Consistency ordinary differential equations network for person re-identification

Yin Huang[a], **Jieyu Ding**[b,c,*]
[a]Qingdao University, School of Computer Science and Technology, Qingdao, China, 266071
[b]Qingdao University, School of Mathematics and Statistics, Qingdao, China, 266071
[c]Qingdao University, Center for Computational Mechanics and Engineering Simulation, Qingdao, China, 266071

Abstract. Person re-identification (Re-ID) aims to retrieve images of a given individual across different camera views. When images contain substantial identity-irrelevant information, the discriminability of person features degrades, leading to reduced recognition accuracy or even failure to recognize individuals. To address this issue, a novel consistency ordinary differential equations network (CODEN) is proposed, utilizing a dynamical system and incorporating a consistency regularization loss for Re-ID. Specifically, CODEN primarily consists of two key modules: the dynamic feature extraction (DFE) module and the stable trajectory evolution (STE) module. The DFE module parameterizes ordinary differential equations to model the dynamical system and dynamically extracts global saliency information by solving these equations using numerical integrators. The STE module perturbs the initial value of the dynamical system and employs a consistency regularization loss to learn robust feature representations, resisting interference from identity-irrelevant information. Experiments are conducted on three datasets, including Market-1501, DukeMTMC-reID, and CUHK03. The experimental results demonstrate the effectiveness of CODEN in enhancing person Re-ID performance.

Keywords: person re-identification, dynamical system, ordinary differential equation, consistency regularization loss.

*Jieyu Ding, djy@qdu.edu.cn

1 Introduction

Person re-identification (Re-ID) aims to match individual images captured from non-overlapping camera views.[1] Re-ID technology, to a certain extent, overcomes the inefficiency of manual retrieval and compensates for the visual limitations of current fixed cameras. It has gradually become one of the research hotspots in the fields of computer vision and pattern recognition due to its wide range of applications, including surveillance and criminal investigations.[2–4]

Existing person Re-ID methods based on deep learning can be roughly categorized into global-based methods, local-based methods, and cues-based methods. Global-based methods primarily rely on neural networks to extract features that capture the overall appearance of a pedestrian.[5,6] Local-based methods typically use grid segmentation, dividing the pedestrian image into several

blocks and extracting features from different regions for identification.[7,8] Cues-based methods utilize external visual cues, such as pose landmarks[9] and semantic parsing,[10] to detect key pedestrian regions. These works improve the performance of Re-ID to some extent. However, pedestrian images often include interference from identity-irrelevant factors, such as foreground occlusions and background clutter, which increases the difficulty of discriminative representation learning for Re-ID in realistic scenarios.

Adversarial defense strategies are techniques designed to protect neural networks from noise perturbations.[11–13] Some methods treat identity-irrelevant information in images as adversarial attacks and employ defensive strategies to protect Re-ID models from such perturbations. For instance, Zhao et al.[14] progressively generate adversarial examples and use an occlusion suppression strategy to enhance the representation of features in visible body regions. Similarly, Dong et al.[15] perturb feature maps to create multiple adversarial representations and utilize generative adversarial networks to learn robust pedestrian features. However, these methods primarily rely on random rectangular perturbation strategies to generate adversarial representations, often overlooking the relationship between the perturbed features and the discriminative information. For example, due to the irregular shape of pedestrians, randomly perturbing regions may disrupt key features such as clothing, arms, face, legs, and body shape, which are highly discriminative. This can introduce ambiguity in certain situations, ultimately impairing recognition performance.

In this paper, we propose a consistency ordinary differential equations network (CODEN), which leverages a dynamical system enhanced by consistency regularization loss for Re-ID. Specifically, CODEN models global information through a continuous dynamical system governed by ordinary differential equations (ODEs) to extract pedestrian features. In addition, CODEN obtains perturbed features from samples along the system trajectories to simulate the noisy information

introduced by non-target pedestrians. Furthermore, it employs a consistency regularization loss to prevent the network from focusing on undesirable information and to avoid confusing the learning dynamical system.

The main contributions of this paper can be summarized as follows:

- We propose CODEN, a dynamical system enhanced by consistency regularization loss for Re-ID, which mitigates the impact of identity-irrelevant information interference.
- We introduce a consistency regularization loss that enforces the dynamical system to filter out irrelevant factors and focus on the essential information needed to distinguish individuals.
- Extensive experiments on the Market-1501, DukeMTMC-reID, and CUHK03 datasets demonstrate the superiority of CODEN.

The remainder of this paper is organized as follows: The related works are introduced in Section 2; The details of the proposed method are elaborated in Section 3; The experimental comparison and analysis are reflected in Section 4; The conclusion of the present paper is summarized in Section 5.

2 Related work

2.1 Person re-identification

According to the feature learning technologies, person re-ID methods can be roughly categorized into hand-crafted feature-based methods and deep learning-based methods. Hand-crafted feature-based methods focus on manually designing features, which includes both feature extraction and metric learning. The feature extraction aims to obtain discriminative visual features, such

as SIFT,[16] HOG,[17] and LBP.[18] Metric learning aims to measure the similarity among features in the feature space, such as LFDA,[19] LMNN,[20] and DR-KISS.[21] However, such methods require careful design of features and have limited performance when dealing with large-scale databases.

Deep learning-based methods primarily employ data-driven approaches to extract pedestrian features. According to the way of extracting features, deep learning-based methods can be roughly categorized into global-based methods, local-based methods and cues-based methods. Global-based methods extract a compact global embedding for each person image. For example, Luo et al.[5] introduced a strong baseline by employing a range of effective training strategies. Chang et al.[22] factorised the visual appearance of a person into latent discriminative factors at multiple semantic levels. Li et al.[23] performed joint learning of soft pixel attention and hard regional attention. Ghorbel et al.[6] employed two branches to separately learn pixel-level local regions and additional missed features. Local-based methods apply automatic or manual slicing of feature maps to mine the spatial information. For example, Sun et al.[7] horizontally partitioned the feature map to extract features. Zhang et al.[8] simultaneously learned diverse body features and discriminative part features. Tian et al.[24] employed global branches to supervise local branches, enhancing the distinctiveness of feature representations. Yang et al.[25] leveraged local correlations to aggregate distinctive information for local features.

