CORE-ReID V2: Advancing the Domain Adaptation for Object Re-Identification with Optimized Training and Ensemble Fusion

Trinh Quoc Nguyen [*] , Oky Dicky Ardiansyah Prima , Syahid Al Irfan , Hindriyanto Dwi Purnomo , Radius Tanone

Article

CORE-ReID V2: Advancing the Domain Adaptation for Object Re-Identification with Optimized Training and Ensemble Fusion

Trinh Quoc Nguyen [1,2,*], **Oky Dicky Ardiansyah Prima** [1], **Syahid Al Irfan** [1,2], **Hindriyanto Dwi Purnomo** [3] **and Radius Tanone** [3]

[1] Graduate School of Software and Information Science, Iwate Prefectural University, Takizawa-shi 020-0693, Iwate, Japan

[2] CyberCore Co., Ltd., Morioka-shi 020-0045, Iwate, Japan

[3] Department of Information Technology, Satya Wacana Christian University, Salatiga 50711, Indonesia

* Correspondence: g236v201@s.iwate-pu.ac.jp

Abstract: This study presents CORE-ReID V2, an enhanced framework building upon CORE-ReID. The new framework extends its predecessor by addressing Unsupervised Domain Adaptation (UDA) challenges in Person ReID and Vehicle ReID, with further applicability to Object ReID. During pre-training, CycleGAN is employed to synthesize diverse data, bridging image characteristic gaps across different domains. In the fine-tuning, an advanced ensemble fusion mechanism, consisting of the Simplified Efficient Channel Attention Block (SECAB) and Efficient Channel Attention Block (ECAB), enhances both local and global feature representations while reducing ambiguity in pseudo-labels for target samples. Experimental results on widely used UDA Person ReID and Vehicle ReID datasets demonstrate that the proposed framework outperforms state-of-the-art methods, achieving top performance in Mean Average Precision (mAP) and Rank-k Accuracy (Top-1, Top-5, Top-10). Moreover, the framework supports lightweight backbones such as ResNet18 and ResNet34, ensuring both scalability and efficiency. Our work not only pushes the boundaries of UDA-based Object ReID but also provides a solid foundation for further research and advancements in this domain.

Keywords: person re-identification; vehicle re-identification; unsupervised learning; visual surveillance; domain adaptation; deep learning

1. Introduction

Object Re-identification (ReID) focuses on retrieving specific object instances across diverse viewpoints [1–5] and has gained significant attention within the computer vision community due to its broad range of practical applications. Substantial advancements have been made in both supervised [6–9] and unsupervised ReID tasks [5,10,11], with most approaches employing backbone models originally designed for generic image classification tasks [12,13].

Unsupervised domain adaptation (UDA) for object ReID aims to transfer knowledge learned from a labeled source domain to accurately measure inter-instance affinities in an unlabeled target domain. Typical ReID tasks, such as person ReID and vehicle ReID, involve source and target domain datasets that do not share identical class identities. State-of-the-art UDA methods [11,14–18] generally adopt a two-stage training paradigm: (1) supervised pre-training on the source domain and (2) unsupervised fine-tuning on the target domain. During the second stage, pseudo-labeling strategies have demonstrated effectiveness in recent works [11,16,17]. These strategies iteratively alternate between generating pseudo-class labels through clustering target-domain instances and refining the network by training on these pseudo-classes. This iterative process enables the pre-trained source-domain network to adapt to the target domain by capturing inter-sample relationships, despite the inherent noise in pseudo-class labels.

The integration of global features, which encapsulate coarse semantic information, and local features, which provide fine-grained details, has proven effective in enhancing algorithm performance. CORE-ReID [11] introduced the Ensemble Fusion framework which combines global and local features with the Efficient Channel Attention Block (ECAB). ECAB leverages inter-channel relationships to guide the model's attention toward salient structures within the input image. Although CORE-ReID achieves competitive results in Person ReID under the UDA setting, it has three main limitations. Firstly, ECAB is exclusively to local features, leaving global features unenhanced. Secondly, CORE-ReID only supports deep and complex backbone networks, such as ResNet50, ResNet101, and ResNet152, while neglecting shallower architectures like ResNet18 and ResNet34, which offer computational efficient and are well-suited for resource-constrained environments. Finally, CORE-ReID is limited to the Person ReID task, restricting its applicability to other ReID scenarios.

In this paper, we present CORE-ReID V2, an enhanced version of CORE-ReID that addresses its limitations and introduces several novel contributions. CORE-ReID V2 not only achieves superior performance in UDA Person ReID but also extends its applicability to Object ReID tasks and supports lightweight backbone networks, such as ResNet18 and ResNet34, making it suitable for real-time systems and mobile devices. Building on the principles of LF2 [17] and CORE-ReID, we design a mean-teacher-based framework that iteratively learns multi-view features and refines noisy pseudo-labels through multiple clustering steps. We introduce the Ensemble Fusion++ module in CORE-ReID V2, which adaptively enhances both local and global features. This module applies ECAB to local features and the Simplified Efficient Channel Attention Block (SECAB) to global features, resulting in a fused representation that provides a more comprehensive feature set. Furthermore, we improve clustering outcomes by incorporating the KMeans++ [19] initialization strategy, which balances randomness and centroid selection to enhance cluster quality. To validate the framework, we pre-train the model on a source domain that integrates camera-aware style-transferred data for Person ReID and domain-aware style-transferred data for Vehicle ReID. Additionally, we adopt a teacher-student architecture for iterative domain adaptation, where the teacher network captures global features while the student network refines diverse local features, both contributing to a more effective pseudo-labeling process. To summarize, the key contributions of CORE-ReID V2 are as follows:

- **Advanced Data Augmentation Techniques:** The framework integrates novel data augmentation strategies, such as Local Grayscale Patch Replacement and Random Image-to-Grayscale Conversion for UDA task. These methods introduce diversity in the training data, enhancing the model's stability.
- **Dynamic and Flexible Backbone Support:** CORE-ReID V2 extends compatibility to smaller backbone architectures, including ResNet18 and ResNet34, without compromising performance. This flexibility allows for deployment in resource-constrained environments while maintaining high accuracy.
- **Expansion to Vehicle and further Object ReID:** Unlike its predecessor, which focused solely on person re-identification, CORE-ReID V2 extends its scope to Vehicle Re-identification and further general Object Re-identification. This expansion demonstrates its versatility and adaptability across various domains.
- **Introduction of Ensemble Fusion++:** The framework incorporates the SECAB into the global feature extraction pipeline to enhance feature representation by dynamically emphasizing informative channels, thereby improving discrimination between instances.

2. Related Work

Extensive research has been conducted on UDA for Object ReID and knowledge transfer techniques, such as knowledge distillation, which enable well-trained models to transfer expertise and improve learning complex domain scenarios. Methods generally fall into two categories: domain translation, which aligns visual styles between domains, and pseudo-labeling, which iteratively cluster target samples to generate pseudo-labels. While pseudo-labeling has demonstrated superior

performance, both methods face challenges related to domain shifts and noisy labels. Moreover, the fusion of global and local features has proven effective in various tasks, including classification, object detection, and semantic segmentation, by integrating contextual information with fine-grained details. Building on these insights, our work refines the CORE-ReID [11] fusion module to achieve a better balance between global and local features, leading to improved performance across multiple ReID tasks, including Person and Vehicle ReID.

2.1. UDA for Object ReID

UDA has gained significant attention for its ability to reduce reliance on costly manual annotations. By utilizing labeled data from a source domain, UDA enhances model performance in a target domain without requiring target-specific annotations. Research in Object ReID has primarily concentrated on Person ReID and Vehicle ReID [20]. Existing UDA approaches for ReID can be broadly grouped into two categories: domain translation-based methods and pseudo-label-based methods [21,22].

Domain translation-based methods: These methods align the visual style of labeled source domain images with that of the target domain. The translated images, along with their original ground-truth labels, are then used for training [23].

