SAM-driven MAE Pre-training and Background-aware Meta-learning for Unsupervised Vehicle Re-identification

Dong Wang[1], Qi Wang[2,3,4], Weidong Min[2,3,4](✉), Di Gai[2,3,4], Qing Han[2,3,4], Longfei Li[2], Yuhan Geng[5]

Abstract Distinguishing identity-unrelated background information from discriminative identity information poses a challenge in unsupervised vehicle re-identification (Re-ID) tasks. Additionally, Re-ID models suffer from the challenge of varying degrees of background interference caused by continuous scene variations. The segment anything model (SAM), recently proposed, has demonstrated exceptional performance in zero-shot segmentation tasks. The combination of SAM and vehicle Re-ID models can achieve the efficient separation of vehicle identity and background information. This paper proposes a method that combines SAM-driven mask autoencoder (MAE) pre-training and background-aware meta-learning for unsupervised vehicle Re-ID. The method consists of three sub-modules. First, the segmentation capacity of SAM is utilized to separate the vehicle identity region from the background one. Given that SAM cannot be robustly employed in exceptional situations, such as ambiguity and occlusion, in vehicle Re-ID downstream tasks, a space-constrained vehicle background segmentation method is presented to obtain accurate background segmentation results. Second, SAM-driven MAE pre-training is designed. It utilizes the aforementioned segmentation results to select patches that belong to the vehicle and mask all other patches, allowing MAE to learn identity-sensitive features in a self-supervised manner. Finally, a background-aware meta-learning method is developed to fit varying degrees of background interference under different scenarios by combining different background region ratios. Extensive experiments confirm that the proposed method demonstrates state-of-the-art performance in reducing background interference variations.

Keywords Unsupervised vehicle Re-ID, Space-constrained vehicle background segmentation, SAM-driven MAE pre-training, Background-aware meta-learning.

1 Introduction

Vehicle re-identification (Re-ID) aims to perform feature similarity matching on specific vehicle targets in a cross-camera system [1–3]. Previous studies [4–6] have discovered that background information limits the capability of Re-ID models to distinguish identity information, especially in unsupervised vehicle Re-ID tasks that lack annotation. Vehicles with the same identity contain varying degrees of background information in different surveillance scenarios, which make Re-ID models sensitive to background variations. Thus, the issue of background variations greatly limits the implementation of vehicle Re-ID tasks and poses many challenges.

The primary challenge lies in the fact that distinguishing between identity-unrelated background information and discriminatory identity information poses an obstacle for vehicle Re-ID models. Existing methods [7, 8] focus on removing background interference information during the training process, thereby enhancing the sensitivity of Re-ID models to identity information. The identity and background information contained in vehicle images do not exist independently but have interdependent relationships in space. Therefore, directly removing background information may cause the learned features to lose high-dimensional spatial information, thereby reducing the robustness of Re-ID models to background variations. Recently, masked autoencoder (MAE; [9]) have been applied in vision pre-training tasks. MAE perform random masking operations on the training set and decode masked patches by encoding unmasked patches, prompting the training model to learn information related to unmasked patches. However, MAE cannot be efficiently applied to downstream vehicle Re-ID tasks. The random masking strategy exacerbates the interference of background information in downstream Re-ID models because of the possibility of discarding vehicle patches with identity information and retaining identity-unrelated background patches. Inspired by the segment anything model (SAM; [10]), this

work aims to obtain high-quality background segmentation results through low-cost prompt engineering as a guide for MAE to selectively preserve identity patches.

The second challenge is how to make Re-ID models adapt to varying degrees of background interference caused by different scene variations. Many researchers [11, 12] have regarded different background information as different domains, and this approach uses cross-domain transfer techniques to promote the alignment of different degrees of background interference. This type of method requires multiple-style transfers of samples in different scenarios, so it is difficult to apply to large-scale datasets, such as VeRi-Wild [13]. Recently, meta-learning-based methods [14–16] have achieved ideal results in overcoming domain generalization problems in Re-ID tasks. This paper believes that treating different degrees of background interference as different domains can help model learning adapt to background changes through meta-learning methods. The objective of this paper is to explore a background-aware meta-learning strategy by utilizing the region ratio of background information as vehicle identity information so that the Re-ID model can adapt to varying degrees of background information interference.

SAM-driven MAE pre-training and a background-aware meta-learning method are developed to overcome the aforementioned challenges. Experiments confirm the effectiveness of the proposed method on two publicly available datasets (i.e., VeRi-776 [17] and VeRi-Wild [13]). The main contributions of this work are summarized as follows:

(1) To ensure the robustness of SAM in performing zero-shot segmentation tasks in the vehicle Re-ID dataset, this paper proposes a space-constrained vehicle background segmentation method to optimize the background segmentation results by introducing a simple visual encoder in SAM for mining the spatial relationship between the vehicle and background region.

(2) SAM-driven MAE pre-training is proposed to enable downstream Re-ID models to learn background-unrelated identity features. Specifically, MAE is guided to selectively encode the vehicle patches by analyzing the input samples and optimized segmentation results. Then, through decoder reconstruction, the encoder indirectly learns vehicle context information related to unmasked patches.

(3) A background-aware meta-learning method is designed to make the Re-ID model adapt to varying degrees of background interference on the basis of different background region ratios.

2 Related work

2.1 Unsupervised Vehicle Re-ID

Unsupervised vehicle Re-ID task aims at mining the vehicle identity information without labeled annotations. Existing methods employ clustering-based pseudo labels as supervised information to optimize the whole unsupervised training process. Some researchers [18–20] have improved the pseudo-label generation method to improve the performance of Re-ID. Yu et al. [18] maintained a global feature dictionary and considered the similarity between samples from three aspects based on the feature dictionary to obtain more reliable identity information than density based clustering. Lu et al. [19] considered that using only global features to generate pseudo labels is unreliable, and therefore using multi-view vehicle features to improve the identifiability of feature representation and eliminate label noise. Unsupervised learning from scratch is difficult for models, so some researchers [21–23] have utilized unsupervised domain adaptation (UDA) methods to enable Re-ID models to learn identity-distinguishing features from unlabeled images. Dai et al. [22] proposed a dynamic task oriented de entanglement network (DTDN), which narrows the domain gap by establishing task-relevant and eliminating task-irrelevant relationships between the target and source domains. Wei et al. [23] proposed a domain encoder based on Transformer, which directly introduces domain information into the network to generate more robust domain-specific feature representations. Recently, MAE have been proposed for pre-training in a self-supervised manner, which achieves astonishing performance in various downstream tasks. Inspired by MAE, our motivation is to explore a robust MAE pre-training method suitable for downstream unsupervised vehicle Re-ID tasks.

1 School of Software, Nanchang University, Nanchang, 330047, China. E-mail: D. Wang, dongwang@email.ncu.edu.cn.

2 School of Mathematics and Computer Science, Nanchang University, Nanchang, 330031, China. E-mail: Q. Wang, wangqi@ncu.edu.cn; W. Min, minweidong@ncu.edu.cn(✉);D. Gai, gaidi@ncu.edu.cn; Q. Han, hanqing@ncu.edu.cn; L. Li, lilongfei@email.ncu.edu.cn.

3 Institute of Metaverse, Nanchang University, Nanchang, 330031, China. E-mail: W. Min, minweidong@ncu.edu.cn(✉); Q. Wang, wangqi@ncu.edu.cn; Gai, gaidi@ncu.edu.cn; Q. Han, hanqing@ncu.edu.cn.

4 Jiangxi Key Laboratory of Smart City,Nanchang, 330031, China. E-mail: W. Min, minweidong@ncu.edu.cn(✉); Q. Wang, wangqi@ncu.edu.cn; Gai, gaidi@ncu.edu.cn; Q. Han, hanqing@ncu.edu.cn.

5 University of Michigan, Ann Arbor, 48109, United States. E-mail: Y. Geng, gengyh@umich.edu.

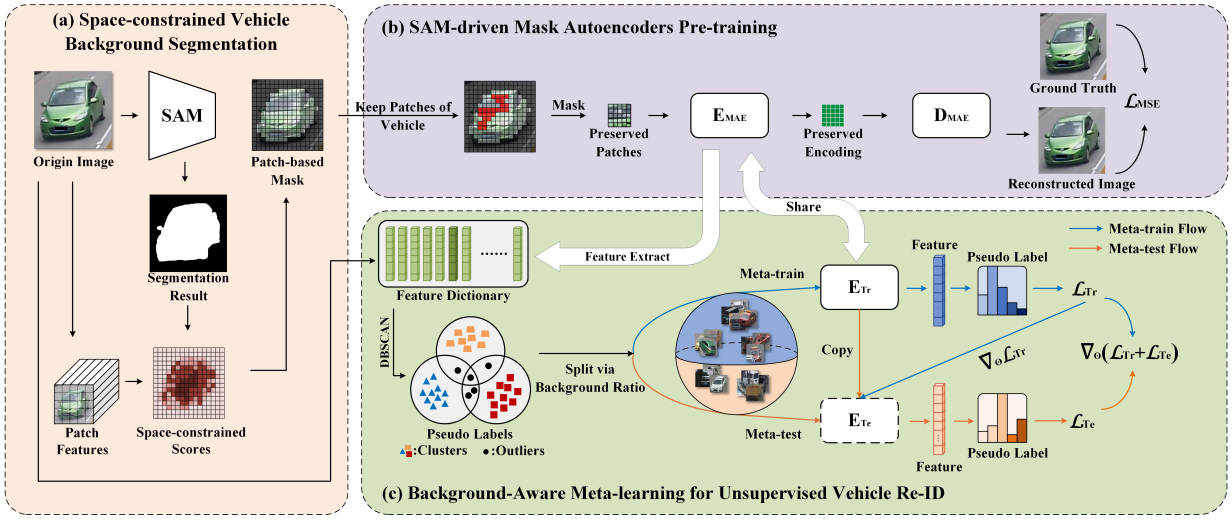


Fig. 1 Overview of the proposed method.