Cues-based methods focus on external tools to guide feature extraction. Song et al.[26] employed masks of pedestrian regions to separately learn features from the body and background. Miao et al.[9] utilized pose landmarks to align extracted features between pedestrian images. Guo et al.[10] applied a human parsing model to refine the representation. Tang et al.[27] introduced a gradual background suppression strategy to reduce background clutter and mitigate its impact. In contrast, our method leverages the potential advantages of a dynamical system to learn discriminative global

features without requiring external visual cue tools.

2.2 Adversarial attack and defense

The process of adversarial attack and defense is mutually antagonistic yet complementary. Advanced attack techniques drive the development of targeted defense mechanisms, while robust defense strategies, in turn, spur the advancement of attack methodologies. The primary objective of adversarial attacks is to induce misclassification in models by introducing perturbations to input samples. Szegedy[11] first introduced the concept of adversarial examples, demonstrating that imperceptible perturbations can lead to incorrect classifications by neural networks. Person Re-ID systems are vulnerable to adversarial attacks. For instance, Bai et al.[28] investigated the adversarial effects in Re-ID and proposed the adversarial metric attack. Wang et al.[29] generated adversarial patterns on the clothing of pedestrians. Wang et al.[12] organized features of different levels in a pyramid structure to extract general and transferable features for the adversarial perturbations.

The primary goal of adversarial defense methods is to protect models from adversarial attacks, with a focus on improving the robustness of neural networks. Some works have applied adversarial defense techniques to enhance the robustness of identity recognition systems. For example, Zhao et al.[14] treated occlusion as a form of noise perturbation and utilized a suppression mechanism to enhance attention to features from non-occluded body regions. Wang et al.[30] detected adversarial attacks by examining context inconsistency. Wang et al.[13] applied a random patch method to synthesize adversarial examples from two images, mitigating the impact of occlusions on pedestrian identity matching. Dong et al.[15] generated adversarial samples at the feature level to simulate challenges such as information loss, misalignment, and noisy data. In contrast, our method obtains diverse perturbed features by sampling along the system trajectory, thereby alleviating the ambi-

guity caused by random rectangular perturbation strategies. In addition, a effective regularization technique is introduced to enable the dynamical system to filter out irrelevant factors for Re-ID.

3 Proposed method

We propose a novel method, CODEN, for Re-ID that leverages a dynamical system enhanced by consistency regularization loss to mitigate the impact of interference from identity-irrelevant information. The proposed method primarily comprises three main components: the dynamic feature extraction (DFE) module, the stable trajectory evolution (STE) module, and a backbone. The DFE module is designed as a dynamical system that extracts global features $\boldsymbol{q}(t_T)$ through temporal evolution. According to the intermediate features $\boldsymbol{q}(t_0 + t_\epsilon)$ and features $\boldsymbol{q}(t_T)$ in system trajectories, the STE module employs consistency regularization loss to enhance the focus on the essential information needed to differentiate individuals. The backbone utilizes ResNet50,[31] pre-trained on ImageNet,[32] as a feature extractor. It extracts feature maps $\boldsymbol{Q} \in \mathbb{R}^{C\times W\times H}$, where C, W, and H represent the number of channels, width, and height, respectively. Global average pooling and batch normalization are then applied to obtain the features $\boldsymbol{q}_0 \in \mathbb{R}^C$. After feature extraction, the identity loss, triplet loss, center loss, and consistency regularization loss are used to ensure the model learns discriminative features. A schematic diagram illustrating the CODEN is presented in Fig. 1.

3.1 Dynamic feature extraction module

Due to factors such as camera viewpoints and human poses, identity-relevant information is often limited, while pedestrian images may include interference from identity-irrelevant information. For such cases, relying solely on the basic ResNet50[31] for learning features results in suboptimal

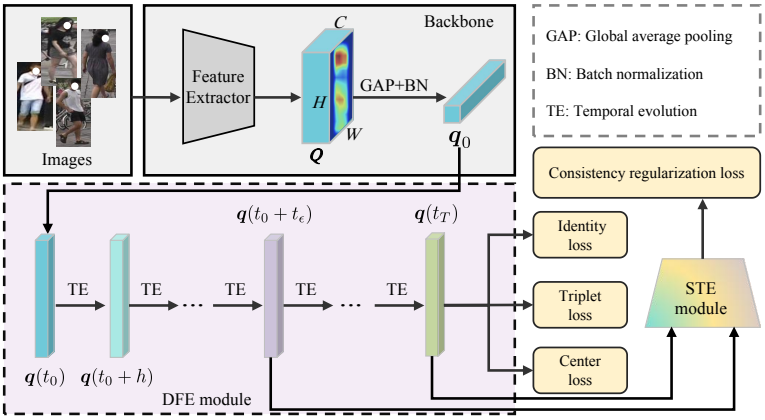

Fig 1 Schematic illustration of CODEN. It primarily consists of the backbone, the DFE module, and the STE module. In the backbone, we utilize a feature extractor to extract the person features. In the DFE module, we consider the features extracted by the backbone as the initial value in the dynamical system and further extract the global features. Based on the global features, the identity loss, triplet loss, and center loss are calculated. In the STE module, we input the global features and intermediate features to compute the consistency regularization loss.

discriminative performance and is prone to introducing interference factors. The non-intersecting property refers to integral curves originating from a point and is influenced by neighboring curves. Dynamical systems controlled by ODEs possess this non-intersecting property, making them more robust against adversarial attacks compared to convolutional neural networks.[33,34] Consequently, we introduce an ODE-controlled dynamical system into the person Re-ID task and design a DFE module to mitigate the impact of interfering factors and extract pedestrian saliency information. The core motivation of the DFE module is that features of pedestrians with the same identity are close to each other in the representation space. This module constrains the features along the integral curves of neighboring pedestrians with the same identity, thereby promoting identity feature learning.

