16.785% to 18.524%, with a PPV of 32.785%  $\pm$  0.889 and a TPR of 12.235%  $\pm$  0.658, indicating relatively poor detection performance in the target domain. In contrast, the DiscoGAN model shows improved performance across all detection points, with an F1 score ranging from 37.124% to 40.215%, a PPV of 70.151%  $\pm$  3.854, and a TPR of 25.362%  $\pm$  1.854, demonstrating stronger capability in capturing and recognizing features in the target domain. The proposed model outperforms the other models at all detection points, with F1 scores ranging from 45.225% to 51.298%, a PPV of 77.624%  $\pm$  1.856, and a TPR of 32.154%  $\pm$  0.721. These results are significantly higher than those of the comparative models, showcasing its robust capability in independent pedestrian feature detection. Based on the experimental analysis, the CycleGAN model struggles with detecting independent pedestrian features in the target domain, particularly in terms of recall, where it fails to effectively identify and capture key features. Although the DiscoGAN model shows considerable improvement in precision and recall, its performance remains limited, especially in scenarios with complex features, where the detection capabilities of the model are not fully utilized. In contrast, the proposed model significantly surpasses the other two models, excelling across all metrics, particularly precision and F1 score. This demonstrates that the integration of adversarial generative domain adaptation networks with self-supervised learning significantly enhances the model's adaptability and feature detection accuracy in the target domain.

**Table 3.** Experimental results combining complementary information from consecutive frames of pedestrian query images

Method	Evaluation Metric (%)				
	AUC	Accuracy	Recall	Precision	F1
Before combination	83.26	77.41	86.25	81.25	83.41
After combination	84.56	79.56	79.56	83.65	85.26

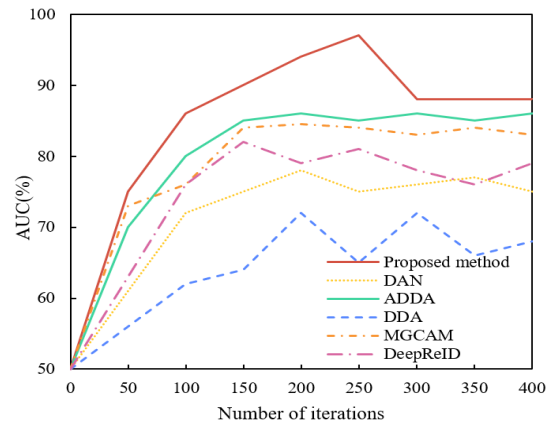
The data in Table 3 indicates that the integration of complementary information from consecutive frames of pedestrian query images leads to an improvement in all performance metrics of the model. Before the integration, the Area Under the Curve (AUC) was 83.26%, accuracy was 77.41%, recall was 86.25%, precision was 81.25%, and the F1 score was 83.41%. After the integration, the AUC increased to 84.56%, and accuracy rose to 79.56%. Although recall slightly decreased to 79.56%, precision improved to 83.65%, and the F1 score also increased to 85.26%. These results suggest that combining complementary information from consecutive frames contributes to an overall enhancement of performance, particularly in terms of AUC, accuracy, and F1 score. This enhancement demonstrates that consecutive frames provide additional temporal information, enabling the model to better capture the dynamic features of pedestrians, thereby improving its recognition capabilities in complex scenarios. Although there was a slight decrease in recall, the increase in precision and F1 score indicates that the model maintained high recognition accuracy while reducing false positives. This finding suggests that the proposed method not only effectively leverages the advantages of adversarial generative domain adaptation networks but also further enhances the model's adaptability in the target domain through deep self-supervised learning, ultimately improving overall pedestrian re-identification performance.