2.2 Background Segmentation

Accurately segmenting the main objects and background elements in a given image is crucial in computer vision. To enable the segment models to segment specific objects, some researchers [24–26] consider that the model should be provided with certain prompt information as guidance. Wu et al. [24] proposed a hierarchical modular attention network (HULANet) to achieve distribution alignment of text and image prediction through a text description driven attention mechanism. Xie et al. [26] used natural language and image features to jointly constrain the predicted object region, achieving more accurate segmentation results by establishing connections between the object, background, and text. With the popularity of large-scale training in the field of computer vision, segmentation of large models has also been proposed, such as SAM and SegGPT [27]. Among them, SAM achieved impressive zero-shot performance by building a three-stage data engine and training on masks exceeding 1B. The powerful segmentation performance of SAM can easily be migrated to the background segmentation task in vehicle images. However, some difficulties such as occlusion and blurring in the vehicle Re-ID tasks may result in incorrect segmentation results for SAM. Therefore, how to enable SAM to provide more accurate background and identity segmentation information in the vehicle Re-ID task is also the key issue to be considered in this paper.

2.3 Background-based Vehicle Re-ID

Due to the vehicle Re-ID is a cross-scene image retrieval task, vehicles with the same identity may suffer varying degrees of background interference in different scenes. Some works [4, 28, 29] deem that background information interference should be eliminated before the image is input into the network. Peng et al. [28] proposed a cross-camera adaptation framework (CCA), which utilizes StarGan to transfer cameras to the dataset and reduce the impact of background information on identity feature learning. Khoramshahi et al. [29] subtracted the original image from the non fine-grained information image generated based on Variational AutoEncoder to obtain a vehicle image that removes background interference and highlights salient information. Recently, some new methods [5, 6, 30] have achieved excellent performance in separating background information interference at the feature level. Lu et al. [6] extracted background-unrelated global features by jointly considering token features of the original image and semantic features based on vehicle masks. Zhu et al. [30] subtracted the global feature similarity from the background feature similarity based on camera ID during the similarity measurement phase to eliminate similarity bias caused by background information. The aforementioned methods only reduce background interference in the retrieval process by filtering background regions without considering the varying degrees of interference in different scenarios. The purpose of this paper is to design a novel meta-learning method that allows the Re-ID model to adapt to various degrees of background interference.

3 Proposed Method

3.1 Overview

This section designs a SAM-driven MAE and background-aware meta-learning method for unsupervised vehicle Re-ID. The overall workflow of the proposed method consists

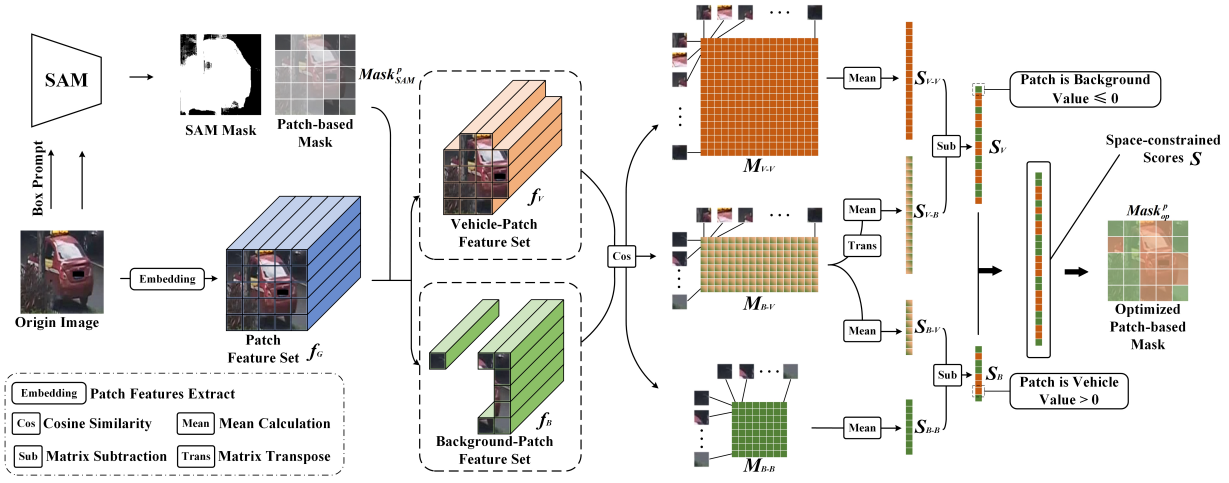


Fig. 2 Detailed process of space-constrained vehicle background segmentation.

of three modules, namely, space-constrained vehicle background segmentation, SAM-driven MAE pre-training, and background-aware meta-learning for unsupervised vehicle Re-ID, as presented in Fig. 1. In the first module, all unlabeled training samples are injected to SAM to obtain preliminary background segmentation results. Considering the unstable segmentation performance of SAM on occluded and blurred samples, we calculate space-constrained scores to optimize all segmentation results. For the SAM-driven MAE pre-training module, the optimized segmentation results are used as a guide to randomly preserve the identity patches. The whole pre-training process is conducted in a self-supervised manner; the masked image with some preserved patches is encoded using E_{MAE}, and unmasked patches is decoded using D_{MAE} to reconstruct image. Pre-training loss L_{MSE} is utilized to ensure the quality of reconstructed images. In the downstream unsupervised vehicle Re-ID task of the third module, encoder E_{MAE} serves as the baseline for extracting features from the training set and inputting them into DBSCAN [31] for clustering to obtain corresponding pseudo labels. Subsequently, the training set is dynamically divided into meta-train and meta-test sets on the basis of the range of the background region ratio. The whole meta-learning process utilizes the parameters of meta-train model E_{TR} as the initial parameters of meta-test model E_{TE} and receives supervision for losses L_{TR} and L_{TE}.

3.2 Space-constrained Vehicle Background Segmentation

With the rise of SAM, segmentation tasks based on zero-shot have become available through low-cost prompts, such as bounding boxes, without the need to train the specific segmentation model on a particular dataset. However, SAM cannot directly obtain precise segmentation results because of the low resolution, blurriness, and occlusion in the vehicle Re-ID dataset. Thus, a space-constrained vehicle background segmentation method is proposed in this paper to provide precise patch-based segmentation results for downstream Re-ID tasks. SAM is employed to roughly divide all patches of the image into vehicle identity and background information regions and further constrain and optimize the segmentation results by considering the spatial correlation between the two regions. The detailed process of the proposed method is shown in Fig. 2.

First, pixel-level background segmentation mask $Mask_{SAM} \in R^{H\times W}$ is obtained by inputting original image $I \in R^{H\times W\times 3}$ and the corresponding bounding box prompt into SAM. Second, the division rule of patches is defined to obtain patch-based mask $Mask^{p}_{SAM} \in R^{\frac{P}{H}\times\frac{P}{W}}$ (i.e., when more than half of the pixels in the patch are located in the vehicle identity region, the patch is considered a vehicle identity patch). To effectively mine the spatial correlation between patches, we extract the feature set $f_G \in R^{N\times D}$ of all patches from original image I, where $N = (H \times W)/P^2$ represents the total number of patches and D is the dimension of the feature. Based on patch-level segmentation labels for $Mask^{p}_{SAM}$, f_G can be divided into background-patch feature set f_B and vehicle-patch feature set f_V. We compute the cosine similarity between two patch feature sets to obtain similarity matrices M_{V-V}, M_{B-V}, and M_{B-B}, which can be formulated as Eq. 1:

$$\begin{cases} M_{V-V} = \cos\left(f_V, \mathrm{trans}\left(f_V\right)\right) \\ M_{B-V} = \cos\left(f_B, \mathrm{trans}\left(f_V\right)\right) \\ M_{B-B} = \cos\left(f_B, \mathrm{trans}\left(f_B\right)\right) \end{cases} \quad \text{s.t. } f_V \in R^{N_V\times D}, f_B \in R^{N_B\times D}, N_V + N_B = N \tag{1}$$

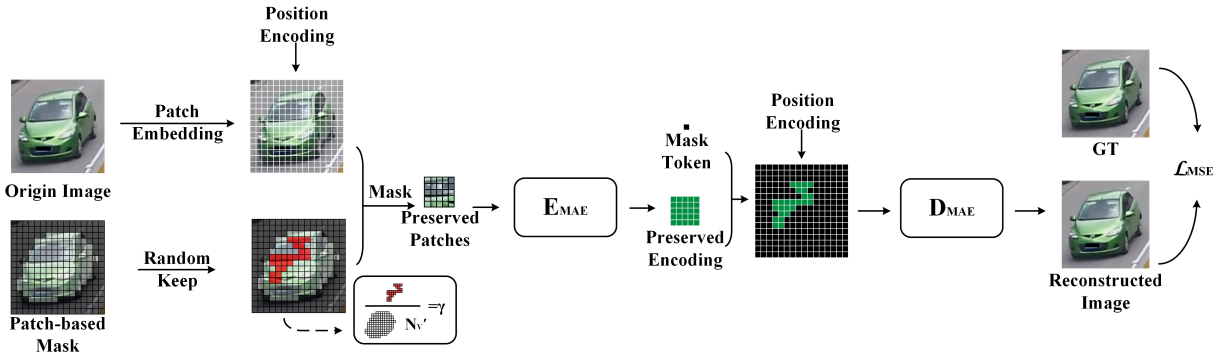


Fig. 3 Detailed process of SAM-driven MAE pre-training.

where V and B refer to the vehicle and background, respectively; M_{X-Y} represents the similarity matrix between X- and Y-patch feature sets, that is, $X, Y \in (V, B)$; $cos(\cdot)$ and $trans(\cdot)$ refer to the cosine similarity calculation and matrix transpose operation, respectively; and N_V and N_B are the patch numbers of the vehicle and background, respectively.