The DFE module primarily consists of an ODE block and an integrator. The ODE block is designed to establish a parameterized ODEs that describes the complex temporal evolution of pedestrian features. The input to the ODE block is the features $\boldsymbol{q}(t)$, and the output is the evolution function $\mathrm{d}\boldsymbol{q}(t)/\mathrm{d}t$ over time. The integrator dynamically solves these equations, extracting

discriminative pedestrian features. It outputs the features at the next time step, $\boldsymbol{q}(t+h)$, where h represents the step size. A schematic diagram of the DFE module architecture is shown in Fig. 2.

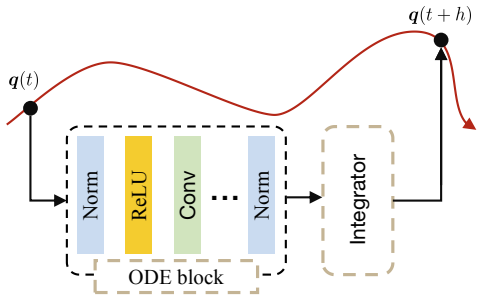

Fig 2 Schematic illustration of DFE module. It primarily consists of an ODE block and an integrator. In the ODE block, we input the features from the previous time step and calculate the temporal evolution function of the dynamical system. In the integrator, we input the temporal evolution function to obtain the features for the next time step.

In the ODE block, we formulate the ODE to describe the evolution of the features, as given by Eq. (1):

$$\begin{cases} \frac{\mathrm{d}\boldsymbol{q}(t)}{\mathrm{d}t} = f_{\boldsymbol{\theta}}(t, \boldsymbol{q}(t)), \quad t \in [t_0, t_T] \\ \boldsymbol{q}(t_0) = \boldsymbol{q}_0 \end{cases}, \tag{1}$$

where t denotes the evolution time over the interval $[t_0, t_T]$, $\boldsymbol{q} \in \mathbb{R}^d$ represents the d-dimensional state of the differential equation, and $f_{\boldsymbol{\theta}}(\cdot)$ is a trainable function parameterized by weights $\boldsymbol{\theta}$, consisting primarily of 1×1 convolutional layers, group normalization layers, and ReLU activation functions.

In the integrator, we employ the fourth-order Runge-Kutta (RK4) scheme to compute pedestrian features, as described by Eq. (2):

$$\boldsymbol{q}_{n+1} = \boldsymbol{q}_n + \frac{h}{6}\left(\boldsymbol{k}_1 + 2\boldsymbol{k}_2 + 2\boldsymbol{k}_3 + \boldsymbol{k}_4\right), \tag{2}$$

where

$$
\begin{aligned}
\boldsymbol{k}_1 &= f_{\boldsymbol{\theta}}\left(t_n, \boldsymbol{q}_n\right), \\
\boldsymbol{k}_2 &= f_{\boldsymbol{\theta}}\left(t_n + \frac{h}{2}, \boldsymbol{q}_n + \frac{h}{2}\boldsymbol{k}_1\right), \\
\boldsymbol{k}_3 &= f_{\boldsymbol{\theta}}\left(t_n + \frac{h}{2}, \boldsymbol{q}_n + \frac{h}{2}\boldsymbol{k}_2\right), \\
\boldsymbol{k}_4 &= f_{\boldsymbol{\theta}}\left(t_n + h, \boldsymbol{q}_n + h\boldsymbol{k}_3\right),
\end{aligned}
\tag{3}
$$

where h is the step size.

3.2 Stable trajectory evolution module

The shapes of pedestrians are often irregular, which makes random rectangular perturbation strategies likely to introduce ambiguities, potentially diminishing the benefits of adversarial representations. Therefore, the STE module utilizes features sampled along the system trajectories as perturbations to the initial value of the dynamical system, simulating the noisy information introduced by non-target pedestrians. It then employs a consistency regularization loss to control differences between the perturbed features and their non-perturbed counterparts, reducing the impact of irrelevant factors on Re-ID performance.

Considering an intermediate time $t_0 + t_\epsilon \in (t_0, t_T)$, the corresponding features $\boldsymbol{q}(t_0 + t_\epsilon)$ can be expressed as:

$$
\boldsymbol{q}(t_0 + t_\epsilon) = \boldsymbol{q}(t_0) + \int_{t_0}^{t_0 + t_\epsilon} f_{\boldsymbol{\theta}}(t, \boldsymbol{q}(t))\mathrm{d}t. \tag{4}
$$

The features $\boldsymbol{q}(t_0 + t_\epsilon)$ can be regarded as the result of perturbations applied to $\boldsymbol{q}(t_0)$. If $\boldsymbol{q}(t_0 + t_\epsilon)$ is taken as the initial condition in Eq. (1), the solution is $\boldsymbol{q}(t_T + t_\epsilon)$. The difference between the

perturbed features $\boldsymbol{q}(t_T + t_\epsilon)$ and the non-perturbed features $\boldsymbol{q}(t_T)$ can be expressed as:

$$\|\boldsymbol{q}(t_T + t_\epsilon) - \boldsymbol{q}(t_T)\|_2 = \left\| \int_{t_T}^{t_T + t_\epsilon} f_{\boldsymbol{\theta}}(t, \boldsymbol{q}(t)) \mathrm{d}t \right\|_2 . \tag{5}$$