**Table 4.** Experimental results of different methods for pedestrian re-identification tasks

Method	Evaluation Metric (%)				
	AUC	Accuracy	Recall	Precision	F1
ResNet50	54.12	68.55	92.36	71.26	80.21
Vision Transformer	78.23	74.51	86.26	78.69	81.36
PCB	67.52	71.23	89.41	73.69	80.33
DeepReID	64.12	71.26	93.26	83.65	81.57
MGCAM	84.56	81.56	90.25	85.64	86.69
DDA	71.23	72.69	96.58	72.36	82.51
ADDA	86.25	84.26	92.64	83.69	88.26
DAN	81.23	81.26	91.26	83.26	87.15
Proposed method	92.36	85.46	94.58	86.45	90.23

The data in Table 4 reveals significant differences in performance across various methods used in pedestrian re-identification tasks. The ResNet50 model shows relatively weak performance, particularly in AUC and accuracy, with an AUC of 54.12%. The Vision Transformer demonstrates some improvement in AUC and accuracy, reaching 78.23% and 74.51%, respectively, although its performance in recall and precision is relatively average. Part-based Convolutional Baseline (PCB) and Deep Learning for Person Re-Identification (DeepReID) models perform well in terms of recall but still fall short in AUC and accuracy. The Mask-

Guided Contrastive Attention Model (MGCAM) and Deep Adaptation Network (DAN) methods exhibit more balanced performance overall, with MGCAM achieving an AUC of 84.56% and an F1 score of 86.69%, indicating strong recognition capability. The Adversarial Discriminative Domain Adaptation (ADDA) method attains an AUC of 86.25% and an F1 score of 88.26%, showing good performance in both precision and recall. The proposed method demonstrates the most outstanding performance, with an AUC of 92.36%, accuracy of 85.46%, and an F1 score of 90.23%, surpassing all other methods across these metrics, highlighting its strong advantages in pedestrian re-identification tasks. Based on the experimental analysis, the proposed method exhibits clear superiority in pedestrian re-identification tasks. Compared to other methods, this approach excels in key metrics such as AUC, accuracy, and F1 score, particularly with an AUC of 92.36% and an F1 score of 90.23%, demonstrating its exceptional recognition capabilities. This indicates that the integration of adversarial generative domain adaptation networks and deep self-supervised learning effectively enhances the model's adaptability and recognition accuracy in complex cross-domain scenarios. In contrast, although traditional methods such as ResNet50 and PCB perform well in recall, they struggle with other key metrics, proving insufficient to address the challenges of cross-domain pedestrian re-identification tasks.



**Figure 5.** AUC value trends for different methods in pedestrian re-identification tasks

The data presented in Figure 5 demonstrates a clear trend in the AUC values for different methods in pedestrian re-identification tasks as the number of iterations increases. The proposed method reached its peak AUC value of 97% at 250 iterations, followed by a slight decline, stabilizing around 88%. In comparison, the ADDA method achieved an AUC value of 86% at 200 iterations, remaining stable thereafter. The MGCAM method reached an AUC of 84.5% at 150 iterations, followed by minor fluctuations, ultimately stabilizing around 83%. The DAN and DeepReID methods exhibited relatively slow and modest growth in AUC throughout the iterations, stabilizing between 75% and 79%. The Domain Discrepancy Alignment (DDA) method showed the smallest increase in AUC, reaching only 68% after 400 iterations. The analysis of the AUC value trends for different methods reveals that the proposed method exhibited rapid improvement in the early stages of training, achieving a peak AUC of 97% at 250 iterations. This indicates that this method is capable of quickly learning and capturing effective features during the initial



training phase, demonstrating superior training efficiency. However, the slight decline in AUC after reaching its peak suggests that the model may have encountered some overfitting issues during the later stages of training. In contrast, although the ADDA method did not reach as high a peak as the proposed method, its AUC value remained relatively stable, indicating robust performance during model training. The MGCAM and DAN methods, while not achieving the same level of AUC growth as the proposed method and ADDA, still exhibited good stability and consistency. Overall, the proposed method demonstrated the best performance in pedestrian re-identification tasks, particularly in its rapid improvement during early training, showcasing its efficient feature-learning capability.