Several methods attempt to map source and target distributions to mitigate domain shifts [24–28]. Saenko et al. [24] introduced a domain adaptation technique based on cross-domain transformations by learning a regularized non-linear transformation that brings source domain points closer to the target domain. In [25], the Geodesic Flow Kernel (GFK) was proposed to address domain shifts by integrating an infinite number of subspaces that capture geometric and statistical changes between the source and target domains. Similarly, Fernando et al. [26] developed a mapping function to align the source subspace with the target subspace for improved adaptation. Correlation Alignment (CORAL) [27] addressed domain shifts by computing the covariance statistics of each domain and applying a whitening and re-coloring linear transformation to align the source feature with the target domain. The Disentanglement Then Reconstruction (DTR) framework [28] enhanced alignment by disentangling the distributions and reconstructing them to ensure consistency across domains.

Another line of research [29–32] adopts adversarial approaches to learn transformations in the pixel space between domains. Methods like PixelDA [29], PTGAN [30], and SBSGAN [31] enforce pixel-level constraints to preserve color consistency during domain translation. CoGAN [32] extends this concept by learning joint distributions, such as the joint distribution of color and depth images or face images with varying attributes.

Other methods focus on discover a domain-invariant feature space to bridge domain gaps [22,33–38]. SPGAN [22] and CGAN-TM [35] improve feature-level similarity between translated and original images. Deep Adaption Network (DAN) [37] employes the Maximum Mean Discrepancy (MMD) [39–41] to align feature distributions across domains. Similarly, Ganin et al. [33] and Ajakan et al. [42] introduced a domain confusion loss to encourage the learning of domain-invariant features. Hoffman et al. [36] proposed the Intermediate Domain Module (IDM) to generate intermediate domain representations dynamically by mixing the hidden features of the source and target domains through two domain features. CyCADA [38] combines both pixel-level and feature-level adaptation to improve domain adaptation.

Pseudo-label-based methods: The second category, pseudo-labeling methods [11,14–17,43–46], models the relationships between unlabeled target-domain data with generated pseudo labels. Fan et al. [43] proposed the progressive unsupervised learning (PUL) method that alternates between assigning labels to unlabeled samples and optimizing the network using the generated targets. This iterative refinement aligns the model's representations more closely with the target domain, enhancing adaptation over time. Lin et al. [44] developed a bottom-up clustering framework enhanced by a repelled loss mechanism, which aims to increase the discriminative power of learned features while mitigating intra-cluster variations. Similarly, UDAP [14] proposed a self-training scheme that minimizes loss functions iteratively using clustering-based pseudo labels. SSG [16], LF^2

[17], and CORE-ReID [11] further contributed to this category, which introduce techniques to assign pseudo labels to both global and local features. Ge et al. [15] proposed Mutual Mean Teaching (MMT), which combines offline hard pseudo labels and online soft pseudo labels in an alternating training process, enhancing the model's ability to adapt to domain shifts. This technique improves the model's capacity to handle domain shifts by iteratively refining both the pseudo labels and feature representations throughout training. SpCL [45] advanced this field by using a hybrid memory module that stored centroids of labeled source domain images alongside un-clustered target instances and target domain clusters. This hybrid memory provides additional supervision to the feature extractor, while minimizing a unified contrastive loss over the three types of stored information. Additionally, Zheng et al. [46] developed the Uncertainty-Guided Noise Resilient Network (UNRN), which evaluates the reliability of predicted pseudo labels for target domain samples. By incorporating uncertainty estimates into the training process, UNRN improves performance with noisy annotations, thereby enhancing performance in domain adaptation scenarios.

Pseudo-labeling methods analyze data at different levels of detail, allowing them to capture small differences within the target domain. As a result, they achieve better performance than domain translation-based approaches and continue to lead in accuracy across most public datasets [11,15,45,46].

2.2. Knowledge Transfer

Knowledge transfer, the process of passing knowledge from a well-trained neural network (referred to as the teacher model) to another model (the student model), has gained considerable attention in recent years [47–50]. The core idea is to provide consistent training supervision for both labeled and unlabeled data through predictions generated by various models. Knowledge transfer techniques help student networks become more accurate and generalize better, as the teacher model's output implicitly contains rich information about the relationships between training samples and their underlying distribution [51]. In this way, knowledge transfer serves as a regularizer, enhancing the student model's performance. For example, Laine and Aila [52] introduced the Mean Teacher model which averaged model weights across multiple training iterations to guide supervision for unlabeled data. In contrast, Deep Mutual Learning (DML) [53], proposed by Zhang et al., shifts from the traditional teacher-student framework by employing a group of student models that train collaboratively, providing mutual supervision and facilitating the exploration of diverse feature representations. Ge et al. introduced MMT [15], which adopts an alternative training method that uses both offline refined hard pseudo-labels and online refined soft pseudo-labels. MEB-Net [54] further builds on this by using three networks (six models in total) to conduct mutual mean teacher training and generate pseudo-labels.

2.3. Feature Fusion

The feature fusion of global and local features has proven highly effective across various computer vision tasks, including classification [55–58], object detection [59–63], semantic segmentation [64–67], and more [68].

In image classification, global features capture the overall structure and appearance, while local features focus on fine-grained details. Combining both types provides complementary information, enhancing the model's ability to generalize across variations such as pose, lighting, and occlusions. For example, Tian et al. [55] proposed a vehicle model recognition system using an iterative discrimination CNN based on selective multi-convolutional region feature extraction. Their SMCR model combines global and local features to boost classification accuracy. Similarly, Lu et al. [56] introduced a script identification framework that leverages both global CNNs, trained on segmented images, and local CNNs, trained on image patches. He et al. [57] presented a traffic sign recognition approach that integrates global and local features using histograms of oriented gradients (HOG), color histograms, and edge features. Suh et al. [58] employed fusion layers to concatenate global and local features for shipping label image classification, improving image quality verification. These

studies demonstrated that fusing global and local features consistently improves classification performance compared to models relying on whole-image analysis alone.

In object detection, global features provide spatial awareness of objects within a scene, while local features capture subtle patterns, such as textures and edges, which are essential for accurate detection under occlusions. Cong et al. [59] proposed an end-to-end co-salient object detection network that uses collaborative learning to enhance inter-image relationships. Their model includes a global correspondence module to extract interactive information across images and a local correspondence module to capture pairwise relationships. Li et al. [60] developed an anchor-free object detector that leverages a global-local feature extraction transformer (GLFT) to capture semantic information from both micro- and macro-level perspectives.

In semantic segmentation, the fusion of global and local features improves pixel-level predictions by combining overall scene context with localized information, especially in complex environments. Yang et al. [64] introduced AFNet, which uses a multi-path encoder to extract diverse features, a multi-path attention fusion module, and a fine-grained attention fusion module to combine high-level abstract and low-level spatial features. Tian et al [65] extended this concept with two encoders to extract both global high-order interactive features and local low-order features. These encoders form the backbone of the global and local feature fusion network (GLFFNet), enabling effective segmentation of remote sensing images through a dual-encoder structure. Later, Zhou et al [66] proposed a local–global multi-scale fusion network (LGMFNet) for building segmentation in SAR images. LGMFNet includes a dual encoder-decoder structure, with a transformer-based auxiliary encoder complementing the CNN-based primary encoder. The global–local semantic aggregation module (GLSM) is also introduced to bridge the two encoders, enabling semantic guidance across multiple scales through a specialized fusion decoder.

Inspired by these advances, feature fusion techniques have gained traction in domain adaptation for Object Re-identification. Self-Similarity Grouping (SSG) [16] is the first approach to applied both global and local features for unsupervised domain adaptation (UDA) in Person ReID. However, SSG faces two challenges: first, using a single network for feature extraction often introduces noisy pseudo-labels, and second, it performs clustering independently on global and local features, potentially assigning multiple inconsistent pseudo-labels to the same sample. To address these limitations, LF^2 [17] was proposed to fuse global and local features into a unified representation, reducing noise and improving clustering consistency. Building on this idea, CORE-ReID [11] introduced an Ensemble Fusion module equipped with the ECAB, which effectively fuses global and local features.

3. Materials and Methods

Despite advancements in domain translation-based methods, these often suffer from a persistent domain gap between translated images and real target domain images, which can adversely impact performance. To address this issue, our approach employs a pseudo-labeling strategy, which enables data analysis at multiple levels of granularity. This method has demonstrated superior performance compared to domain translation-based techniques [11,15,45,46].