Applying the Gronwall-Bellman inequality, Eq. (5) can be reformulated as:

$$\|\boldsymbol{q}(t_T + t_\epsilon) - \boldsymbol{q}(t_T)\|_2 \leq \left\| S e^{t_\epsilon} e^{\int_{t_T}^{t_T + t_\epsilon} |f_{\boldsymbol{\theta}}(t, \boldsymbol{q}(t))| \mathrm{d}t} \right\|_2 , \tag{6}$$

where $S > 0$ represents a hyper-parameter used to adjust the difference between $\boldsymbol{q}(t_T + t_\epsilon)$ and $\boldsymbol{q}(t_T)$. To minimize the difference, the consistency regularization loss is designed, as shown in Eq. (7):

$$\mathcal{L}_{\mathrm{s}} = \sum_{i=1}^{N} \left\| S e^{t_\epsilon} e^{\int_{t_T}^{t_T + t_\epsilon} |f_{\boldsymbol{\theta}}(t, \boldsymbol{q}_i(t))| \mathrm{d}t} \right\|_2 , \tag{7}$$

where N represents the number of images. When $\mathcal{L}_{\mathrm{s}}$ is sufficiently small, the model's outputs for each pedestrian with the same identity will stabilize around $\boldsymbol{q}(t_T)$.

3.3 Optimization

During the training phase, the identity loss $\mathcal{L}_{\mathrm{id}}$, triplet loss $\mathcal{L}_{\mathrm{tri}}$, center loss $\mathcal{L}_{\mathrm{cen}}$, and consistency regularization loss $\mathcal{L}_{\mathrm{con}}$ are used to optimize the network parameters. We adopt cross entropy loss as the identity loss to supervise feature learning, with label smoothing[35] employed to prevent overfitting. The identity loss $\mathcal{L}_{\mathrm{id}}$ can be expressed as:

$$\mathcal{L}_{\mathrm{id}} = -\sum_{k=1}^{N} \hat{y}_k \log s(\hat{p}_k), \tag{8}$$

where N denotes the number of images. The function $s(\cdot)$ signifies the softmax operation. $\hat{p}$ represents the predicted values generated by an identity classifier. The smoothed labels, denoted as $\hat{y}$, are computed with the k-th element calculated as follows:

$$\hat{y}_k = (1-\epsilon)y + \frac{\epsilon_y}{N}, \tag{9}$$

where $0 < \epsilon_y < 1$ is the smoothing rate, typically set to 0.1, and y denotes the sample labels.

We employ triplet loss with hard sample mining,[5] which selects the farthest positive sample and the closest negative sample within a training batch to separate positive and negative samples by a certain margin. P identities and K images for each identity are randomly selected to form a batch of size $P \times K$. The triplet loss $\mathcal{L}_{\text{tri}}$ is defined as follows:

$$\mathcal{L}_{\text{tri}} = \sum_{i=1}^{PK} \left[\max_{p \in \mathcal{P}} d_i^{a,p} - \min_{n \in \mathcal{N}} d_i^{a,n} + m \right]_+, \tag{10}$$

where m represents the margin between positive and negative sample pairs. $[x]_+ = \max(x, 0)$. a is the index of the anchor image in the batch. p and n are the indices of the most challenging positive sample and the most challenging negative sample of index a respectively. $\mathcal{P}$ and $\mathcal{N}$ represent the positive sample set and negative sample set of index a respectively, that is, $\mathcal{P} = \{p | y_p = y_a\}$ and $\mathcal{N} = \{n | y_n \neq y_a\}$. d denotes the Euclidean distance between feature representations.

We adopt center loss,[36] which penalizes the distance between features and their corresponding class centers to encourage instances of the same class to be closer to their class center. Specifically,

the loss $\mathcal{L}_{\text{cen}}$ is defined as follows:

$$\mathcal{L}_{\text{cen}} = \frac{1}{2}\sum_{j=1}^{N}\left\|\boldsymbol{q}_j - \boldsymbol{c}_{y_j}\right\|_2^2, \tag{11}$$

where y_j denotes the label of the j-th image, and $\boldsymbol{c}_{y_j}$ represents the learnable center of class y_j in the feature space.

Finally, the total loss $\mathcal{L}_{total}$ of our method during training is as follows:

$$\mathcal{L}_{total} = \lambda_{\text{id}}\mathcal{L}_{\text{id}} + \lambda_{\text{tri}}\mathcal{L}_{\text{tri}} + \lambda_{\text{cen}}\mathcal{L}_{\text{cen}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}}, \tag{12}$$

where λ_{id}, λ_{tri}, λ_{cen}, and λ_{con} represent the weights assigned to the corresponding losses.

4 Experiment

In this section, we first detail the datasets used and the experimental settings. Next, we conduct an ablation study to examine the effectiveness of each module. Then, we evaluate the impact of hyper-parameters on the CODEN. Subsequently, the proposed method is compared with several existing methods. Finally, we provide a qualitative analysis of the experimental results.

4.1 Datasets and experimental settings

Datasets. We evaluate our method on three widely used Re-ID datasets, including Market-1501,[37] DukeMTMC-reID,[38] and CUHK03.[39] The details of the datasets are provided in Table 1.

Market-1501 dataset contains 32,668 images of 1,501 identities. All images are collected by 5 high-resolution and 1 low-resolution cameras deployed in the Tsinghua University campus. The

train set consists of 12,936 images from 751 identities. The test set consists of 19,732 images from 750 identities.

DukeMTMC-reID dataset contains 36,411 images of 1,404 identities from 8 cameras at Duke University. The train set consists of 16,522 images from 702 identities, and the test set consists of 19,889 images from 702 identities.

CUHK03 dataset contains images from 1,467 identities captured by 6 cameras at the Chinese University of Hong Kong. It is annotated into labeled and detected datasets. CUHK03-labeled dataset consists of 7,368 train images and 5,328 test images, while CUHK03-detected dataset consists of 7,365 train images and 5,332 test images. We partitioned the dataset according to the new protocol,[39] with 767 identities in the train set and the remaining 700 identities in the test set.

Table 1 The detailed information of datasets. ID represents the number of identities. IMG represents the number of images.