These similarity matrices are implemented by mean operations in the column dimension to obtain S_{V-V}, S_{B-V}, S_{V-B}, and S_{B-B}. S_{V-B} is obtained by a similar operation after transposing matrix M_{B-V}. S_{X-Y} indicates the proxy similarity score of each element in the X-patch feature set and the entire Y-patch feature set. On the basis of the four proxy similarity scores, each patch in the image is compared using the similarity between the vehicle and background regions to determine which region it should be in. Score S_X is calculated by subtracting S_{X-B} from S_{X-V} to facilitate a score comparison. S_V and S_B n the original patch order are merged to obtain space-constrained scores S. The detailed calculation process is expressed as Eq. 2:

$$S = \begin{cases} S_V = S_{V-V} - S_{V-B} \\ S_B = S_{B-V} - S_{B-B} \end{cases} \quad \text{s.t. } S_V \in R^{N_V}, S_B \in R^{N_B}, S \in R^N \tag{2}$$

After the values in S are obtained by subtracting S_{X-B} from S_{X-V}, the positive and negative situations of each value provide a basis for determining which region each patch should be in. Specifically, when the space-constrained score S_i of the i-th patch is greater than 0, the patch is considered a vehicle patch; when S_i is less than 0, the patch is considered a background patch; and when S_i is equal to 0, the patch is a noise patch with the same similarity as the vehicle and background regions and treated as a background patch. Through this processing, optimized patch-based mask $Mask^p_{op}$ is obtained, and it provides precise patch-based background segmentation information.

3.3 SAM-driven MAE Pre-training

Learning robust identity representations is crucial for unsupervised vehicle Re-ID tasks. However, existing unsupervised Re-ID models cannot easily separate identity-unrelated background information during the representation learning process. The main reason is that most models increasingly focus on background information errors during each iteration of training. Enhancing the sensitivity of Re-ID models to discriminative identity information is the key to solving the abovementioned problem. This section designs a SAM-driven MAE pre-training method that enhances feature extraction of vehicle identity regions through a SAM-guided pre-trained model based on MAE architecture. The pre-trained model has high sensitivity to vital vehicle identity information in downstream tasks, and its detailed process is illustrated in Fig. 3.

In the pre-training encoding step, given image I is divided into N patches of size P^2, and patch embedding is performed. Assuming that the obtained embeddings are directly inputted into MAE, MAE's random masking strategy ensures that all patches have the same possibility of being preserved. When the background patch is preserved, the encoder may learn background-related interference information. This paper optimizes the original random masking strategy of MAE by using the optimized patch-based mask $Mask^p_{op}$ obtained in Section 3.2 as guide information to randomly preserve partial vehicle patches. Given that the number of vehicle patches is N'_V, the preserved ratio is set to γ. Our masking strategy selects a total of $N'_V \times \gamma$ patches during each iteration process and inputs the preserved patches into the encoder E_{MAE} for encoding operations to obtain the corresponding encodings.

In the decoding step, the preserved encodings are restored to the position of the corresponding patches. The positions of the previously masked patches are supplemented by the same learnable mask token. After positional embeddings

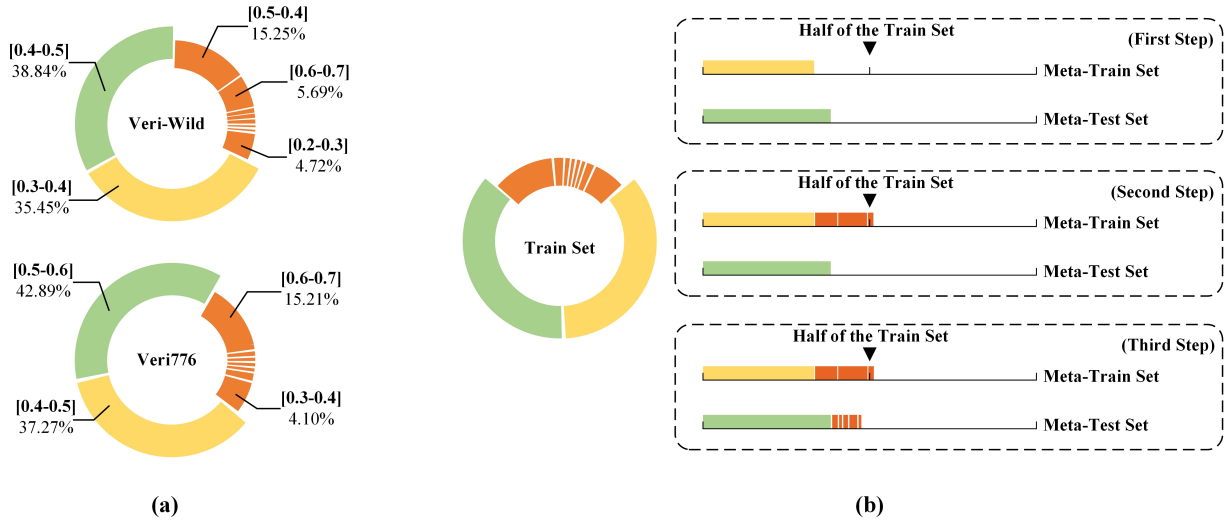


Fig. 4 Visualization of the detailed information of the meta-train and meta-test sets' splitting strategy. (a) Proportion of images in the ratio range of each background region in the VeRi-776 and VeRi-Wild training sets. (b) Proposed meta-set splitting strategy.

are added, the set composed of these encoding and tokens are inputted into decoder D_{MAE}. During the reconstruction process, decoder D_{MAE} can only predict all masked patches on the basis of the preserved encodings provided by encoder E_{MAE}. The reconstructed vehicle regions do not suffer from background information interference because the background patches are not retained, thereby improving the efficiency of reconstruction. The pre-trained model is updated and optimized by calculating the mean-squared error between reconstructed image I_{rec} and original image I. The loss function L_{MSE} of SAM-driven MAE pre-training is shown as Eq. 3:

$$L_{MSE} = \frac{1}{W \times H} \sum_{i=1}^{W \times H} \left(I^i - I^i_{rec}\right)^2 \tag{3}$$

where I^i and I^i_{rec} are the i-th pixels of the original and reconstructed images, respectively.

In the whole self-supervised pre-training process, decoder D_{MAE} performs contextual semantic inference with high correlation on the basis of the given vehicle patch encodings, and the ground truth continuously corrects the reconstruction results. This process makes encoder E_{MAE} sensitive to vehicle identity information, thus providing a robust pre-trained model that distinguishes between identity and identity-unrelated background information for downstream unsupervised vehicle Re-ID tasks.

3.4 Background-aware Meta-learning for Unsupervised Vehicle Re-ID

Although existing unsupervised vehicle Re-ID methods have impressive performance, they still suffer from varying degrees of background interference caused by scene variations. The region ratio of the same identity vehicle body in 2D pixel space varies because of varying degrees of background interference. The reduced sensitivity of unsupervised Re-ID models to background variations leads to considerable differences in intraclass features, thereby reducing the accuracy of feature learning. This paper proposes a background-aware meta-learning approach that splits the original training set into meta-train and meta-test sets in accordance with varying background interference. The degree of background interference is simulated by calculating the ratio of the background region of vehicles in each image. The proposed meta-learning learns background-invariant features, and it consists of four steps: meta-set split, meta-train, meta-test, and meta-optimize.

Meta-sets split. Given training set U, this paper uses DBSCAN to generate pseudo labels for it. To adjust the Re-ID model learning to different degrees of background interference, we simulate completely different background interference distributions in the meta-train and meta-test sets. On the basis of the optimized patch-based mask $Mask^p_{op}$ (obtained in Section 3.2) of all images, the ratio of the background region $r \in (0, 1)$ in the corresponding image is computed. The ratio of the background region is split into an average of 10 intervals (every 0.1 represents an interval), and all images in the training set are divided into 10 subsets depending on which interval r is in.

As shown in Fig. 4(a), the background region of most of the images in the vehicle Re-ID dataset is concentrated in few intervals. Direct random division based on the intervals may result in an extremely unbalanced number of images in

the meta-train and meta-test sets. A balanced split strategy is adopted in this paper, as shown in Fig. 4(b): First, the two subsets with the largest number of images are randomly split into meta-train and meta-test sets. Second, the other subsets are randomly divided into meta-train set one by one until the number of images in the meta-train set exceeds half of the total number of images. Last, all remaining subsets are allocated to the meta-test set.

Meta-train. In the meta-train step, encoder E_{TR} uses the pre-trained model E_{MAE} in Section 3.3 for parameter initialization, samples the meta-train set, and employs E_{TR} to compute the meta-train loss. The proposed method uses triplet loss L_{Tri} and cross-entropy loss L_{CE} with label smoothing as total loss L_{TR} at the meta-train stage to improve model performance. The computation process can be formulated as Eq. 4:

$$\begin{cases} L_{Tri} = \max\left(d_p - d_n + \alpha\right) \\ L_{CE} = -\sum_{i=1}^{N} \tilde{y}_l \cdot \log\left(q_i\right) \\ L_{TR} = L_{Tri} + L_{CF} \end{cases} \tag{4}$$

where d_p and d_n represent the distance of positive and negative sample pairs in the mini-batch, respectively, and α is the margin of triplet loss. $\tilde{y}_i = \beta y_i + (1-\beta)v$ represents constant β label smoothing for pseudo-label y_i, v is a uniform vector, and q_i is the classification prediction for the image.