While existing pseudo-labeling frameworks such as Deep Mutual Learning (DML) [53], MMT [15], and MEB-Net [54] have proven effective, they suffer from limitations due to their heavy reliance on pseudo-labels generated by the teacher model. These pseudo-labels can be noisy or inaccurate, adversely affecting model training. To mitigate this issue, we utilize a teacher-student network paradigm, where the student network is trained on labeled source domain data, and the teacher network is iteratively refined using the Mean Teacher method. Furthermore, we incorporate the Ensemble Fusion++ module, which enhances feature extraction by adaptively refining both local and global representations, thereby it is expected to produce more stable and reliable pseudo-labels than existing approaches.

In CORE-ReID [11], the Efficient Channel Attention Block (ECAB) was primarily applied to local features, restricting the full potential of the Ensemble Fusion module. In CORE-ReID V2, we extend

and enhance this module to ensure that both global and local features undergo comprehensive optimization. This enhancement results in a more balanced and discriminative feature representation, improving generalization across diverse ReID tasks. Additionally, the improved Ensemble Fusion++ is not only effective for Person ReID but also demonstrates strong domain adaptation capabilities in Vehicle ReID, further validating its versatility in Object ReID.

This chapter outlines the methodology and materials used in CORE-ReID V2 for unsupervised domain adaptation (UDA) in Object ReID. The proposed framework consists of two main stages: (1) pre-training on a labeled source domain and (2) fine-tuning on an unlabeled target domain.

3.1. Overview

3.1.1. CORE-ReID V1 and CORE-ReID V2

CORE-ReID V1: A Baseline for Unsupervised Domain Adaptation in Person Re-identification: CORE-ReID V1 was introduced as a framework to address Unsupervised Domain Adaptation (UDA) in Person Re-identification (ReID). It effectively tackled domain shifts between camera views by leveraging Camera-Aware Style Transfer for synthetic data generation, Random Grayscale Patch Replacement for data augmentation, and K-Means Clustering for pseudo-labeling. Additionally, the Ensemble Fusion module with Efficient Channel Attention Block (ECAB) played a crucial role in integrating local and global features, improving the model's performance in cross-domain scenarios.

Despite its success, CORE-ReID V1 had several limitations:

1. Limited Application Domain: The framework was specifically designed for Person ReID, restricting its applicability to other ReID tasks such as Vehicle ReID and Object ReID.
2. Synthetic Data Generation Challenge: The Camera-Aware Style Transfer method relied on predefined camera information, making it ineffective when the number of cameras was unspecified.
3. Inefficient Data Augmentation: The Random Grayscale Patch Replacement technique only operated locally, limiting its effectiveness in learning color-invariant features.
4. Clustering Limitations: The K-Means clustering used random centroid initialization, leading to poor centroid placement, slow convergence, high variance in clustering results, and imbalanced cluster sizes.
5. Feature Fusion Issue: The ECAB module enhanced only local features, neglecting improvements to global representations.
6. Restricted Backbone Support: The framework exclusively supported deep networks such as ResNet50, ResNet101, and ResNet152, making it computationally expensive and unsuitable for lightweight applications.

CORE-ReID V2: Expanding Scope, Enhancing Performance: To overcome these limitations, CORE-ReID V2 is proposed as a enhancement over CORE-ReID V1, expanding its capabilities to Vehicle ReID and Object ReID while introducing architectural and methodological improvements.

1. Expanded Application Scope: Unlike CORE-ReID V1, which was restricted to Person ReID, CORE-ReID V2 extends its applicability to Vehicle ReID and Object ReID, making it a versatile framework for various ReID tasks.
2. Advanced Synthetic Data Generation: CORE-ReID V2 incorporates both Camera-Aware Style Transfer and Domain-Aware Style Transfer, allowing effective synthetic data generation even when the number of cameras is unknown.
3. Improved Data Augmentation: A new grayscale patch replacement strategy considers both local grayscale transformation and global grayscale conversion, leading to better feature generalization across domains.
4. Enhanced Clustering with Greedy K-Means++: Instead of relying on random initialization, CORE-ReID V2 employs Greedy K-Means++, which selects optimized centroids to improve cluster spread; minimizes redundancy, requiring fewer iterations; enhances stability and

consistency, reducing randomness; ensures better centroid distribution, leading to improved clustering performance.

5. Ensemble Fusion++ for Comprehensive Feature Enhancement: CORE-ReID V2 introduces Ensemble Fusion++, which integrates both ECAB and SECAB, ensuring that global features are enhanced alongside local features, leading to a more balanced and comprehensive feature representation.

6. Flexible Backbone Support: CORE-ReID V2 broadens its applicability by supporting lightweight networks such as ResNet18 and ResNet34, alongside ResNet50, ResNet101, and ResNet152. This allows deployment in computationally constrained environments, such as real-time and edge-based applications.

CORE-ReID V2 represents a substantial advancement over CORE-ReID V1 by expanding its scope beyond Person ReID, improving clustering stability, introducing adaptive feature enhancement mechanisms, and supporting lightweight architectures. Table 1 shows the summary of these improvements.

Table 1. Summary the main advancement of CORE-ReID V2 over CORE-ReID V1.

Category	CORE-ReID V1		CORE-ReID V2
	Current Status	**Drawbacks/ Issues**	
Applied Domain	Person ReID	Only support Person ReID.	Expansion from Person ReID to Vehicle ReID and further Object ReID.
Synthetic Data Generation	Camera-Aware Style Transfer	Do not work in case the number of cameras is not specified.	Camera-Aware Style Transfer and Domain-Aware Style Transfer (for the case the number of cameras is not specified).
Data Augmentation	Random gray scale patch replacement	Only replace random gray scale patch in the image locally.	Locally gray scale patch replacement and global gray scale conversion.
K-Means Clustering	Random initialization	Problems from random initialization (1) Poor centroid placement (2) Slow convergence (3) Stuck in local minima (4) High variance in results (5) Imbalanced cluster sizes	Greedy K-Means++ initialization helps: (1) Selects centroids with optimized spread (2) Minimizes redundancy, requiring fewer iterations (3) Improves initialization stability (4) Reduces randomness and provides consistent clusters (5) Ensures better centroid distribution
Ensemble Fusion	Ensemble Fusion with ECAB	Only the local features are enhanced in the Ensemble Fusion.	Ensemble Fusion++ (with ECAB and SECAB) helps enhance both local and global features.
Supported Backbones	ResNet50, 101, 152	Do not support small backbones such as ResNet18, 34.	ResNet18, 34, 50, 101, 152

3.1.2. Problem Definition and Methodology

Problem definition: We represent the customized labeled source domain data as $\mathbb{D}_S = \{(x_{S,i}, y_{S,i})|_{i=1}^{N_S}\}$, where $x_{S,i}$ and $y_{S,i}$ denote the ith source image and its corresponding ground truth identity label, respectively, and N_S is the total number of source images. Similarly, the unlabeled target domain data is denoted as $\mathbb{D}_T = \{x_{T,i}|_{i=1}^{N_T}\}$, where $x_{T,i}$ indicates the ith target image, and N_T is the number of target images. Identity labels are unavailable for the images in the target domain dataset, and it is important to note that the identities across the source and target domains do not overlap. The objective of Unsupervised Domain Adaptation (UDA) for Object ReID is to transfer

knowledge from the source domain S to the target domain T. To accomplish this, we propose the CORE-ReID V2 framework, designed to achieve effective knowledge transfer through a pseudo-label-based method.