		Train		Test	
Dataset	Camera	ID	IMG	ID	IMG
Market-1501	6	751	12,936	750	19,732
DukeMTMC-reID	8	702	16,522	702	17,661
CUHK03	6	767	7,368/7,365	700	5,328/5,332

Evaluation metrics. Performance evaluation is conducted utilizing the cumulative matching curve (CMC) and mean average precision (mAP) metrics. The CMC protocol is used to evaluate the model's performance at Rank-k, where Rank-k represents the expectation of finding the correct target at the k-th rank. The mAP protocol is used to evaluate the performance of multiple matching images in the gallery set. These metrics are jointly utilized for a comprehensive evaluation of the overall performance.

Implementation details. The experiments are conducted on a machine equipped with an NVIDIA 3060 GPU, using the PyTorch framework. During the data preprocessing phase, all

images are resized to 256×128 pixels. Each batch comprises 60 images, with 4 identities per batch, and each identity includes 15 images. The dimension of the pedestrian features is set to 2,048. The weight hyper-parameters for the model are configured as follows: $\lambda_{\text{id}} = 1$, $\lambda_{\text{tri}} = 1$, $\lambda_{\text{cen}} = 5 \times 10^{-4}$, and $\lambda_{\text{con}} = 0.35$. The stable hyper-parameter is set to $S = 0.15$. For the numerical integration of the ODEs, the initial time is set to $t_0 = 0$, the final time to $t_T = 0.04$, and the intermediate time to $t_\epsilon = 0.01$, the step size to $h = 0.01$. The training process employs the Adam optimizer for model parameter updates over 180 epochs. The initial learning rate is set to 3.5×10^{-4} and decays by a factor of 0.1 after training for 40 and 70 epochs. The re-ranking technique is not employed for fair comparison.

4.2 Ablation studies

To evaluate the contribution of each module in the proposed CODEN, a series of ablation experiments were conducted by progressively adding or removing modules. ResNet50[31] was used as the baseline model. The results of these experiments are presented in Table 2.

Table 2 Ablation studies of the CODEN on the Market-1501 and DukeMTMC-reID datasets.

	Market-1501			DukeMTMC-reID		
Method	mAP	R-1	R-5	mAP	R-1	R-5
Baseline	85.9	94.4	98.1	76.5	87.0	94.2
ODEN	87.4	95.2	98.4	78.9	88.9	95.0
CODEN	**88.0**	**95.4**	**98.6**	**79.1**	**89.6**	**95.2**

Effectiveness of the DFE module. The effectiveness of the DFE module was investigated. ODEN integrates the DFE module into the baseline. Compared to the baseline, ODEN exhibits an improvement in mAP by 2.1% and in Rank-1 by 1.0% on the Market-1501, and an improvement in mAP by 2.6% and in Rank-1 by 2.4% on the DukeMTMC-reID. These results demonstrate that the

DFE module significantly enhances performance by combining feature learning with a dynamical system.

Effectiveness of the STE module. The effectiveness of the STE module was investigated. The CODEN integrates both the DFE and STE modules. Compared to ODEN, CODEN shows improvements in mAP from 87.4% to 88.0% and in Rank-1 from 95.2% to 95.4% on Market-1501, and improvements in mAP from 78.9% to 79.1% and in Rank-1 from 88.9% to 89.6% on DukeMTMC-reID. The STE module reduces identity-irrelevant information interference by introducing feature perturbations and employing consistency regularization loss, which helps form a more discriminative feature space and enhances the robustness of recognition tasks.

Impact of the ODE block structrue. The impact of the ODE block structure with different numbers of convolutional layers in CODEN on performance was investigated, with results presented in Table 3. The optimal performance of CODEN is achieved with two convolutional layers. Deviations from this number, whether by increasing or decreasing the layers, result in a slight decrease in accuracy. This suggests that overly simplistic dynamic models may not capture essential pedestrian features, while excessively complex models could introduce unnecessary complications in network weight optimization. Therefore, maintaining a moderate number of layers in dynamic models is crucial for ensuring optimal performance.

Table 3 Impact of ODE block structure for CODEN on the Market-1501 and DukeMTMC-reID datasets.

	Market-1501			DukeMTMC-reID		
Number of layers	mAP	R-1	R-5	mAP	R-1	R-5
1	**88.1**	95.0	98.5	79.0	89.1	95.0
2	88.0	**95.4**	**98.6**	**79.1**	**89.6**	95.2
3	87.7	95.3	98.4	79.0	88.8	**95.3**
4	87.6	95.3	98.3	78.9	88.8	95.0

Impact of the integral scheme. The impact of the integral scheme has been investigated.

We compared the impact of the Euler scheme, the Midpoint scheme, and the RK4 scheme on the Market-1501 dataset. The experimental results are presented in Table 4. Compared to the Euler scheme, the RK4 scheme improves mAP and Rank-1 accuracy by 0.4% and 0.3%, respectively. When compared to the Midpoint scheme, it enhances mAP and Rank-1 accuracy by 0.3% and 0.1%, respectively. These results indicate that while the RK4 scheme is more computationally intensive, it offers more accurate and stable recognition. Therefore, the appropriate integral scheme can be selected based on the specific requirements for accuracy and efficiency in different scenarios.

Table 4 Impact of different integral schemes for CODEN on the Market-1501 dataset. Error represents the accumulated error of the scheme. t_{train} represents the training time for each iteration.

Scheme	Error	t_{train}	mAP	R-1	R-5	R-10
Euler	$\mathcal{O}(h)$	0.35 s/it	87.6	95.0	98.5	99.1
Midpoint	$\mathcal{O}(h^2)$	0.36 s/it	87.7	95.3	98.5	99.1
RK4	$\mathcal{O}(h^4)$	0.39 s/it	88.0	95.4	98.6	99.1

Cross-dataset evaluation. The generalization ability of a method is a crucial factor in its suitability for practical deployment. To evaluate the CODEN's generalization capability, we conducted a cross-dataset evaluation for recognizing unseen pedestrians and scenes. Models trained on one dataset are evaluated on a different dataset. The experimental results are presented in Table 5. Compared to the baseline, CODEN achieves improvements of 8.9% in mAP and 9.5% in Rank-1 from Market-1501 to DukeMTMC-reID, and 6.0% in mAP and 7.4% in Rank-1 from DukeMTMC-reID to Market-1501. This indicates that the design of the DFE and STE modules is effective, and CODEN enhances the ability to effectively recognize unseen pedestrians as well as scenes.