Meta-test. In the meta-train step, parameters θ_{TR} of E_{TR} are used to construct a temporary model E_{TE} with meta-train loss L_{TR}. Parameters θ_{TE} of E_{TE} can be obtained from Eq. 5:

$$\theta_{TE} = \theta_{TR} - lr\frac{\partial L_{TR}}{\partial \theta_{TR}} \tag{5}$$

where lr is the learning rate. Then, E_{TE} is employed to calculate meta-test loss L_{TE} for the images sampled in the meta-test set, similar to Eq. 4.

Meta-optimize. Overall optimization of the model can be achieved based on the learning and adaptation of the model to different background interference tasks in meta-train and meta-test flow. The total loss and overall model parameter updates are shown in Eq. 6 and Eq. 7, respectively.

$$L_{Total} = L_{TR} + L_{TE} \tag{6}$$

$$\frac{\partial L_{Total}}{\partial \theta_{TR}} = \frac{\partial L_{TR}}{\partial \theta_{TR}} + \frac{\partial L_{TE}}{\partial \theta_{TE}}\frac{\partial \theta_{TE}}{\partial \theta_{TR}} \tag{7}$$

The aforementioned process constructs meta-train and meta-test tasks with different degrees of background interference. It continuously motivates the Re-ID model to adapt to different degrees of background interference during iterative training and learn other robust background-invariant features.

The overall training process of the proposed SAM-driven MAE pre-training and background-aware meta-learning for unsupervised vehicle Re-ID method is summarized in Algorithm. 1.

Algorithm 1 Procedure of proposed method.

Input: Unlabeled training set U, bounding box prompt T, batch size b, segment anything model SAM, encoder E_{MAE}, decoder D_{MAE}.
Output: Optimized model parameters θ^*
//Space-constrained Vehicle Background Segmentation;
for image I in U **do**
  Input I and $T(I)$ into SAM to obtain patch-based mask $Mask^p_{SAM}$;
  Compute space-constrained scores S based on $Mask^p_{SAM}$ with Eq. 1 and Eq. 2;
  Obtain optimized patch-based mask $Mask^p_{op}$ based on S;
end for
//SAM-driven MAE Pre-training;
for image $iter$ in $pre-train_iters$ **do**
  Sample mini-batch with b in U to obtain u;
  Randomly preserve some vehicle patches in u through $Mask^p_{SAM}$ and input E_{MAE} to obtain patch-feature encoding f_{MAE};
  Fill f_{MAE} with mask tokens and input D_{MAE} to obtain the reconstructed image;
  Compute L_{MSE} with Eq. 3;
  Update parameters for E_{MAE} and D_{MAE} based on L_{MSE};
end for
//Background-aware meta-learning for Re-ID;
Generate pseudo-labels for U with DBSCAN;
Split U into U_{TR} and U_{TE};
for image $iter$ in $train_iters$ **do**
  Samples mini-batch with b from U_{TR} and U_{TE} to obtain u_{TR} and u_{TE}, respectively;
  Build E_{TR} using pre-trained E_{MAE} parameters and performing meta-train flow;
  Build E_{TE} and performing meta-test flow;
  Optimize θ_{TR} with gradient computed by Eq. 7;
end for
$\theta^* \leftarrow \theta_{TR}$;
Result: θ^*

4 Experiments

4.1 Datasets and Evaluation Protocols

Datasets. Extensive experiments are conducted on two widely used datasets: VeRi-776 and VeRi-Wild. The contents of the VeRi-776 dataset were collected from 20 cameras covering a real traffic monitoring area of 1 km^2 within 24 h. It has a total of 50,117 images of 776 vehicles, including 37,778 images in the training set, 1,678 images in the query, and 10,661 images in the gallery. VeRi-Wild is a large-scale vehicle Re-ID dataset. It contains 416,314 images of 40,671 vehicles that were obtained from 174 cameras that recorded images within a month. The images were captured under the influence of various environmental factors, such as backgrounds, lighting, viewpoints, and weather. The training set includes 277,794 images of 30,671 vehicles. VeRi-Wild divides the test set into three subsets. The small subset includes 41,816 images of 3,000 vehicles, the medium subset includes 69,389 images of 5,000 vehicles, the large subset includes 138,517 images of 10,000 vehicles.

Evaluation Protocols. Mean average precision (mAP) and cumulative matching characteristics (CMC) are employed to evaluate the performance of unsupervised vehicle Re-ID methods. mAP is a widely used evaluation metric in object detection tasks. It measures average precision by balancing accuracy and recall. CMC, on the other hand, focuses on the ranking-based performance of the model. It measures the accuracy of the top K matching results for a given query image. In the experiments, mAP, Rank-1, and Rank-5 are calculated to compare the performance of the evaluated methods.

4.2 Implementation Details

In the space-constrained vehicle background segmentation step, a generalizable detection model with a small annotation cost is trained based on YOLOv8, which is employed to provide accurate bounding box prompts for SAM. In SAM-driven MAE pre-training, all samples are trained in 50 epochs, and the batch size is set to 64. For the unsupervised vehicle Re-ID downstream task, each image is augmented by random horizontal flipping, padding, cropping, and erasing. The total epochs of the Re-ID model are set to 60, with each epoch consisting of 600 iterations. Each iteration involves learning from a mini-batch of 64 samples, each containing four images for each of the 16 pseudo-classes. CLIP-B/16 [32] is used as the network encoder and participates in the steps in Sections 3.2 and 3.3. In both steps, all images are resized to 256×256, and the Re-ID model is updated by the Adam optimizer. Considering device limitations, we choose 60,000 images from the VeRi-Wild dataset and use them as the training set. All experiments are conducted with the Ubuntu18.04 operating system and in Pytorch environment with 4 Tesla P40.

4.3 Ablation Study

Impact of different preserved rate. Table 1 presents the effects of different preserved rates on the downstream Re-ID tasks. The preserved rate in this paper is calculated based on the number of preserved patches in the vehicle identity region of the segmentation result. The experimental results indicate that as the preserved rate increases, the performance of the Re-ID model gradually decreases. The purpose of SAM-driven MAE is to provide preserved vehicle patches and allow the encoder–decoder architecture to learn information about masked vehicle patches. Therefore, the pre-training model can only learn a small amount of vehicle identity information when massive vehicle patches are preserved. In this case, the pre-training model becomes susceptible to background interference because the masked patches contain abundant background information. Overall, the preservation of massive vehicle patches during pre-training limits the feature learning effectiveness of downstream Re-ID tasks. Thus, the preserved rate of pre-training is uniformly set to 25% in the subsequent experiments.

Table 1 Performance comparison of SAM-driven MAE pre-training with different preserved rates in VeRi-776.

Preserved Rates	Rank-1	Rank-5	mAP
75%	72.0	77.9	26.3
50%	78.0	84.6	32.9
25%	**83.3**	**88.0**	**36.7**

Table 2 Performance comparison of meta-learning strategies based on different attributes in VeRi-776. The meaning of the "Bg ratio" is the ratio of background region.

Split Attributes	Rank-1	Rank-5	mAP
Vehicle Model	82.4	86.5	36.7
Color	81.2	86.1	36.5
Bg ratio (Pixel Level)	83.8	87.4	37.9
Bg ratio(Ours)	**86.8**	**90.6**	**38.4**

Effect of meta-learning strategy. As shown in Table 2, different attributes are used to replace the background region ratio in the proposed method when a meta-set split is performed to verify the effectiveness of the proposed method. The vehicle model and color attributes of each image are predicted by CLIP to ensure fairness in self-supervised learning. The attribute label sets is defined as:

Table 3 Ablation study of the different baselines with proposed method on VeRi-776 and VeRi-Wild datasets. Among them, "TransReID-I" and "TransReID-D" respectively represent the TransReID baseline pre-trained on ImageNet and DeiT, "TMGF-L" represents the TMGF baseline pre-trained on Luperson, and "MAE-Random" represents the CLIP based baseline constructed in this paper.

Different Modules	VeRi-776			VeRi-Wild								
				Test3000			Test5000			Test10000		
	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP
Ours(TransReID -I)	78.9	83.8	36.6	58.6	80.1	31.1	53.0	75.6	28.7	40.7	64.0	21.8
Ours(TransReID-D)	79.8	85.8	36.9	61.4	81.5	32.8	50.7	74.1	27.3	43.3	66.3	23.2
Ours(TMGF-L)	79.5	85.7	33.7	60.6	82.0	33.1	53.2	74.9	28.4	43.4	65.7	23.1
Ours(MAE-Random)	**86.8**	**90.6**	**38.4**	**63.3**	**83.6**	**34.0**	**54.5**	**77.1**	**29.4**	**44.4**	**67.4**	**23.9**

Table 4 Ablation study of the different modules on VeRi-776 and VeRi-Wild datasets.