Methodology: we adopt a pseudo-label-based approach by dividing the process into two stages: pre-training the model on the source domain using a fully supervised strategy, followed by fine-tuning it on the target domain through an unsupervised learning approach (Figure 1).
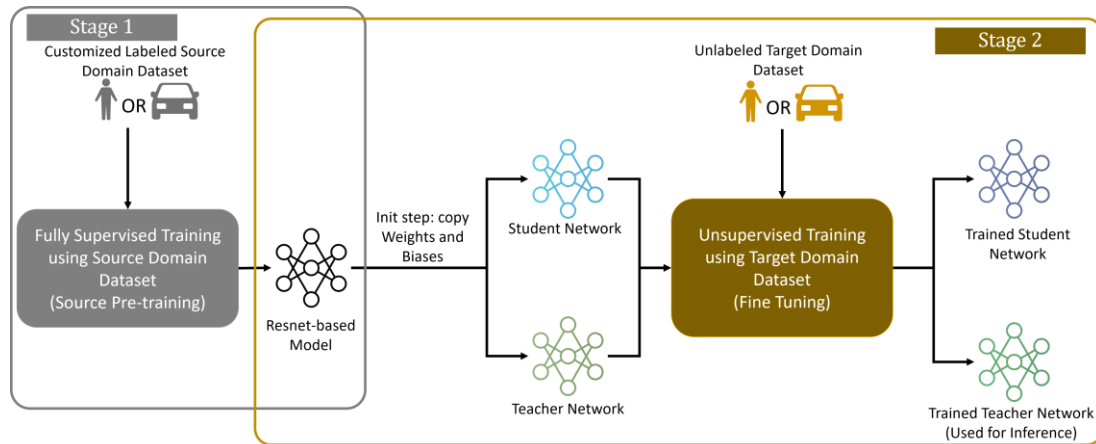

Figure 1. The overall method proposed in this study. First, the model is trained on a customized labeled source domain dataset, after which the parameters of the pre-trained model are transferred to both the student and teacher networks as an initialization step for the next stage. During fine-tuning, the student model is trained, and the teacher model is updated through the Mean Teacher method. To optimize computational efficiency, only the teacher model is employed for inference.

Depending on the specific task (Person ReID or Vehicle ReID), the appropriate dataset is utilized. Our method uses a pair of teacher–student networks. After training the model on a customized labeled source domain dataset, the parameters of the pre-trained model are copied to both the student and teacher networks as an initialization step for the fine-tuning stage. During fine-tuning, we first train the student model and then optimize the teacher model using the Mean Teacher method [69]. This is because averaging model weights across multiple training steps generally yields a more accurate model than relying solely on the final weights [70]. Following the Mean Teacher method, the teacher model uses Exponential Moving Average (EMA) weight parameters of the student model instead of directly sharing weights. This approach allows the teacher network to aggregate information after every step, rather than every epoch, improving consistency. To minimize computational costs, only the teacher model is used during inference.

3.2. Source-Domain Pre-Training

3.2.1. Image-to-Image Translation

Inheriting from CORE-ReID, we employ CycleGAN to generate additional training samples by treating the stylistic variations across different cameras as distinct domains for Person ReID task. This involves training image-to-image translation models using CycleGAN for images captured from various camera views within the dataset. Our goal is to train on a source domain S and evaluate the algorithm during the fine-tuning phase on a different target domain T. By incorporating test data into the training set, similar to DGNet++ [71], we can fully leverage the available data in S. In the Person ReID task, for a source domain dataset containing images from C different cameras, we utilize $C(C-1)$ generative models to produce data in both $X \rightarrow Y$ and $Y \rightarrow X$ directions. The final training set is a combination of the original real images and the style-transferred images from both

the training and test sets within the source domain dataset. These style-transferred images retain the labels of their corresponding real images.

In the case of Vehicle ReID, due to the simpler nature of vehicle features and the large number of cameras used (some datasets do not provide the number of cameras used), we adopt domain-aware transfer models instead of camera-aware models. As a result, only a single transfer model is needed to generate style-transferred images from the source domain to the target domain (Figure 2).
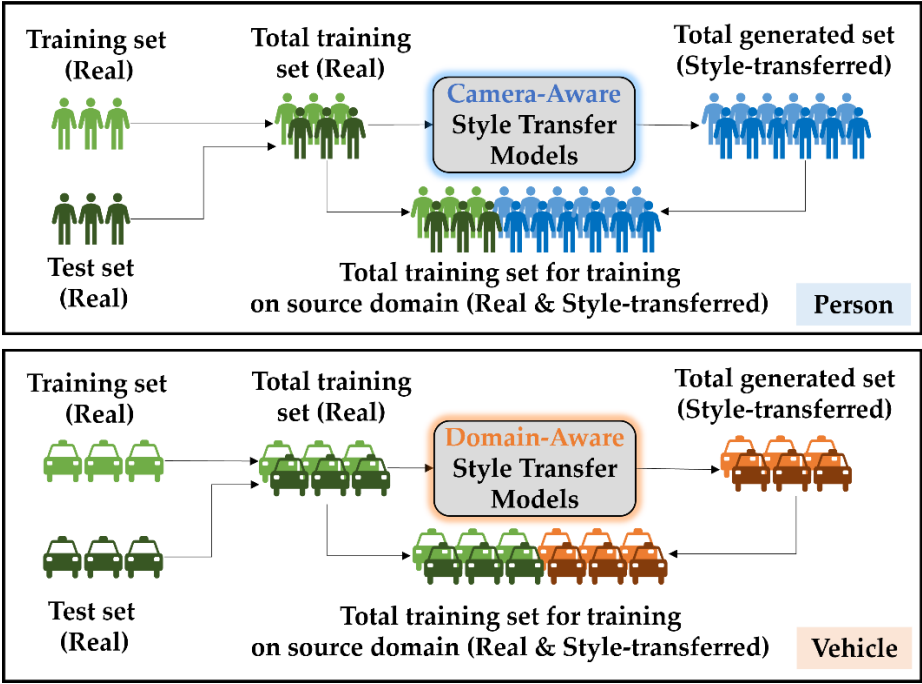
Figure 2. Our process for creating a complete training set for the source domain is as follows: in Person ReID task, we first combine the training set (represented by green icons) and the test set (represented by dark green icons) from the source dataset to create a comprehensive set of real images. This combined set is then used to train a camera-aware style transfer model, which generates style-transferred images (blue icons for the training set and dark blue icons for the test set) that reflect the stylistic characteristics of the target cameras. The final training set for the source domain is formed by merging the real images (green and dark green icons) with the style-transferred images (blue and dark blue icons). For Vehicle ReID, due to the simpler nature of vehicle features and the extensive number of cameras involved (with some datasets not specifying the number of cameras), we use domain-aware transfer models instead of camera-aware models. These models generate style-transferred images (orange icons for the training set and dark orange icons for the test set) that capture the target domain's style. The final source domain training set is then constructed by integrating the real images (green and dark green icons) with the style-transferred images (orange and dark orange icons).

Figure 3 illustrates two representative examples from both the training and test sets in the Market-1501 and CUHK03 datasets, where image styles have been modified according to camera views. This adjustment showcases our approach to data augmentation, where images are transformed to mimic the visual characteristics associated with each camera's unique viewpoint and color distribution. By aligning image styles with camera perspectives, this method effectively reduces the inconsistencies in appearance caused by differences in lighting, angles, and color shifts across camera views. This approach helps the model generalize better, thus mitigating overfitting in Convolutional Neural Networks (CNNs). Furthermore, incorporating camera-specific style information allows the model to learn more robust pedestrian features that are less sensitive to variations across different camera setups, leading to enhanced performance in ReID tasks.
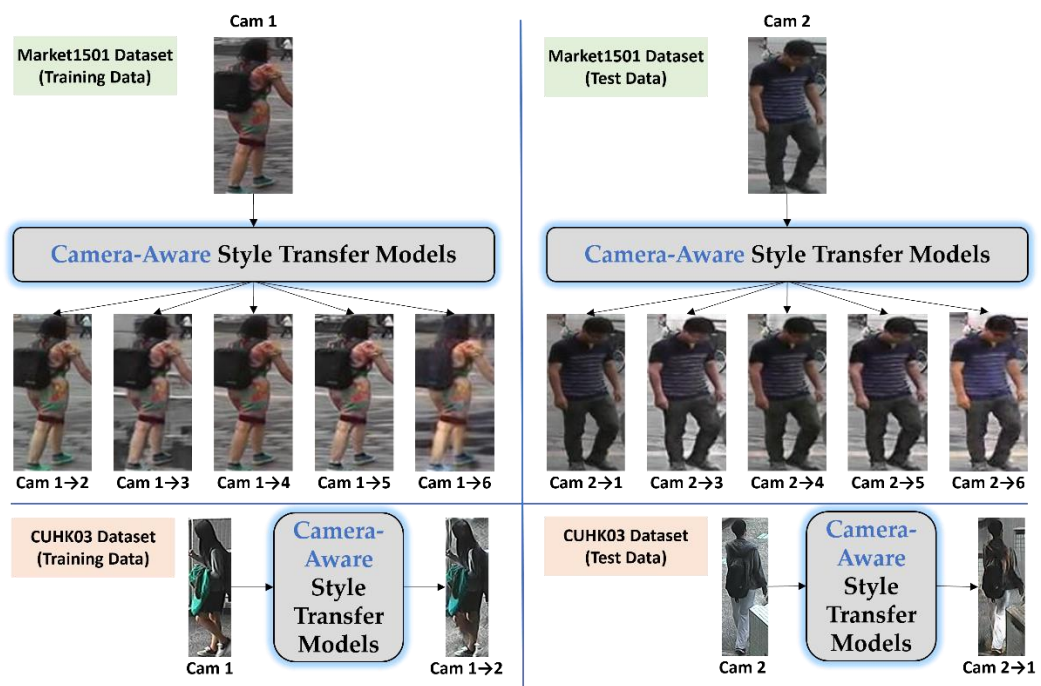
Figure 3. The camera-aware style-transferred samples from Market-1501 and CUHK03 datasets. Each original image, captured by a specific camera, has been transformed to match the styles of the other five cameras in Market-1501 and one camera in CUHK03, covering both training and test data. By applying the style transfer models, these transformations produce style-transferred images as outputs based on the real input images.