## 5. CONCLUSION

A pedestrian re-identification image keypoint detection method based on adversarial generative domain adaptation networks was proposed in this study, along with a pedestrian re-identification and tracking algorithm that integrates deep self-supervised adversarial domain adaptation networks. These innovative approaches significantly enhanced the accuracy of pedestrian re-identification in cross-domain scenarios. The experimental results demonstrate that the proposed method outperforms traditional methods across various metrics, particularly excelling in AUC, accuracy, and F1 score. Additionally, the experiments incorporating complementary information from consecutive frames of pedestrian query images further validate the model's effectiveness in practical applications, leading to notable improvements in detection precision and overall performance. The analysis of AUC value trends across multiple iterations reveals that the proposed method exhibits rapid learning capabilities in the early stages and maintains high recognition accuracy across several experiments.

This study provides a robust solution for pedestrian re-identification, particularly in complex cross-domain scenarios. By combining GANs with self-supervised learning, significant contributions have been made toward enhancing the model's adaptability and recognition accuracy. These achievements not only address challenges in real-world applications but also offer new directions for future research. However, certain limitations persist in this study, especially regarding model stability and overfitting issues. The experimental results indicate that, although the model performs well during the early stages of training, a decline in performance may occur later. Furthermore, this study primarily focuses on image-level detection and recognition, leaving the potential of leveraging video-level information from consecutive frames to be further explored. Future research could be expanded in several directions: firstly, improving model stability to reduce the risk of overfitting; secondly, further exploring multi-modal data fusion techniques to enhance robustness in pedestrian re-identification within complex environments; and thirdly, strengthening the extraction and utilization of video information to improve the model's recognition capabilities in dynamic scenarios.

## ACKNOWLEDGEMENT

This project was funded by the Shaanxi Provincial

Department of Science and Technology (Grant No.: 2023YBNY215/1512).

## REFERENCES

- [1] Miyamoto, A., Yabe, A., Hradil, P., Hakola, I. (2024). Feasibility study on intelligent bridge combined with smart monitoring techniques. *Structure and Infrastructure Engineering*, 20(7-8): 1133-1148. <https://doi.org/10.1080/15732479.2023.2276896>
- [2] Xue, Z., Yao, T. (2024). Enhancing occluded pedestrian re-identification with the MotionBlur data augmentation module. *Mechatronics and Intelligent Transportation Systems*, 3(2): 73-84. <https://doi.org/10.56578/mits030201>
- [3] Pedraza, C., Vega, F., Manana, G. (2018). PCIV, an RFID-based platform for intelligent vehicle monitoring. *IEEE Intelligent Transportation Systems Magazine*, 10(2): 28-35. <https://doi.org/10.1109/MITS.2018.2806641>
- [4] Kotapati, G., Ali, M.A., Vatambeti, R. (2023). Deep learning-enhanced hybrid fruit fly optimization for intelligent traffic control in smart urban communities. *Mechatronics and Intelligent Transportation Systems*, 2(2): 89-101. <https://doi.org/10.56578/mits020204>
- [5] Manzoor, S., An, Y.C., In, G.G., Zhang, Y., Kim, S., Kuc, T.Y. (2023). SPT: Single pedestrian tracking framework with re-identification-based learning using the Siamese Model. *Sensors*, 23(10): 4906. <https://doi.org/10.3390/s23104906>
- [6] Lei, M., Song, Y., Zhao, J., Wang, X., Lyu, J., Xu, J., Yan, W. (2022). End-to-end network for pedestrian detection, tracking and re-identification in real-time surveillance system. *Sensors*, 22(22): 8693. <https://doi.org/10.3390/s22228693>
- [7] Gwak, J., Park, G., Jeon, M. (2017). Viewpoint invariant person re-identification for global multi-object tracking with non-overlapping cameras. *KSII Transactions on Internet and Information Systems (TIIS)*, 11(4): 2075-2092. <https://doi.org/10.3837/tiis.2017.04.014>
- [8] Kumar, S.A., Yaghoubi, E., Das, A., Harish, B.S., Proença, H. (2020). The p-destre: A fully annotated dataset for pedestrian detection, tracking, and short/long-term re-identification from aerial devices. *IEEE Transactions on Information Forensics and Security*, 16: 1696-1708. <https://doi.org/10.1109/TIFS.2020.3040881>
- [9] Bouzid, A., Sierra-Sosa, D., Elmaghraby, A. (2022). Directional statistics-based deep metric learning for pedestrian tracking and re-identification. *Drones*, 6(11): 328. <https://doi.org/10.3390/drones6110328>
- [10] Adaimi, G., Kreiss, S., Alahi, A. (2021). Deep visual re-identification with confidence. *Transportation Research Part C: Emerging Technologies*, 126: 103067. <https://doi.org/10.1016/j.trc.2021.103067>
- [11] Nigam, J., Rameshan, R.M. (2022). Fast scale invariant tracker and re-identification for first-person social videos. *IETE Journal of Research*, 68(4): 2812-2825. <https://doi.org/10.1080/03772063.2020.1729258>
- [12] Wen, J.Z., Liu, H.Y., Li, J.B. (2024). PTDS centertrack: Pedestrian tracking in dense scenes with re-identification and feature enhancement. *Machine Vision and Applications*, 35(3): 54. <https://doi.org/10.1007/s00138-024-01520-8>