Different Modules	VeRi-776			VeRi-Wild								
				Test3000			Test5000			Test10000		
	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP
MAE(Random)	68.2	74.8	24.5	55.7	75.2	30.5	48.0	67.9	26.0	39.1	59.4	20.9
Ours(w/o Bg-Meta)	83.3	88.0	36.7	56.5	75.1	30.6	48.5	68.7	26.2	39.8	60.4	21.1
Ours(w/o SAM-driven MAE)	83.4	87.8	37.9	52.7	71.6	29.2	44.7	64.6	24.7	36.3	56.0	19.7
Ours(w/ Patch-Seg)	72.6	78.4	29.9	54.9	75.9	29.4	47.4	68.8	25.1	38.3	59.3	20.0
Ours	**86.8**	**90.6**	**38.4**	**63.3**	**83.6**	**34.0**	**54.5**	**77.1**	**29.4**	**44.4**	**67.4**	**23.9**

Color: black, white, silvery, red, yellow, blue, green, golden, khaki, pink;

Vehicle Model: sedan, bus, van, truck, hatchback, suv, mpv, jeep;

Meta-learning methods based on the two attributes have not achieved notable results because of the considerable appearance differences inherent in images of different colors and models. A comparison of the two methods that uses the pixel-level background information and the proposed patch-level background region ratio shows that the proposed method has better performance because the pixel-level background segmentation information obtained directly from SAM is inaccurate, and inaccurate segmentation information misleads the learning of the Re-ID model. According to the analysis above, the proposed method can effectively represent background interference information that is difficult to describe, so it can effectively help the Re-ID model adapt to background variations.

Baseline comparison. Due to the need for patch level features in the proposed method, a baseline based on ViT and CLIP is considered for performance comparison. Among them, TransReID [33] (modified to unsupervised architecture) and TMGF [34] are based on ViT, while "MAE-Random" is based on CLIP. As shown in Table 3, compared to ViT based baselines, the "MAE-Random" architecture baseline is more adaptable to the proposed method. This is because CLIP is a visual-linguistic pre-trained model. CLIP with advanced semantic information guidance is different from ViT that only focuses on visual information. It is easier to separate background information from identity information, thus achieving better performance.

Effect of different modules. To investigate the effectiveness of each module of the proposed method, we conduct ablation experiments on the performance of different modules, as shown in Table 4. The explanation for each module is as follows:

(1) **"MAE(Random)"** indicates the use of MAE pre-training on the basis of random masks, followed by unsupervised downstream training of vehicle Re-ID.

(2) **"Ours(w/o Bg-Meta)"** indicates the use of SAM-driven MAE pre-training, followed by unsupervised downstream training of vehicle Re-ID.

(3) **"Ours(w/o SAM-driven MAE)"** indicates that without any MAE pre-training, unsupervised vehicle Re-ID with background-aware meta-learning is directly performed.

(4) **"Ours(w/ Patch-Seg)"** indicates the direct use of patch-level segmented images as input in downstream training, using the entire method proposed.

(5) **"Ours"** refers to the use of all modules of the proposed method.

According to the experimental results, "Ours(w/o Bg-Meta)" has better performance on the VeRi-Wild dataset compared with the other methods. The Rank1 values of "Ours(w/o Bg-Meta)" are 3.8%, 3.8%, 3.5% higher in the three test sets compared with the Rank1 values of "Ours(w/o SAM-driven MAE)". This result demonstrates that the pro-

Table 5 Comparison of different training strategies for UDA tasks.

Strategy	VeRi-Wild→VeRi-776			VeRi-776→VeRi-Wild								
				Test3000			Test5000			Test10000		
	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP
MAE(S)	88.1	**94.0**	**58.8**	61.7	83.1	34.7	54.2	76.4	29.9	43.7	67.2	24.1
MAE(T)	**88.9**	93.7	56.1	**63.3**	**84.0**	**34.7**	**54.5**	**77.1**	**30.0**	**44.0**	**67.7**	**24.2**
MAE(S+T)	88.7	93.4	52.3	62.7	82.8	34.3	52.9	76.2	29.4	42.7	66.8	23.6

posed SAM-driven MAE pre-training allows downstream Re-ID models to learn additional robust identity features. The results of "Ours(w/o SAM-driven MAE)" show relatively balanced performance in two datasets, proving that the proposed background-aware meta-learning method has excellent adaptability to datasets with different degrees of background interference. Observing the results of "Ours(w/Patch-Seg)", although directly applying patch level image segmentation results downstream can achieve the most direct separation of background information. However, due to the presence of high-dimensional spatial information in the background, directly removing the background from the image cannot achieve effective performance. Compared with the methods that use individual modules, the "Ours" method that employs all modules exhibits the best performance in all the evaluation indicators and test sets. This finding proves that the "Ours" method combines the advantages of adapting to being sensitive to discriminative identity information and adapting to varying degrees of background interference in the Re-ID model.

Analysis of unsupervised domain adaptation training strategies. The training of the unsupervised domain adaptation (UDA) task for vehicle Re-ID is divided into two stages: supervised pre-training in the source domain and unsupervised fine-tuning in the target domain. The proposed method provides a robust pre-training model to make the Re-ID tasks focus on discriminative vehicle identity information. To explore the effectiveness of the proposed SAM-driven MAE pre-training in UDA tasks, we compare three pre-training strategies, as shown in Table 5. The specific explanation for each strategy is as follows:

(1) **"MAE(S)"** indicates a training strategy of conducting SAM-driven MAE pre-training in the source domain, followed by supervised learning in the source domain and unsupervised fine-tuning in the target domain.

(2) **"MAE(T)"** refers to a training strategy that involves supervised learning in the source domain, followed by SAM-driven MAE pre-training in the source domain and unsupervised fine-tuning in the target domain.

(3) **"MAE(S+T)"** indicates a training strategy of using images from the source and target domains for SAM-driven MAE pre-training, followed by supervised training in the source domain and unsupervised fine-tuning in the target domain.

According to the experimental results, the "MAE(T)" strategy performs much better than the "MAE(S)" strategy does. However, for the "VeRi-Wild→VeRi-776" task, the "MAE(S)" strategy has a higher Rank-5 value and mAP than the "MAE(T)" strategy because the VeRi-Wild dataset has many images and complex vehicle information.The self-supervised pre-training conducted on the VeRi-Wild dataset enables the Re-ID model to learn abundant robust feature representations on the relatively simple VeRi-776 dataset. The comparison of these experimental results proves that self-supervised SAM-driven MAE pre-training before supervised training hinders the Re-ID model from adapting the information learned in the source domain to the target domain. However, after supervised training in the source domain, performing SAM-driven MAE pre-training in the target domain can effectively convey the information learned by the model in the source domain. This finding also indirectly confirms that the proposed SAM-driven MAE pre-training can alleviate the domain gap in UDA tasks.

4.4 Comparison with State-of-the-arts

Existing state-of-the-art methods are compared with the proposed method in Table 6. The proposed method is superior to the other methods in all evaluation indicators. MetaCam [39] employs a meta-learning strategy to overcome camera variations by using camera annotations. Compared with MetaCam, the proposed method fully considers the interference caused by background variations without using any annotation information, and it outperforms MetaCam by 8.8% on VeRi-776 and 2.8% on VeRi-Wild (Test3000) in terms of mAP. These results prove the effectiveness of the proposed meta-learning method and its low manual annotation dependency. Compared with the currently best-performing methods of GroupSampling [36] and GCMT [42], the proposed method has better performance on VeRi-776 and VeRi-Wild datasets, respectively. The key reason is that the two methods focus on

Table 6 Comparison of the proposed method with state-of-the-art methods on VeRi-776 and VeRi-Wild datasets.

Methods	VeRi-776			VeRi-Wild								
				Test3000			Test5000			Test10000		
	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP
SSML[18]	74.5	80.3	26.7	49.6	71.0	23.7	43.9	64.9	20.4	34.7	55.4	15.8
CACL[35]	57.0	69.7	24.5	57.5	80.7	30.1	49.3	73.8	26.0	39.3	64.6	20.5
GroupSampling[36]	77.8	83.8	35.0	58.8	80.9	30.9	51.7	74.1	26.9	41.2	64.8	21.5
HHCL[37]	65.2	70.9	30.2	57.5	78.5	30.4	49.1	72.6	26.1	39.1	62.2	20.6
ISE[38]	66.0	72.5	27.7	61.2	81.8	32.9	54.2	75.6	29.0	44.3	66.7	23.5
MetaCam[39]	72.6	80.3	29.6	58.8	79.9	31.2	51.3	74.2	27.2	41.0	64.7	21.6
SpCL[40]	65.6	74.0	28.3	59.8	81.1	31.4	50.9	74.7	27.1	40.8	65.1	21.7
Strong Baseline[41]	58.5	69.0	23.8	50.1	74.9	25.6	42.4	67.9	21.8	32.7	57.2	16.8
MMT[41]	60.9	69.0	25.4	60.6	82.5	31.8	52.1	76.8	27.6	42.4	66.8	22.2
GCMT[42]	80.6	87.2	35.2	54.9	78.3	29.3	45.7	71.3	25.0	36.1	60.3	19.5
AE[43]	32.4	48.9	9.0	12.8	23.5	4.2	10.0	19.1	3.3	7.3	14.9	2.4
Ours	**86.8**	**90.6**	**38.4**	**63.3**	**83.6**	**34.0**	**54.5**	**77.1**	**29.4**	**44.4**	**67.4**	**23.9**

Table 7 Comparison of the proposed method with state-of-the-art UDA vehicle Re-ID methods on source dataset → target dataset tasks.