Given the simpler nature of vehicle features and the large number of cameras involved (with some datasets not specifying the number of cameras), we aim to utilize labels from the source domain along with target-domain-style-transferred images to create a shared-features domain dataset. This approach involves developing domain-aware style transfer models to bridge the feature gap between the two domains by transforming source domain images into target-domain-style outputs. Figure 4 shows two examples of input images from the VeRi-776 [1] and VehicleID [72] datasets, with the pixel differences between the input and output images highlighted in pink.
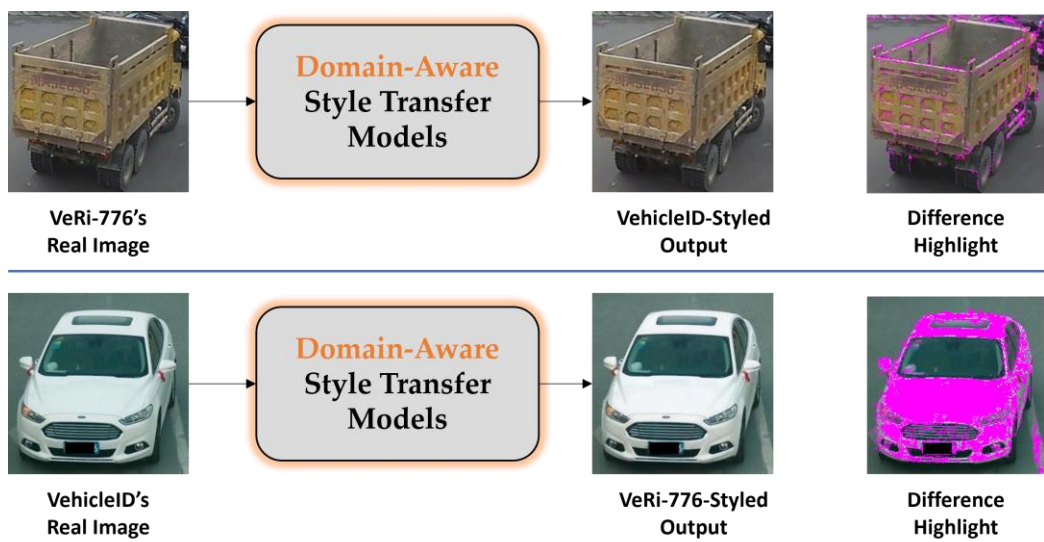

Figure 4. The domain-aware style-transferred samples from VeRi-776 [1] and VehicleID [72] datasets. Each original real image from the source domain dataset has been transformed to match the styles of the target domain dataset. By applying domain-aware style transfer models, the model trained in the pre−training stage will be able to capture the style and features of the target domain.

3.2.2. Fully Supervised Pre-Training

Like many existing UDA approaches [17] that rely on a model pre-trained on a source dataset, we employ a ResNet-based model, pre-trained on ImageNet as the backbone network. In this setup, the original final fully connected (FC) layer is removed and replaced with two new layers. The first one is a batch normalization layer with either 2,048 or 512 features, depending on the specific ResNet architecture. The second layer is an FC layer with M_S dimensions, where M_S represents the number of identities (classes) in the source dataset S (Figure 5).
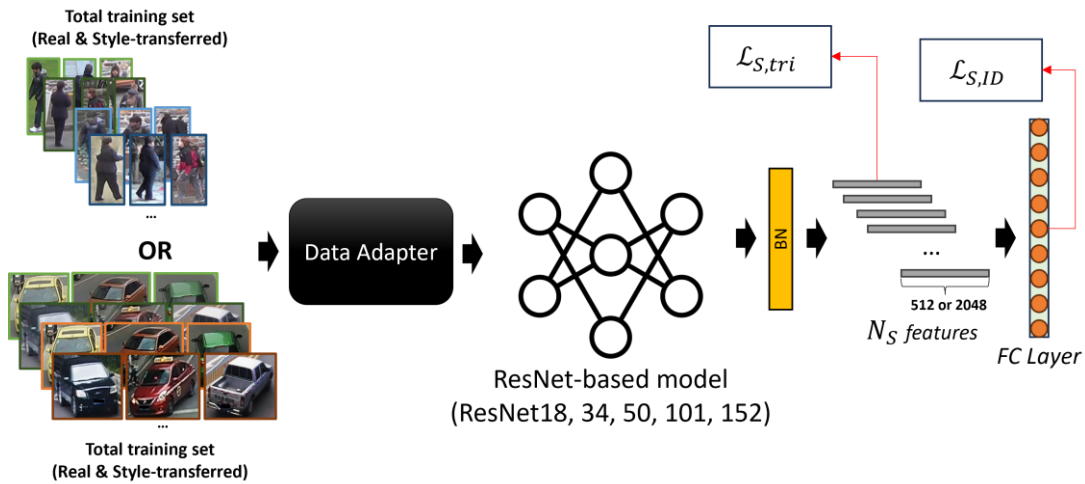

Figure 5. The comprehensive training process employed during the fully supervised pre-training stage. A ResNet-based model, adaptable to various backbone sizes (from ResNet18 and 34 to ResNet50, 101, and 150), serves as the backbone architecture within our training framework.

In our training process, we define the number of identities for the full training set in the source domain as:

$$M_{S,train} = M_{S,train}^{original} + M_{S,test}^{original}, \tag{1}$$

where $M_{S,train}^{original}$ and $M_{S,test}^{original}$ represent the number of identities in the original training and test sets of S, respectively. For each labeled image $x_{S,i}$ and its ground truth identity $y_{S,i}$ in the source domain data $\mathbb{D}_S = \{(x_{S,i}, y_{S,i})|_{i=1}^{N_S}\}$ with N_S representing the total number of images, we train the model using both identity classification (cross-entropy) loss $\mathcal{L}_{S,ID}$ and triplet loss $\mathcal{L}_{S,triplet}$. The identity classification loss is applied to the final fully connected (FC) layer, handling the task as a classification problem, while the triplet loss, applied after batch normalization, is used for feature verification. The loss functions are defined as follows:

$$\mathcal{L}_{S,ID} = \frac{1}{N_S}\sum_{i=1}^{N_S}\mathcal{L}_{ce}(C_S(f(x_{S,i}), y_{S,i}), \tag{2}$$

$$\mathcal{L}_{S,triplet} = \frac{1}{N_S}\sum_{i=1}^{N_S}\max(0, ||f(x_{S,i}) - f(x_{S,i}^+)||_2 - ||f(x_{S,i}) - f(x_{S,i}^-)||_2 + m), \tag{3}$$

where $f(x_{S,i})$ is the feature of the source image $x_{S,i}$, $\mathcal{L}_{ce}$ is the cross-entropy loss, C_S is a learnable classifier in the source domain: $f(x_{S,i}) \rightarrow \{1,2,\dots,M_S\}$. $||\cdot||_2$ denotes the L_2-norm distance, $x_{S,i}^+$ and $x_{S,i}^-$ are the hardest positive and hardest negative feature indices in each mini-batch for the sample $x_{S,i}$, and m represents the triplet distance margin. Using a balance parameter κ, the total loss for source-domain pre-training is:

$$\mathcal{L}_{S,total} = \mathcal{L}_{S,ReID} = \mathcal{L}_{S,ID} + \kappa\mathcal{L}_{S,triplet}. \tag{4}$$