- [13] Bouzid, A., Sierra-Sosa, D., Elmaghraby, A. (2023). A robust pedestrian re-identification and out-of-distribution detection framework. *Drones*, 7(6): 352. <https://doi.org/10.3390/drones7060352>
- [14] Xiao, C., Luo, Z. (2023). Improving multiple pedestrian tracking in crowded scenes with hierarchical association. *Entropy*, 25(2): 380. <https://doi.org/10.3390/e25020380>
- [15] Hu, H., Hachiuma, R., Saito, H., Takatsume, Y., Kajita, H. (2022). Multi-camera multi-person tracking and re-identification in an operating room. *Journal of Imaging*, 8(8): 219. <https://doi.org/10.3390/jimaging8080219>
- [16] Ghiță, A.Ș., Florea, A.M. (2022). Real-time people re-identification and tracking for autonomous platforms using a trajectory prediction-based approach. *Sensors*, 22(15): 5856. <https://doi.org/10.3390/s22155856>
- [17] Huang, W., Luo, M., Zhang, P., Zha, Y. (2021). Full-scaled deep metric learning for pedestrian re-identification. *Multimedia Tools and Applications*, 80: 5945-5975. <https://doi.org/10.1007/s11042-020-09997-x>
- [18] Tagore, N.K., Chattopadhyay, P. (2022). A bi-network architecture for occlusion handling in Person re-identification. *Signal, Image and Video Processing*, 16(4): 1071-1079. <https://doi.org/10.1007/s11760-021-02056-4>
- [19] Qi, Y., Ge, H., Pei, W., Liu, Y., Hou, Y., Sun, L. (2023). Attention-guided spatial-temporal graph relation network for video-based person re-identification. *Neural Computing and Applications*, 35(19): 14227-14241. <https://doi.org/10.1007/s00521-023-08477-1>
- [20] Singh, N.K., Khare, M., Jethva, H.B. (2022). A comprehensive survey on person re-identification approaches: various aspects. *Multimedia Tools and Applications*, 81(11): 15747-15791. <https://doi.org/10.1007/s11042-022-12585-w>
- [21] Khan, S.U., Khan, N., Hussain, T., Baik, S.W. (2024). An intelligent correlation learning system for person Re-identification. *Engineering Applications of Artificial Intelligence*, 128: 107213. <https://doi.org/10.1016/j.engappai.2023.107213>
- [22] Liu, J., Lin, M., Zhao, M., Zhan, C., Li, B., Chui, J.K.T. (2023). Person re-identification via semi-supervised adaptive graph embedding. *Applied Intelligence*, 53(3): 2656-2672. <https://doi.org/10.1007/s10489-022-03570-9>