Methods	VeRi-Wild→VeRi-776			VeRi-776→VeRi-Wild								
				Test3000			Test5000			Test10000		
	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP
HHCL[37]	31.5	42.7	11.6	41.7	65.0	18.5	35.6	58.2	15.7	26.8	47.7	11.8
MetaCam[39]	65.3	76.6	26.7	25.8	41.7	10.2	22.3	36.2	8.5	16.9	30.1	6.4
SpCL[40]	73.7	81.7	35.5	60.8	82.2	32.2	52.7	76.1	28.1	42.3	66.8	22.4
Strong Baseline[41]	65.3	75.1	26.5	57.6	80.8	30.4	50.2	74.5	26.3	39.7	63.6	20.7
MMT[41]	70.1	78.5	31.5	61.9	83.0	32.4	53.2	76.9	28.1	42.8	67.2	22.6
AE[43]	75.9	86.6	40.2	34.0	55.9	15.1	29.0	51.1	12.2	21.6	41.0	8.7
AWB[44]	81.2	86.4	38.2	59.0	80.4	31.9	51.0	74.0	27.4	40.9	64.8	21.9
GLT[45]	78.1	83.5	36.7	56.4	78.9	29.9	48.0	71.9	25.5	38.0	62.3	20.1
Ours	**88.9**	**93.7**	**56.1**	**63.3**	**84.0**	**34.7**	**54.5**	**77.1**	**30.0**	**44.0**	**67.7**	**24.2**

the clustering level in unsupervised learning and ignore the background interference. The proposed method demonstrates state-of-the-art performance in unsupervised vehicle Re-ID tasks.

The latest methods for some UDA tasks are also compared with the proposed method, as shown in Table 7. The proposed method surpasses the other methods by large margins regardless of whether the target domain is VeRi-776 or VeRi-Wild. Specifically, on the VeRi-776 dataset, the Rank-1 and mAP of the proposed method are 7.7% and 17.9% higher than those of the best performing method AWB [44], respectively. In the case of VeRi-Wild (Test 3000), the Rank-1 and mAP of the proposed method are 1.4% and 2.3% higher than those of MMT [41], respectively. AE[43] and GLT[45] methods optimize representation learning in the latent space to reduce label noise and domain differences. However, the abstract nature of representation learning can be difficult to control in iterative training, making it challenging for the model to accurately capture discriminative identity information. To address this, our proposed method utilizes SAM to provide efficient and precise background guidance, increasing the model's sensitivity to identity information and improving overall performance. Additionally, after comparing the performance of the methods on UDA and USL tasks, we observe a massive improvement on VeRi-776. This improvement indicates that our method can effectively learn robust identity information and prompts the pre-training model to apply the knowledge learned from large-scale datasets to downstream UDA tasks.

4.5 Qualitative Analysis

Visualization of the segmentation result. As shown in Fig. 5, the pixel-level segmentation results obtained by SAM for different scenes (i.e., (a), (b), (c), and (d)) are inaccurate, leading to incorrect guidance for downstream tasks. As indicated in the fourth column of Fig. 5, the proposed space-constrained vehicle background segmentation method can be optimized

based on the segmentation results of SAM, further distinguishing between vehicle and background information. For example, in Fig. 5(c), our method corrects the result of SAM that mistakenly divides shrubs and trees into vehicle regions.

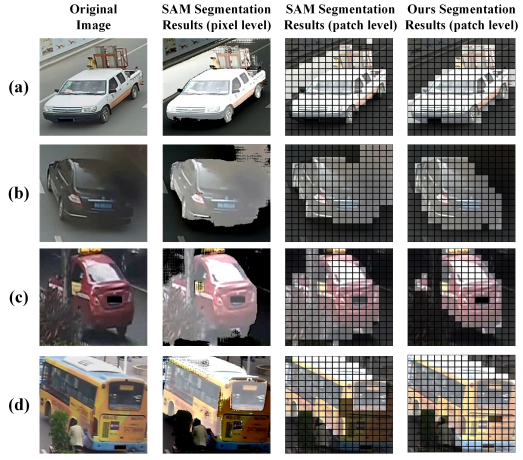


Fig. 5 Visual comparison of segmentation results obtained by directly applying SAM and the proposed method with various vehicle models in three different scenes: (a) complex vehicle structure, (b) blurred scene, (c) static occluded scene with shrub, and (d) dynamic occlusion scene with pedestrian.

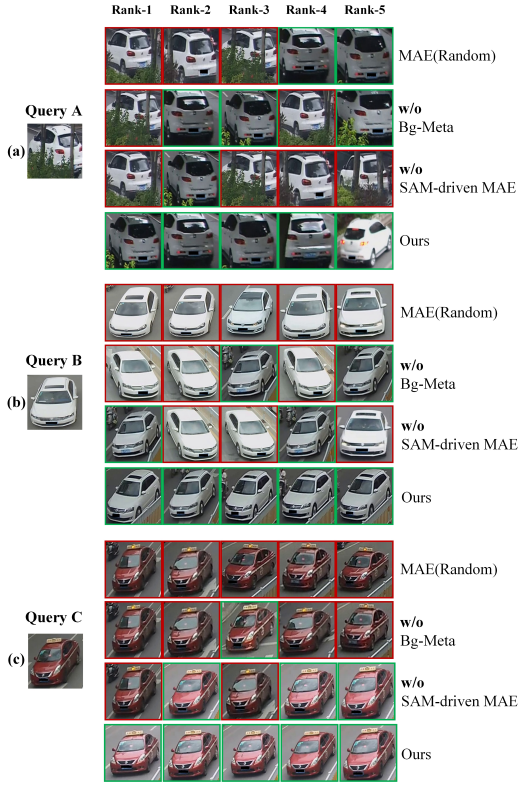


Fig. 6 Top 5 rank lists were retrieved for queries with various vehicle models by different ablation modules. The green and red boxes indicate positive and negative candidate samples, respectively.

Visualization of the rank list. The retrieval results of the four methods in Table 4 are visualized to reveal the effectiveness of the proposed method intuitively. In Fig. 6, the top 5 rank list results for the corresponding query are given. According to the rank lists, the "MAE (Random)" method that does not consider background information cannot distinguish similar structures in an example with different identities. For Query A in Fig. 6(a), due to the inability to identify the same structure of vehicles with different identities passing through shrubs, the "MAE (Random)" method mistakenly identifies the top three candidate samples as positive samples. Compared with the methods that use individual modules, the "Ours" method that employs all modules retrieves the top five of the rank lists correctly, thus alleviating this problem to varying degrees. In terms of Query B in Fig. 6(b), the "Ours" method can accurately identify positive samples in the background of pedestrian interference. For Query C in Fig. 5(c), the the proposed method prioritizes the intricate details of the vehicle region. This enables it to effectively mitigate variations in the background environment resulting from lighting changes and accuracely retrieve the top five positive samples.

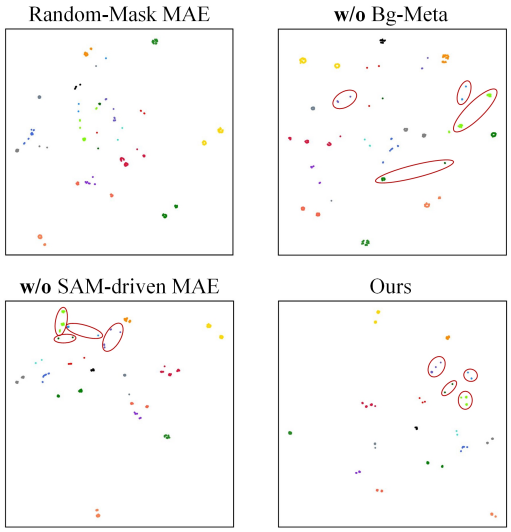


Fig. 7 T-SNE visualization of the feature distribution of different ablation modules. Each point of the same color belongs to the same class.

Visualization of T-SNE. The feature learning ability of the proposed method in qualitative analysis is also assessed. Twenty classes in the training set of VeRi-776 are randomly selected, and their feature distributions are visualized. As shown in Fig. 7, compared with the "MAE(Random)" method, the "Ours(w/o Bg-Meta)" method makes the Re-ID model more sensitive to vehicle appearance information, thereby effec-

tively widening the distance among various classes. However, due to the effects of background variations, the "Ours(w/o Bg-Meta)" method still maintains a large intra-class distance. Compared with "Ours(w/o Bg-Meta)", "Ours(w/o SAM-driven MAE)" makes the Re-ID model adapt to varying degrees of background interference, thus remarkably reducing the distance within each class. When the "Ours" method that combines the advantages of two modules is used, a reliable feature distribution is obtained. For example, the red circles in the "Ours" method have small distances inside, but the distances between circles are large.

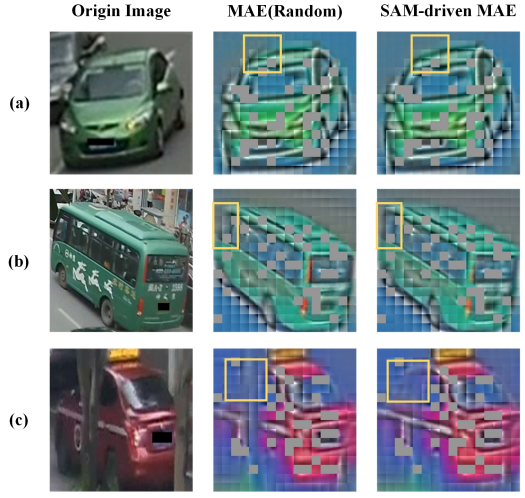


Fig. 8 Visualization of the reconstructed images by "MAE (Random)" and "SAM-driven MAE" methods. The gray patches are preserved patches used for image reconstruction. Both methods preserve patches in the same position during image reconstruction to ensure a fair comparison.

Visualization of reconstructed effects. We compare the reconstruction effects of "MAE (Random)" and "SAM-driven MAE", as illustrated in Fig. 8. The "SAM-driven MAE" method is more accurate than the "MAE (Random)" method in reconstructing fine-grained information about vehicles. As shown in Fig. 8(a) and (b), the "MAE (Random)" method produces blurrier results compared with the proposed method in the reconstruction of vehicle profiles. As indicated in Fig. 8(c), the "SAM-driven MAE" method still reconstructs the vehicle contour for the patches obstructed by trees in the original image, but the "MAE (Random)" method is ineffective. These situations indirectly confirm that our method can effectively provide a robust pre-training model that can distinguish between background and discriminative identity information.