The model is expected to achieve strong performance on fully labeled source-domain data, but its performance significantly drops when applied directly to the unlabeled target domain. Before feeding images into the network, we utilize the "Data Adapter" component (Figure 6) to preprocess them by resizing to a specific size depending on the type of object. We then apply several data augmentation techniques, including edge padding, random cropping, and random horizontal flipping. To address color deviation, we incorporate random color dropout through global and local grayscale transformations [73], preserving key information while minimizing overfitting and enhancing the model's generalization. These approaches specifically balance the model's weighting of color features and color-independent features, resulting in improved feature robustness in the neural network.
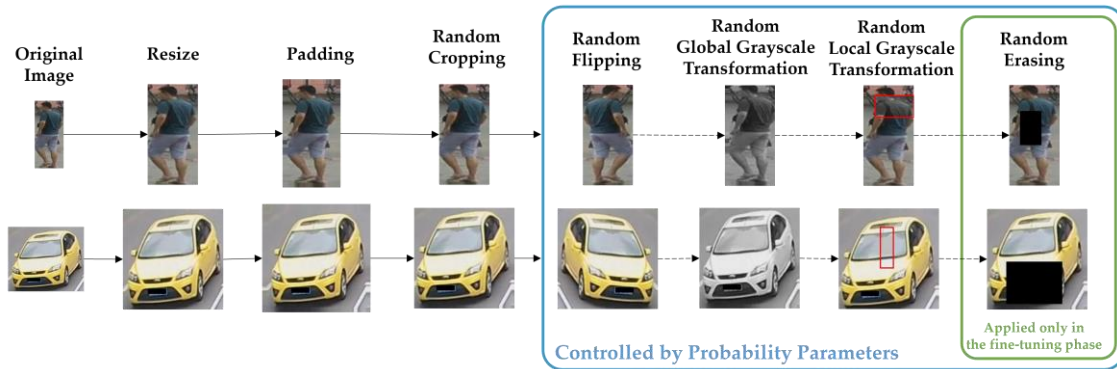

Figure 6. Data adapter component. The transformations of random flipping, random global grayscale, random local grayscale and random erasing will be controlled by probability parameters. In addition, random erasing is only applied in the fine-tuning stage.

We employ the global grayscale transformation to a training batch with a set probability p_{global}, then feed it into the model for training. This process is defined as: $I^* = RGBToGrayscale(I)$, where $RGBToGrayscale()$ represents the grayscale conversion function using the NTSC formula ($0.299 \times Red + 0.587 \times Green + 0.114 \times Blue$), I denotes the input image and I^* is the randomly grayscale image. This function operates by performing pixel-wise accumulation on the red, green, and blue channels of the original RGB image, resulting in a grayscale output. Importantly, the labels remain consistent between the converted grayscale image and the original. The procedure for Local Grayscale Transformation is outlined in Algorithm 1.

Algorithm 1: Global Grayscale Transformation

Input: Input image I;
       Grayscale transformation probability p_{global}.
Output: Randomly grayscale image I^*.
Initialization: $p_t \coloneqq Rand\ (0,1)$.
  1: **if** $p_t \geq p_{global}$ **then**
  2:    $I^* \coloneqq I$.
  3: **else**
  4:    $I^* \coloneqq RGBToGrayscale(I)$.
  5: **return** I^*.
  6: **end**

To enhance model adaptability to significant biases from localized color dropout, we apply a local grayscale transformation to each visible image I in the training batch using the following equation:

$$I^*_{position} = RGBToGrayscale\left(I_{position}\right) = RGBToGrayscale(RandPosition(I)), \tag{5}$$

where $RandPosition()$ generates a random rectangular region within the image I, The transformed sample is represented by I^*. During model training, local grayscale transformation is applied randomly to images in each batch with a probability p_{local}. This involves selecting a random rectangular region within the image and replacing it with the grayscale pixels of that same region. Consequently, images with mixed grayscale levels are generated, aiding the model in learning with color-variant features without altering object structure. The process includes several parameters: s_{min} and s_{max} define the minimum and maximum size ratios of the rectangle relative to the full image area; the rectangle's area S_t is computed by sampling from $S_t \leftarrow Rand(s_{min}, s_{max}) \times S$, where S is the input image area; r_t is a coefficient that sets the rectangle's shape ratio within the interval $(r_{local}, 1/r_{local})$; coordinates x_t and y_t for the rectangle's top-left corner are generated randomly. If the rectangle exceeds image boundaries, new coordinates and dimensions are selected. This approach produces images with grayscale sections that vary in intensity without impacting the core structure, allowing the model to learn features invariant to color variations. The full procedure for local grayscale transformation is detailed in Algorithm 2.

Algorithm 2: Local Grayscale Transformation

Input: Input image I;
   Grayscale transformation probability p_{local};
   Area ratio range (low to high) s_{min} and s_{max};
   Aspect ratio r_{local}.

Output: Randomly transformed image I^*.

Initialization: $p_t \coloneqq Rand(0,1)$;
   $W \coloneqq I.size[0], \ H \coloneqq I.size[1]$;
   $S \coloneqq W * H$.

1: **if** $p_t \geq p_{local}$ **then**
2:   $I^* \coloneqq I$;
    return I^*.
3: **else**
4:   **while** $True$ **do**
5:     $S_t \coloneqq Rand(s_{min}, s_{max}) \times S$;
6:     $r_t \coloneqq Rand(r_{local}, 1/r_{local})$;
7:     $W_t \coloneqq \sqrt{s_t/r_t}, \ H_t \coloneqq \sqrt{S_t \times r_t}$;
8:     $x_t \coloneqq Rand(0, W), \ y_t \coloneqq Rand(0, H)$;
9:     **if** $x_t + W_t \leq W$ and $y_t + H_t \leq H$ **then**
10:       $I_{position} \coloneqq (x_t, y_t, x_t + W_t, y_t + H_t)$;
11:       $I_{position} \coloneqq RGBToGrayscale(I_{position})$;
12:       $I^* \coloneqq I$;
13:       **return** I^*.
14:     **end**
15:   **end**
16: **end**

3.2.3. Implementation Details

To perform camera-aware image-to-image translation for generating synthetic data, we train 30 generative models for the Market-1501 dataset and 2 models for the CUHK03 dataset. These numbers are derived from the formulas $6 \times (6 - 1) = 30$ and $2 \times (2 - 1) = 2$, corresponding to the number of camera pairs in each dataset. For domain-aware image-to-image translation, we train 2 generative models (VeRi-776 to VehicleID and reverse). During training, all input images are first resized to 286×286 pixels, followed by cropping them to 256×256 pixels. We use the Adam optimizer for training all models from scratch, with a batch size of 8. The learning rate is initialized at 0.0002 for the Generator and 0.0001 for the Discriminator. For the first 30 epochs, these rates are kept constant and then linearly decayed to near zero over the subsequent 20 epochs according to a lambda learning rate schedule.