Table 5 Cross-domain evaluation on Market-1501 and DukeMTMC-reID datasets. Market → DukeMTMC represents that the model is trained on source domain Market-1501 and tested on target domain DukeMTMC-reID, and vice versa.

	Market → Duke			Duke → Market		
Method	mAP	R-1	R-5	mAP	R-1	R-5
Baseline	17.9	32.2	47.2	21.5	47.9	63.2
ODEN	25.7	40.9	56.6	26.3	53.3	69.2
CODEN	**26.8**	**41.7**	**58.0**	**27.5**	**55.3**	**71.4**

4.3 Parameter analysis

The CODEN introduces two crucial training hyper-parameters: λ_{con} and S. In order to investigate the impact of these parameters on performance, numerous experiments were conducted on the Market-1501 dataset, and the results are illustrated in Fig. 3.

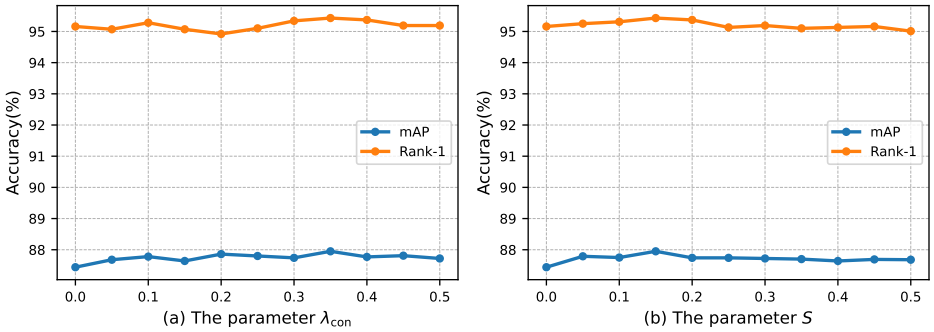

Fig 3 Analysis of parameters: (a) the impact of the parameter λ_{con} and (b) the impact of the parameter S

Impact of parameter λ_{con}. The parameter λ_{con} affects the strength of supervision on global feature stability. We set λ_{con} within the range of 0 to 0.5, and the results are illustrated in Fig. 3(a). As the value of λ_{con} increases, both mAP and rank-1 show an upward trend. CODEN achieves peak performance when $\lambda_{\text{con}} = 0.35$. However, with further increases in λ_{con}, performance declines, possibly due to an excessive emphasis on stability, which compromises the model's discriminative properties. These results indicate that both insufficient and excessive supervision of global feature stability lead to a decline in performance.

Impact of parameter S. The parameter S is a factor influencing the difference between integral curves. We set S within the range of 0 to 0.5, and the results are depicted in Fig. 3(b). When the value of the parameter S is small, the dynamic trajectories propagated forward by the ODEs may rapidly change near the termination time. This behavior increases the rigidity of the dynamical system and reduces recognition performance. Conversely, when the value of the parameter S is large, pedestrian features may aggregate irrelevant information, making it difficult to learn distinctive information. At $S = 0.15$, both the mAP and Rank-1 reach their peaks. The results further validate the effectiveness of the consistency regularization loss.

4.4 Performance comparison

To ensure a fair evaluation of CODEN's performance, we compared it against several existing methods on the Market-1501, DukeMTMC-reID, and CUHK03 datasets. The comparison results are presented in Table 6. CODEN shows competitive performance in these benchmark datasets.

On the Market-1501 dataset, CODEN achieves an impressive mAP of 88.0% and Rank-1 of 95.4%. Compared to the defense-based method ETNDNet,[15] CODEN shows improvements of 0.8% in mAP and 0.1% in Rank-1. Compared to the local-based method PCB+RPP,[7] CODEN shows improvements of 6.4% in mAP and 1.6% in Rank-1. Additionally, CODEN demonstrates competitive advantages compared over other methods.

On the DukeMTMC-reID dataset, the proposed CODEN achieves a mAP of 79.1% and Rank-1 of 89.6%. Compared to the global-based method BOT,[5] CODEN exhibits improvements of 2.7% and 3.2% in mAP and Rank-1, respectively. Compared to the cues-based method PGFA,[9] CODEN shows improvements of 6.4% in mAP and 3.9% in Rank-1. CODEN outperforms other methods, achieving better performance on both evaluation metrics.

Table 6 Comparing with the state-of-the-art methods on the Market-1501, DukeMTMC-reID, and CUHK03 datasets.

	Market-1501		DukeMTMC-reID		CUHK03-detected		CUHK03-labeled	
Method	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1
PCB[7]	77.4	92.3	65.3	81.9	-	-	54.2	61.3
PCB+RPP[7]	81.6	93.8	69.2	83.3	-	-	57.5	63.7
HA-CNN[23]	75.7	91.2	63.8	80.5	41.0	44.4	38.6	41.7
MLFN[22]	74.3	90.0	62.8	81.0	49.2	54.7	47.8	52.8
MGCAM[26]	74.3	83.6	-	-	49.9	49.3	46.8	46.3
BOT[5]	85.9	94.5	76.4	86.4	62.7	65.6	65.0	66.5
PGR[40]	77.2	93.9	66.0	82.6	-	-	-	-
P^2-Net[10]	83.4	94.0	70.8	84.9	64.2	71.6	69.2	75.8
PGFA[9]	81.4	94.6	72.7	85.7	-	-	-	-
FPO[27]	79.2	91.8	-	-	56.3	63.1	60.2	65.6
IGOAS[14]	84.1	93.4	75.1	86.9	-	-	-	-
AFELN[8]	81.3	93.3	72.9	86.5	-	-	61.7	66.5
AND[6]	87.8	92.3	63.7	83.8	56.5	60.6	55.6	61.9
ICAMFL[41]	82.3	93.3	71.6	85.6	59.3	64.6	63.3	67.1
PRE-Net[25]	86.5	95.3	77.8	89.3	-	-	-	-
SRFnet[24]	85.7	94.2	77.9	89.1	69.6	73.3	72.4	75.0
ETNDNet[15]	87.2	95.3	77.9	88.5	-	-	-	-
CODEN	**88.0**	**95.4**	**79.1**	**89.6**	**70.2**	**73.6**	**73.0**	**76.5**