Table 8 Comparison of complexity and performance on VeRi-776 dataset. The meaning of "B-A" is Background-Aware ability.

Method	Stage	B-A	mAP	Rank-1
ISE	1	✘	27.7	66.0
MetaCam	1	✘	29.6	72.6
SSML	1	✘	26.7	74.5
GroupSampling	1	✘	35.0	77.8
GCMT	1	✘	35.2	80.6
Ours	**2**	✔	**38.4**	**86.8**

4.6 Discussion on method complexity

The proposed method employs SAM to obtain low-cost background segmentation information, which guides the model to perform two-stage background information separation learning: SAM-driven MAE pre-training and Background-Aware Meta-learning. As shown in Table 8, compared to current methods, the proposed method utilizes extra end-to-end pre-training and has higher complexity. However, as a reward, the proposed method has a certain level of background-aware ability and has achieved more competitive performance.

4.7 Discussion on Person Re-ID

The performance of unsupervised person Re-ID methods is similarly constrained by image background factors. To verify the proposed method's universality and generalization ability, we compared it with the latest approaches in the field of unsupervised person Re-ID. The experiments were conducted on the Market-1501 dataset, and the specific results are shown in Table 9. Compared to the state-of-the-art method MetaCam, the proposed method achieved improvements of 3.8% and 19.5% in Rank-1 and mAP, respectively. Despite the more complex interference of background elements in pedestrian re-identification datasets, the proposed method still demonstrates competitive performance. This directly attests to the effectiveness of the proposed method in the task of unsupervised person Re-ID.

To better understand the resistance of the proposed method to background interference in unsupervised person Re-ID tasks, we visualized the focal regions of model features and compared them with existing methods. As shown in Fig. 9, compared to other methods, our approach makes the model more sensitive to the human body region and pays less attention to background areas lacking identity information. This is because our method utilizes pre-training and meta-learning, separating identity-independent information from interfering with model representation learning. It effectively guides the model to focus on the unique areas of person images, resulting in the learning of more robust features.

Table 9 Comparison with state-of-the-art unsupervised person Re-ID methods on datasets of Market-1501.

Methods	Rank-1	Rank-5	mAP
MMCL [46]	80.3	89.4	45.5
BUC [47]	66.2	79.6	38.3
ECN [48]	49.0	61.7	24.5
WFDR [49]	62.0	75.1	42.4
SSL [50]	71.7	83.8	37.8
JVTC [51]	79.5	89.2	47.5
AE [43]	63.2	75.4	39.0
MetaCam [39]	83.9	92.3	61.7
HCT [52]	80.0	91.6	56.4
Ours	**87.7**	**93.2**	**81.2**

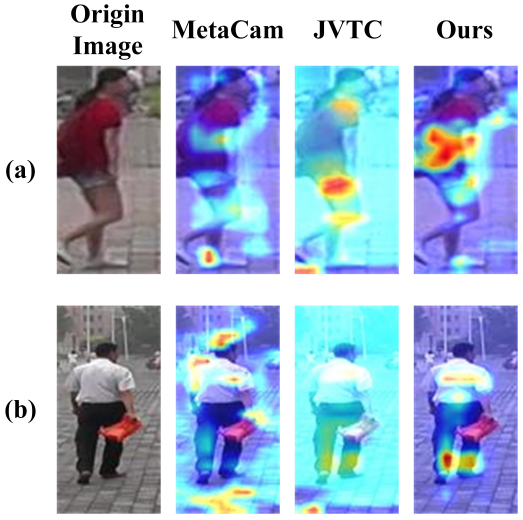


Fig. 9 Visualization of attention maps for features by different methods.

Table 10 Comparison with state-of-the-art unsupervised person Re-ID methods on datasets of Market-1501.

Methods	Rank-1	Rank-5	mAP
UMTS [53]	95.8	-	75.9
AAMI [54]	85.9	91.8	61.3
CAL [55]	95.4	97.9	74.3
FDA-NeT [13]	84.3	92.4	55.5
SAN [56]	93.3	97.1	72.5
FIDI [57]	95.7	-	77.6
DFLNet [58]	93.2	97.6	73.3
EALN [59]	84.4	94.1	57.4
Ours	**96.1**	**97.9**	**87.8**

4.8 Discussion on Supervised Re-ID

The effectiveness of the proposed method has been further validated in the supervised vehicle re-identification task. Specifically, we replaced the real labels of the training set with the pseudo-labels generated through clustering in our proposed method. Subsequently, we compared the performance of this method with existing approaches, and the experimental results are shown in Table 10. In comparison with the top-performing methods, UMTS [53] and CAL [55], our proposed method achieved remarkable improvements of 11.9% and 13.5% in mAP, respectively. This indicates the insensitivity of our proposed method to task variations and its robust generalization capability. Furthermore, it underscores that in a supervised learning context without label noise interference, our proposed method can more effectively capture distinctive identity information.

5 Conclusions

We propose SAM-driven MAE pre-training and background-aware meta-learning for unsupervised vehicle Re-ID. A space-constrained vehicle background segmentation method is presented to obtain high-quality background segmentation results via SAM. To enhance the capacity to distinguish between background information and vehicle identity, we design SAM-driven MAE pre-training to learn identity-sensitive features for downstream unsupervised vehicle Re-ID tasks. For downstream unsupervised vehicle Re-ID tasks, background-aware meta-learning is proposed to enhance the sensitivity of the Re-ID model to varying degrees of background interference by using the background region ratios. Extensive experiments confirm that the proposed method can effectively alleviate the problem of background variations. In our future work, SAM-driven large-scale pre-training that adopts text prompt learning will be further explored and discussed to overcome the complexity of extra pre-training end-to-end.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant No. 62076117 and 62166026, the Jiangxi Key Laboratory of Smart City under Grant No. 20192BCD40002 and the Jiangxi Provincial Natural Science Foundation under Grant No. 20224BAB212011, 20232BAB212008 and 20232BAB202051.

Availability of data and materials

The data presented in this study are available on request from the corresponding author.

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

Authors contributions

Dong Wang: Methodology, Writing-Original Draft, Conceptualization, Implementation. Qi Wang: Funding Acquisition, Project Administration, Writing-Review and Editing. Weidong Min: Funding Acquisition, Project Administration, Supervision. Di Gai: Formal Analysis, Visualization. Qing Han: Visualization, Data Curation. Longfei Li: Validation, Software. Yuhan Geng: Validation, Investigation.

References

[1] Lei J, Qin T, Peng B, Li W, Pan Z, Shen H, Kwong S. Reducing Background Induced Domain Shift for Adaptive Person Re-Identification. *IEEE Transactions on Industrial Informatics*, 2023, 19(6): 7377–7388, doi:10.1109/TII.2022.3210589.

[2] Zhang G, Zhang H, Lin W, Chandran AK, Jing X. Camera Contrast Learning for Unsupervised Person Re-Identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023, 33(8): 4096–4107, doi:10.1109/TCSVT.2023.3240001.

[3] Zhu K, Guo H, Liu S, Wang J, Tang M. Learning Semantics-Consistent Stripes With Self-Refinement for Person Re-Identification. *IEEE Transactions on Neural Networks and Learning Systems*, 2022: 1–12, doi:10.1109/TNNLS.2022.3151487.

[4] Wu M, Zhang Y, Zhang T, Zhang W. Background Segmentation for Vehicle Re-Identification, Berlin, Heidelberg: Springer-Verlag2020, 88–99, doi:10.1007/978-3-030-37734-2_8.

[5] Munir A, Martinel N, Micheloni C. Oriented Splits Network to Distill Background for Vehicle Re-Identification. In *2021 17th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2021, 1–8, doi:10.1109/AVSS52988.2021.9663832.

[6] Lu Z, Lin R, Hu H. MART: Mask-aware reasoning transformer for vehicle re-identification. *IEEE Transactions on Intelligent Transportation Systems*, 2022, 24(2): 1994–2009.

[7] Ning X, Gong K, Li W, Zhang L, Bai X, Tian S. Feature refinement and filter network for person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 2020, 31(9): 3391–3402.

[8] Li Z, Deng Y, Tang Z, Huang J. Sfmnet: Self-guided feature mining network for vehicle re-identification. In *2023 International Joint Conference on Neural Networks (IJCNN)*, IEEE2023, 1–8.

[9] He K, Chen X, Xie S, Li Y, Dollár P, Girshick R. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, 16000–16009.

[10] Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, Xiao T, Whitehead S, Berg AC, Lo WY, et al.. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[11] Lin Y, Wu Y, Yan C, Xu M, Yang Y. Unsupervised person re-identification via cross-camera similarity exploration. *IEEE Transactions on Image Processing*, 2020, 29: 5481–5490.

[12] Wang H, Lu J, Pang F, Zhou J, Zhang K. Bi-directional Style Adaptation Network for Person Re-Identification. *IEEE Sensors Journal*, 2022, 22(12): 12339–12347.

[13] Lou Y, Bai Y, Liu J, Wang S, Duan L. Veri-wild: A large dataset and a new method for vehicle re-identification in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, 3235–3243.

[14] Kamenou E, del Rincón JM, Miller P, Devlin-Hill P. A Meta-Learning Approach for Domain Generalisation Across Visual Modalities in Vehicle Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, 385–393.

[15] Zhang L, Liu Z, Zhang W, Zhang D. Style Uncertainty Based Self-Paced Meta Learning for Generalizable Person Re-Identification. *IEEE Transactions on Image Processing*, 2023, 32: 2107–2119.

[16] Ni H, Song J, Luo X, Zheng F, Li W, Shen HT. Meta distribution alignment for generalizable person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, 2487–2496.