For pre-training, we adopt ResNet101 as the backbone (with support for other ResNet architectures as well). The initial learning rate is set to 0.00035, then reduced by a factor of 0.1 at the 40th and 70th epochs, totaling 350 training epochs with a 10-epoch warmup period. Each training batch consists of 32 identities, with 4 images per identity, resulting in a final batch size of 128. The balance parameter κ for computing the total loss is set to 1. Regarding preprocessing, each image is resized to 256×128 pixels for Person ReID task and 256×256 pixels for Vehicle ReID task. The resized images are padded with 10 pixels using edge padding, followed by random cropping back to their original resized dimensions. Additional augmentation techniques included random horizontal flipping, global grayscale transformation p_{global}, and local grayscale transformation p_{local}, applied with probabilities of 0.5, 0.05, and 0.4, respectively. Images are then converted to 32-bit floating-point pixel values normalized to the $[0,1]$ range. The RGB channels are further normalized by subtracting mean values of $[0.485, 0.456, 0.406]$ and dividing by standard deviations of $[0.229, 0.224, 0.225]$.

3.3. Target-Domain Fine-Tuning

3.3.1. Overall Algorithm

In this phase, we use the pre-trained model to perform comprehensive optimization. We present our CORE-ReID V2 framework (Figure 7) along with Efficient Channel Attention Block (ECAB) and Simplified Efficient Channel Attention Block (SECAB) in Ensemble Fusion++.
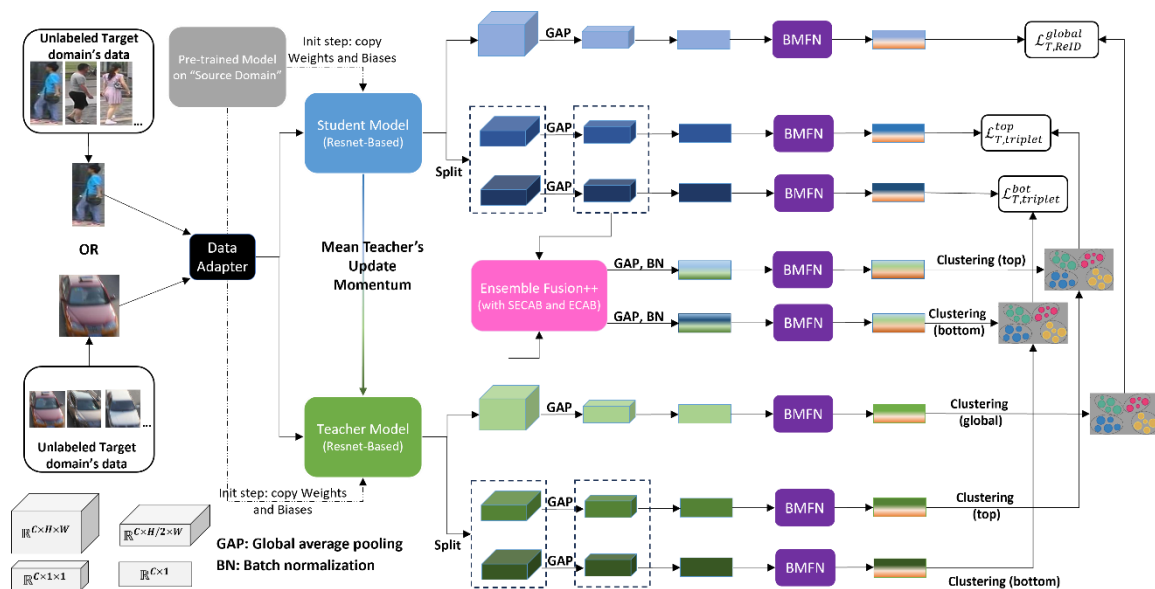

Figure 7. The comprehensive overview of our CORE-ReID V2 framework. The data adapter will pre-process the data depending on the type of object. We integrate local and global features using the enhanced Ensemble Fusion++ component. Specifically, the Efficient Channel Attention Block (ECAB) and Simplified Efficient Channel Attention Block (SECAB) are utilized to boost local and global feature extraction, respectively. By using Bi-directional Mean Feature Normalization (BMFN), the framework effectively merges features from the original image $x_{T,i}$ and its horizontally flipped counterpart $x'_{T,i}$, generating a fused feature $\varphi_l, l \in \{top, bottom\}$. The student network is trained in a supervised manner using pseudo-labels, while the teacher network is updated through a Mean Teacher approach, which computes the temporal average of the student network's weights. Especially, the flipped image features are processed identically to the original image's features until they reach the BMFN stage, ensuring consistent feature fusion.

Building upon the strategies utilized in SSG [16], LF² [17], and CORE-ReID [11], our objective is to enable the model to dynamically integrate both global and local features. This approach allows for feature representations that encompass comprehensive global and detailed local information. To further enhance these feature representations, we incorporate ECAB and SECAB modules during the

fusion process. By organizing multiple clusters based on global and fused features, we aim to generate more reliable pseudo-labels, thereby reducing the risk of ambiguous learning.

To refine these pseudo-labels, we implement a teacher-student network pair grounded in the mean-teacher framework. We feed the same unlabeled image from the target domain into both the teacher and student networks. During iteration i, the student network's parameters, ρ_ς, are updated using Mean Teacher momentum, adjusting them through backpropagation within the target domain training. In parallel, the teacher network's parameters, ρ_τ, are derived as a moving average of the student network's parameters ρ_ς. This is controlled by the temporal momentum coefficient η, which is restricted to the range $[0,1)$. The update rule is defined as:

$$\rho_{\tau,i} = \eta\rho_{\tau,i-1} + (1-\eta)\rho_\varsigma, \tag{6}$$

We utilize the K-means algorithm for clustering to assign pseudo-labels to the data. As a result, each sample $x_{T,i}$ is assigned three pseudo-labels (global, top, and bottom). The target domain dataset is defined as: $\mathbb{D}_T = \{(x_{T,i}, \hat{y}_{T,i,j})|_{i=1}^{N_T}\}$, where $j \in \{global, top, bottom\}$ and N_T represents the total number of images in the target dataset T. The pseudo-label $\hat{y}_{T,i,j} \in \{1,2\ldots,M_{T,j}\}$ indicates that $\hat{y}_{T,i,j}$ is derived from the clustering results $\hat{Y}_j = \{\hat{y}_{T,i,j}|i = 1,2,\ldots,N_T\}$. These are obtained using the combined feature with its flipped counterpart $x'_{T,i}$ generated by BMFN, denoted as $\varphi_l, l \in \{top, bottom\}$. Here, $M_{T,j}$ stands for the number of distinct identities (classes) in the clustering outcome $\hat{Y}_j$.

Before computing the loss function, we use BMFN to extract optimized features from networks f_j^ς, f_j^τ, $j \in \{global, top, bottom\}$ and $\varphi_l, l \in \{top, bottom\}$ from the Ensemble Fusion++. Given an image $x_{T,i}$ in the target dataset, along with its flipped version $x'_{T,i}$, we extract the feature maps F_j^m and the flipped feature maps F'^m_j for $j \in \{global, top\ bottom\}$ and $m \in \{\varsigma, \tau\}$. The BMFN output is computed as follows:

$$f_j^m = \text{BMFN}(F_j^m, F'^m_j) = \frac{\frac{F_j^m + F'^m_j}{2}}{||\frac{F_j^m + F'^m_j}{2}||_2}. \tag{7}$$

$$\varphi_l^m = \text{BMFN}(\theta_l^m, \theta'^m_l) = \frac{\frac{\theta_l^m + \theta'^m_l}{2}}{||\frac{\theta_l^m + \theta'^m_l}{2}||_2}, \tag{8}$$

After obtaining multiple pseudo-labels, we generate three new target-domain datasets to train the student network. The pseudo-labels derived from the local fusion features, denoted as $\varphi_l, l \in \{top, bottom\}$, are used to calculate the softmax triplet loss for the corresponding local features f_l^ς from the student network:

$$\mathcal{L}_{T,triplet}^l = \frac{1}{N_T}\sum_{i=1}^{N_T} \log\left(\frac{e^{||f_l^\varsigma(x_{T,i}|\rho_\varsigma) - f_l^\varsigma(x_{T,i}^-|\rho_\varsigma)||_2}}{e^{||f_l^\varsigma(x_{T,i}|\rho_\varsigma) - f_l^\varsigma(x_{T,i}^-|\rho_\varsigma)||_2} + e^{||f_l^\varsigma(x_{T,i}|\rho_\varsigma) - f_l^\varsigma(x_{T,i}^+|\rho_\varsigma)||_2}}\right), \tag{9}$$

where ρ_τ and ρ_ς are the parameters of the teacher and student networks, respectively. The optimized local feature from the student network is denoted as f_l^ς, with $l \in \{top, bottom\}$. Here, $x_{T,i}^+$ and $x_{T,i}^-$ represent the hardest positive and negative samples relative to the anchor image $x_{T,i}$ in the target domain.