On the CUHK03 dataset, the proposed CODEN achieves a mAP of 70.2% and Rank-1 of 73.6% on the automatically detected data, and a mAP of 73.0% and Rank-1 of 76.5% on the manually labeled data. Compared to the global-based method ICAMFL,[41] CODEN shows improvements of 10.9% in mAP and 9.0% in Rank-1 on the detected data, and 9.7% in mAP and 9.4% in Rank-1 on the labeled data. Compared with other methods, CODEN shows significant performance advantages. These results demonstrate the effectiveness of CODEN.

4.5 Qualitative analysis

To qualitatively validate the effectiveness of CODEN, we conducted a visualization analysis of saliency. Saliency maps are generated by extracting the maximum values along the channel dimension of features, enabling an intuitive understanding of the focus on different regions. Red

regions exhibit higher attention, while blue regions display lower attention. Fig. 4 presents the visualization results, clearly illustrating the superiority of CODEN over the baseline. The baseline shows localized hotspots, which are susceptible to interference and can potentially lead to erroneous identification results. CODEN exhibits multiple highly activated hotspots. This enables a more precise perception of pedestrian details and the extraction of as many relationships and regional features as possible. It showcases its outstanding performance in feature representation and recognition. This further validates the effectiveness and superiority of CODEN in the task of person Re-ID.

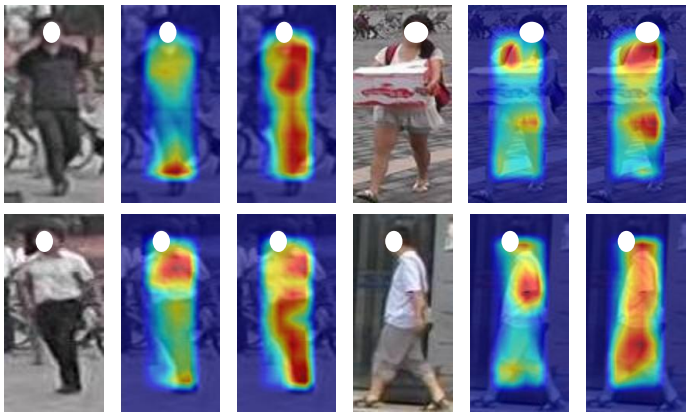

Fig 4 Visualization of regions of interest by the method. (a) Original image. (b) Heatmap generated by the baseline. (c) Heatmap generated by CODEN.

5 Conclusion

Person Re-ID is an important research field with huge application potential and challenges. In this paper, we present the CODEN for Re-ID, which establishes a nonlinear dynamic model using ODEs to better mine pedestrian information. By introducing feature perturbation and incorporating a consistency regularization loss, CODEN effectively regularizes system trajectories, thereby reducing the influence of identity-irrelevant information. Despite the proposed method has achieved

progress on the two evaluation metrics, there is still room for further improvement. For example, the impact of different forms of dynamic models on recognition performance deserves further study.

Disclosures

The authors declare no conflict of interest.

Code, Data, and Materials Availability

The codes and datasets generated and/or analyzed during the current study are available from the corresponding author on reasonable request.

Acknowledgments

This study has been supported by the National Natural Science Foundations of China under Grant Nos. 12172186, 11772166.

References

1 A. Nambiar, A. Bernardino, and J. C. Nascimento, "Gait-based person re-identification: A survey," *ACM Comput. Surv.* **52**(2), 1–34 (2019).

2 D. Zhang, H. Fan, X. Zhou, *et al.*, "Joint global feature and part-based pyramid features for unsupervised person re-identification," *J. Electron. Imaging* **33**(2), 023043–023043 (2024).

3 X. Wang, Y. Zhang, Y. Xu, *et al.*, "Multi-feature fusion network for person reidentification," *J. Electron. Imaging* **32**(2), 023044–023044 (2023).

4 C. Hu, Y. Chen, L. Guo, *et al.*, "Pose-guided node and trajectory construction transformer for occluded person re-identification," *J. Electron. Imaging* **33**(4), 043021–043021 (2024).

5 H. Luo, Y. Gu, X. Liao, *et al.*, "Bag of tricks and a strong baseline for deep person re-identification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 1487–1495 (2019).

6 M. Ghorbel, S. Ammar, Y. Kessentini, *et al.*, "Masking for better discovery: Weakly supervised complementary body regions mining for person re-identification," *Expert Syst. Appl.* **197**(1), 116636 (2022).

7 Y. Sun, L. Zheng, Y. Yang, *et al.*, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proceedings of the European conference on computer vision*, 480–496 (2018).

8 W. Zhang, L. Huang, Z. Wei, *et al.*, "Appearance feature enhancement for person re-identification," *Expert Syst. Appl.* **163**, 113771 (2021).

9 J. Miao, Y. Wu, P. Liu, *et al.*, "Pose-guided feature alignment for occluded person re-identification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 542–551 (2019).

10 J. Guo, Y. Yuan, L. Huang, *et al.*, "Beyond human parts: Dual part-aligned representations for person re-identification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 3642–3651 (2019).