[17] Zheng Z, Ruan T, Wei Y, Yang Y, Mei T. VehicleNet: Learning robust visual representation for vehicle re-identification. *IEEE Transactions on Multimedia*, 2020, 23: 2683–2693.

[18] Yu J, Oh H. Unsupervised vehicle re-identification via self-supervised metric learning using feature dictionary. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE2021, 3806–3813.

[19] Lu Z, Lin R, He Q, Hu H. Mask-aware pseudo label denoising for unsupervised vehicle re-identification. *IEEE Transactions on Intelligent Transportation Systems*, 2023, 24(4): 4333–4347.

[20] He Z, Zhao H, Wang J, Feng W. Multi-Level Progressive Learning for Unsupervised Vehicle Re-identification. *IEEE Transactions on Vehicular Technology*, 2022.

[21] Wang P, Ding C, Tan W, Gong M, Jia K, Tao D. Uncertainty-aware clustering for unsupervised domain adaptive object re-identification. *IEEE Transactions on Multimedia*, 2022.

[22] Dai P, Chen P, Wu Q, Hong X, Ye Q, Tian Q, Lin CW, Ji R. Disentangling task-oriented representations for unsupervised domain adaptation. *IEEE Transactions on Image Processing*, 2021, 31: 1012–1026.

[23] Wei R, Gu J, He S, Jiang W. Transformer-Based Domain-Specific Representation for Unsupervised Domain Adaptive Vehicle Re-Identification. *IEEE Transactions on Intelligent Transportation Systems*, 2022, 24(3): 2935–2946.

[24] Wu C, Lin Z, Cohen S, Bui T, Maji S. Phrasecut: Language-based image segmentation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, 10216–10225.

[25] Yang Z, Wang J, Tang Y, Chen K, Zhao H, Torr PH. Lavt: Language-aware vision transformer for referring image seg-

mentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, 18155–18165.

[26] Xie J, Hou X, Ye K, Shen L. Clims: Cross language image matching for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, 4483–4492.

[27] Wang X, Zhang X, Cao Y, Wang W, Shen C, Huang T. Seggpt: Segmenting everything in context. *arXiv preprint arXiv:2304.03284*, 2023.

[28] Peng J, Jiang G, Chen D, Zhao T, Wang H, Fu X. Eliminating cross-camera bias for vehicle re-identification. *Multimedia Tools and Applications*, 2020: 1–17.

[29] Khorramshahi P, Peri N, Chen Jc, Chellappa R. The devil is in the details: Self-supervised attention for vehicle re-identification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, Springer2020, 369–386.

[30] Zhu X, Luo Z, Fu P, Ji X. VOC-ReID: Vehicle re-identification based on vehicle-orientation-camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, 602–603.

[31] Ester M, Kriegel HP, Sander J, Xu X, et al.. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, 1996, 226–231.

[32] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, et al.. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, PMLR2021, 8748–8763.

[33] He S, Luo H, Wang P, Wang F, Li H, Jiang W. TransReID: Transformer-Based Object Re-Identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, 15013–15022.

[34] Li J, Wang M, Gong X. Transformer Based Multi-Grained Features for Unsupervised Person Re-Identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, 2023, 42–50.

[35] Li M, Li CG, Guo J. Cluster-guided asymmetric contrastive learning for unsupervised person re-identification. *IEEE Transactions on Image Processing*, 2022, 31: 3606–3617.

[36] Han X, Yu X, Li G, Zhao J, Pan G, Ye Q, Jiao J, Han Z. Rethinking sampling strategies for unsupervised person re-identification. *IEEE Transactions on Image Processing*, 2022, 32: 29–42.

[37] Hu Z, Zhu C, He G. Hard-sample guided hybrid contrast learning for unsupervised person re-identification. In *2021 7th IEEE International Conference on Network Intelligence and Digital Content (IC-NIDC)*, IEEE2021, 91–95.

[38] Zhang X, Li D, Wang Z, Wang J, Ding E, Shi JQ, Zhang Z, Wang J. Implicit sample extension for unsupervised person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, 7369–7378.

[39] Yang F, Zhong Z, Luo Z, Cai Y, Lin Y, Li S, Sebe N. Joint noise-tolerant learning and meta camera shift adaptation for unsupervised person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, 4855–4864.

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

[41] Ge Y, Chen D, Li H. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. *arXiv preprint arXiv:2001.01526*, 2020.

[42] Liu X, Zhang S. Graph Consistency Based Mean-Teaching for Unsupervised Domain Adaptive Person Re-Identification. In ZH Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, International Joint Conferences on Artificial Intelligence Organization2021, 874–880, main Track.

[43] Ding Y, Fan H, Xu M, Yang Y. Adaptive Exploration for Unsupervised Person Re-Identification. *TOMM*, 2020, 16(1): 3:1–3:19, doi:10.1145/3369393.

[44] Wang W, Zhao F, Liao S, Shao L. Attentive waveblock: complementarity-enhanced mutual networks for unsupervised domain adaptation in person re-identification and beyond. *IEEE Transactions on Image Processing*, 2022, 31: 1532–1544.

[45] Zheng K, Liu W, He L, Mei T, Luo J, Zha ZJ. Group-aware label transfer for domain adaptive person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, 5310–5319.

[46] Wang D, Zhang S. Unsupervised person re-identification via multi-label classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, 10981–10990.

[47] Lin Y, Dong X, Zheng L, Yan Y, Yang Y. A bottom-up clustering approach to unsupervised person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 2019, 8738–8745.

[48] Zhong Z, Zheng L, Luo Z, Li S, Yang Y. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, 598–607.

[49] Yu HX, Zheng WS. Weakly supervised discriminative feature learning with state information for person identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, 5528–5538.

[50] Lin Y, Xie L, Wu Y, Yan C, Tian Q. Unsupervised person re-identification via softened similarity learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, 3390–3399.

[51] Li J, Zhang S. Joint visual and temporal consistency for unsupervised domain adaptive person re-identification. In *Computer Vision–ECCV 2020: 16th European Conference,*

Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16, Springer2020, 483–499.

[52] Zeng K, Ning M, Wang Y, Guo Y. Hierarchical clustering with hard-batch triplet loss for person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, 13657–13665.

[53] Jin X, Lan C, Zeng W, Chen Z. Uncertainty-aware multi-shot knowledge distillation for image-based object re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 2020, 11165–11172.

[54] Zhou Y, Shao L. Aware attentive multi-view inference for vehicle re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, 6489–6498.

[55] Rao Y, Chen G, Lu J, Zhou J. Counterfactual attention learning for fine-grained visual categorization and re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, 1025–1034.

[56] Jin X, Lan C, Zeng W, Wei G, Chen Z. Semantics-aligned representation learning for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 2020, 11173–11180.

[57] Yan C, Pang G, Bai X, Liu C, Ning X, Gu L, Zhou J. Beyond triplet loss: person re-identification with fine-grained difference-aware pairwise loss. *IEEE Transactions on Multimedia*, 2021, 24: 1665–1677.

[58] Bai Y, Lou Y, Dai Y, Liu J, Chen Z, Duan LY, Pillar I. Disentangled Feature Learning Network for Vehicle Re-Identification. In *IJCAI*, 2020, 474–480.

[59] Lou Y, Bai Y, Liu J, Wang S, Duan LY. Embedding adversarial learning for vehicle re-identification. *IEEE Transactions on Image Processing*, 2019, 28(8): 3794–3807.

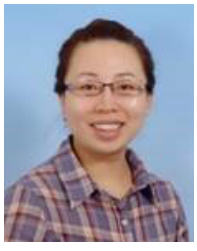
Author biography

Dong Wang received the B.E. degree in software engineering from Nanchang University, China in 2022. He is currently pursuing the M.E. degree at Nanchang University, China. His current research interests include computer vision.

Qi Wang received the M.E. and Ph.D. degrees in School of Information Engineering from Nanchang University, China in 2018 and 2021, respectively. He is currently a lecturer, School of Software, Nanchang University, China. He also is an assistant researcher in Jiangxi Key Laboratory of Smart City, China. His current research focuses on computer vision and deep learning, particularly object re-identification.

Weidong Min Weidong Min received the B.E., M.E. and Ph.D. degrees in computer application from Tsinghua University, China in 1989, 1991 and 1995, respectively. He is currently a Professor, School of Mathematics and Computer Science, and the Dean, Institute of Metaverse, Nanchang University, China. He is the Dean, Jiangxi Key Laboratory of Smart City, China. He is an Executive Director of China Society of Image and Graphics. His current research interests include image and video processing, virtual reality, artificial intelligence, big data, and distributed system.

Di Gai received the M.E. and Ph.D. degrees in College of Computer Science and Technology from Jilin University, China, in 2018 and 2021, respectively. He is currently a lecturer, School of Software, Nanchang University, China. He also is an assistant researcher in Jiangxi Key Laboratory of Smart City, China. His research interests include medical image processing and pattern recognition, especially on image fusion.



Qing Han obtained the B.E. and M.E. degrees of computer application at Tianjin Polytechnic University in China in 1997 and 2006, respectively. She is now an associate professor at School of Mathematics and Computer Science, Nanchang University, China. Her research interests include image and video processing, network management.

Longfei Li received the B.E. degree in software engineering from Jiangxi Normal University, China in 2021. He is currently pursuing the M.E. degree at Nanchang University, China. His current research interests include computer vision.

Yuhan Geng Yuhan Geng received the B.S. degree in Bioinformatics from The Chinese University of Hong Kong, Shenzhen, China in 2023. She is currently pursuing the M.S. degree at University of Michigan, Ann Arbor, United States. Her current research interests include computer vision.