11 C. Szegedy, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199* (2013).

12 H. Wang, G. Wang, Y. Li, *et al.*, "Transferable, controllable, and inconspicuous adversarial attacks on person re-identification with deep mis-ranking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 342–351 (2020).

13 S. Wang, R. Liu, H. Li, *et al.*, "Occluded person re-identification via defending against attacks from obstacles," *IEEE Trans. Inf. Forensics Secur.* **18**, 147–161 (2022).

14 C. Zhao, X. Lv, S. Dou, *et al.*, "Incremental generative occlusion adversarial suppression network for person reid," *IEEE Trans. Image Process.* **30**, 4212–4224 (2021).

15 N. Dong, L. Zhang, S. Yan, *et al.*, "Erasing, transforming, and noising defense network for occluded person re-identification," *IEEE Trans. Circuits Syst. Video Technol.* **34**(6), 4458–4472 (2023).

16 D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision* **60**, 91–110 (2004).

17 N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of IEEE computer society conference on computer vision and pattern recognition*, 886–893 (2005).

18 T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(7), 971–987 (2002).

19 S. Pedagadi, J. Orwell, S. Velastin, *et al.*, "Local fisher discriminant analysis for pedestrian re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3318–3325 (2013).

20 S. Liao, Y. Hu, X. Zhu, *et al.*, "Person re-identification by local maximal occurrence representation and metric learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2197–2206 (2015).

21 D. Tao, Y. Guo, M. Song, *et al.*, "Person re-identification by dual-regularized kiss metric learning," *IEEE Trans. Image Process.* **25**(6), 2726–2738 (2016).

22 X. Chang, T. M. Hospedales, and T. Xiang, "Multi-level factorisation net for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2109–2118 (2018).

23 W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2285–2294 (2018).

24 H. Tian and J. Hu, "Self-regulation feature network for person reidentification," *IEEE Trans. Instrum. Meas.* **72**, 1–8 (2023).

25 G. Yan, Z. Wang, S. Geng, *et al.*, "Part-based representation enhancement for occluded person re-identification," *IEEE Trans. Circuits Syst. Video Technol.* **33**(8), 4217–4231 (2023).

26 C. Song, Y. Huang, W. Ouyang, *et al.*, "Mask-guided contrastive attention model for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1179–1188 (2018).

27 Y. Tang, X. Yang, N. Wang, *et al.*, "Person re-identification with feature pyramid optimization and gradual background suppression," *Neural Networks* **124**, 223–232 (2020).

28 S. Bai, Y. Li, Y. Zhou, *et al.*, "Adversarial metric attack and defense for person re-identification," *IEEE Trans. Pattern Anal. Mach. Intell.* **43**(6), 2119–2126 (2020).

29 Z. Wang, S. Zheng, M. Song, *et al.*, "advpattern: Physical-world attacks on deep person re-identification via adversarially transformable patterns," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8341–8350 (2019).

30 X. Wang, S. Li, M. Liu, *et al.*, "Multi-expert adversarial attack detection in person re-identification using context inconsistency," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15097–15107 (2021).

31 K. He, X. Zhang, S. Ren, *et al.*, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778 (2016).

32 O. Russakovsky, J. Deng, H. Su, *et al.*, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vision* **115**, 211–252 (2015).

33 R. T. Chen, Y. Rubanova, J. Bettencourt, *et al.*, "Neural ordinary differential equations," in *Proceedings of advances in neural information processing systems*, 6571–6583 (2018).

34 H. Yan, J. Du, V. Y. Tan, *et al.*, "On robustness of neural ordinary differential equations," *arXiv preprint arXiv:1910.05513* (2019).

35 C. Szegedy, V. Vanhoucke, S. Ioffe, *et al.*, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826 (2016).

36 Y. Wen, K. Zhang, Z. Li, *et al.*, "A discriminative feature learning approach for deep face recognition," in *Proceedings of the European conference on computer vision*, 499–515 (2016).

37 L. Zheng, L. Shen, L. Tian, *et al.*, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE international conference on computer vision*, 1116–1124 (2015).

38 E. Ristani, F. Solera, R. Zou, *et al.*, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proceedings of the European conference on computer vision*, 17–35 (2016).

39 Z. Zhong, L. Zheng, D. Cao, *et al.*, "Re-ranking person re-identification with k-reciprocal encoding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1318–1327 (2017).

40 J. Li, S. Zhang, Q. Tian, *et al.*, "Pose-guided representation learning for person re-identification," *IEEE Trans. Pattern Anal. Mach. Intell.* **44**(2), 622–635 (2019).

41 M. Wang, H. Ma, and Y. Huang, "Information complementary attention-based multidimension feature learning for person re-identification," *Eng. Appl. Artif. Intell.* **123**, 106348 (2023).

Yin Huang was born in 1995. He is currently a PhD candidate at the College of Computer Science and Technology, Qingdao University, China. He received his Master's degree from Liaoning University of Technology, China, in 2021. His research interest is multi-body dynamics and computer vision.

Jieyu Ding is currently a professor at the School of Mathematics and Statistics, and the Center for Computational Mechanics and Engineering Simulation of Qingdao University, China. She received her PhD degree in Shanghai University, and Shanghai Institute of Applied Mathematics and Mechanics, China, in 2008. Her research interest includes multi-body system dynamics, design optimization, and optimal control.

List of Figures

List of Tables

3 Impact of ODE block structure for CODEN on the Market-1501 and DukeMTMC-reID datasets.

4 Impact of different integral schemes for CODEN on the Market-1501 dataset. Error represents the accumulated error of the scheme. t_{train} represents the training time for each iteration.

5 Cross-domain evaluation on Market-1501 and DukeMTMC-reID datasets. Market $\rightarrow$ DukeMTMC represents that the model is trained on source domain Market-1501 and tested on target domain DukeMTMC-reID, and vice versa.

6 Comparing with the state-of-the-art methods on the Market-1501, DukeMTMC-reID, and CUHK03 datasets.