In a similar fashion to supervised learning, we utilize the cluster results $\hat{Y}_{global}$ of the globally clustered feature f_{global}^ς as pseudo-labels to compute the classification loss $\mathcal{L}_{T,ReID}^{global}$ and the global triplet loss $\mathcal{L}_{T,triplet}^{global}$. These losses are defined as follows:

$$\mathcal{L}_{T,ID}^{global} = \frac{1}{N_T}\sum_{i=1}^{N_T}\mathcal{L}_{ce}(C_T(f_{global}^\varsigma(x_{T,i}), \hat{y}_{T,i,global}), \tag{10}$$

$$\mathcal{L}_{T,triplet}^{global} = \frac{1}{N_T}\sum_{i=1}^{N_T}\max(0, ||f_{global}^\varsigma(x_{T,i}) - f_{global}^\varsigma(x_{T,i}^+)||_2 - ||f_{global}^\varsigma(x_{T,i}) - f_{global}^\varsigma(x_{T,i}^-)||_2 + m), \tag{11}$$

where C_T represents the fully connected classification layer of the student network, mapping $f_{global}^{\varsigma}(x_{T,i})$ to the set $\{1,2,\dots,M_{T,global}\}$. The notation $||\cdot||_2$ indicates the L_2-norm distance.

The total loss is computed by combining the different losses with weighting parameters α,β,γ:

$$\begin{aligned}
\mathcal{L}_{T,total} &= \mathcal{L}_{T,ReID}^{global} + \gamma\mathcal{L}_{T,triplet}^{top} + \delta\mathcal{L}_{T,triplet}^{bottom} \\
&= \alpha\mathcal{L}_{T,ID}^{global} + \beta\mathcal{L}_{T,triplet}^{global} + \gamma\mathcal{L}_{T,triplet}^{top} + \delta\mathcal{L}_{T,triplet}^{bottom}.
\end{aligned} \tag{12}$$

During the inference phase, the Ensemble Fusion++ process is bypassed, using only the optimized teacher network to reduce computational overhead. Specifically, the global feature map from the teacher network is split into two segments, referred to as top and bottom features (which also acts similarly in the student network). These segments undergo global average pooling, after which the two local features and the global feature are concatenated. Finally, L_2 normalization and the BMFN method are applied to obtain the optimal feature representation for inference.

3.3.2. Ensemble Fusion++ Component

To extract the fusion features, we horizontally divide the final global feature map of the student network into two segments (top and bottom), resulting in ς_{top} and ς_{bottom} after applying global average pooling. Unlike the Ensemble Fusion component in CORE-ReID [Nguyen, 2024 #29], the final global feature map τ_{global} from the teacher network is further enhanced using the proposed SECAB module. These features ς_{top} and ς_{bottom} from the student network, along with τ_{global} from the teacher network are then utilized for adaptive feature fusion through the Ensemble Fusion++ module, which includes learnable parameters.

The inputs ς_{top} and ς_{bottom} are processed by ECAB, while τ_{global} is processed by SECAB for adaptive fusion. The enhanced attention maps (ψ_{top} and ψ_{bottom}) generated by ECAB are combined with the output τ'_{global} through element-wise multiplication, resulting in the ensemble fusion feature maps: τ'^{top}_{global} and τ'^{bottom}_{global}. These maps undergo Global Average Pooling (GAP) and batch normalization, yielding the fusion features θ_{top} and θ_{bottom}. These features are then fed into the BMFN for predicting pseudo-labels using clustering algorithms in subsequent steps.

The process within Ensemble Fusion++ (Figure 8) can be summarized as follows:

$$\tau'^{top}_{global} = \psi_{top} \otimes \tau'_{global} = ECAB(\varsigma_{top}) \otimes \left[\tau_{global} \otimes SECAB(\tau_{global})\right], \tag{13}$$

$$\tau'^{bot}_{global} = \psi_{bottom} \otimes \tau'_{global} = ECAB(\varsigma_{bottom}) \otimes \left[\tau_{global} \otimes SECAB(\tau_{global})\right], \tag{14}$$
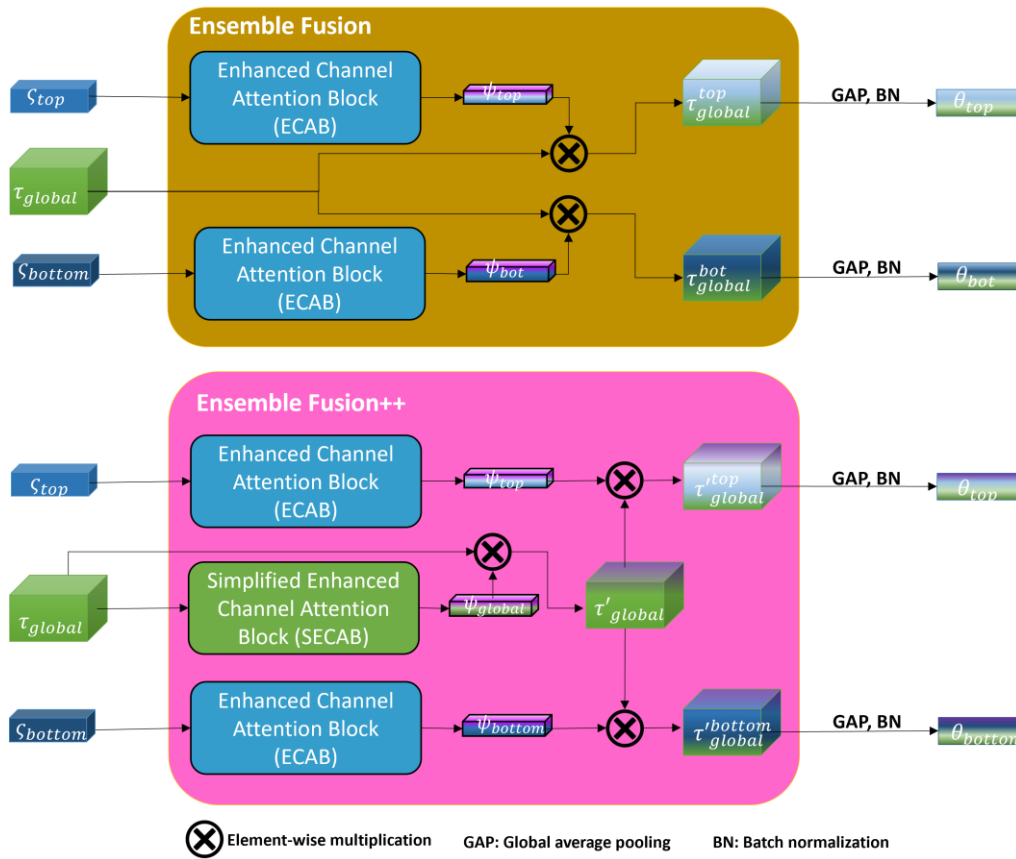
Figure 8. The comparison between Ensemble Fusion in CORE-ReID [11] and proposed Ensemble Fusion++ component. ς_{top} and ς_{bottom} features are passed through the ECAB, τ_{global} feature is passed via the SECAB to produce the channel attention maps by exploiting the inter-channel relationship of features which helps to enhance the features.

3.3.3. SECAB

The importance of attention has been extensively explored in previous literature [74–76]. Attention not only guides where to focus but also enhances the representation of relevant features. Inspired by the ECAB [11], we introduce a new component named SECAB (Figure 9), a straightforward yet impactful attention module for feed-forward convolutional neural networks to enhance the global feature. This module enhances representation power through attention mechanisms that emphasize crucial features while suppressing unnecessary ones. We generate a channel attention map by leveraging inter-channel relationships within features. Each channel of a feature map serves as a feature detector, and channel attention directs focus towards the most meaningful aspects of an input image. To efficiently compute channel attention, we compress the spatial dimension of the input feature map.
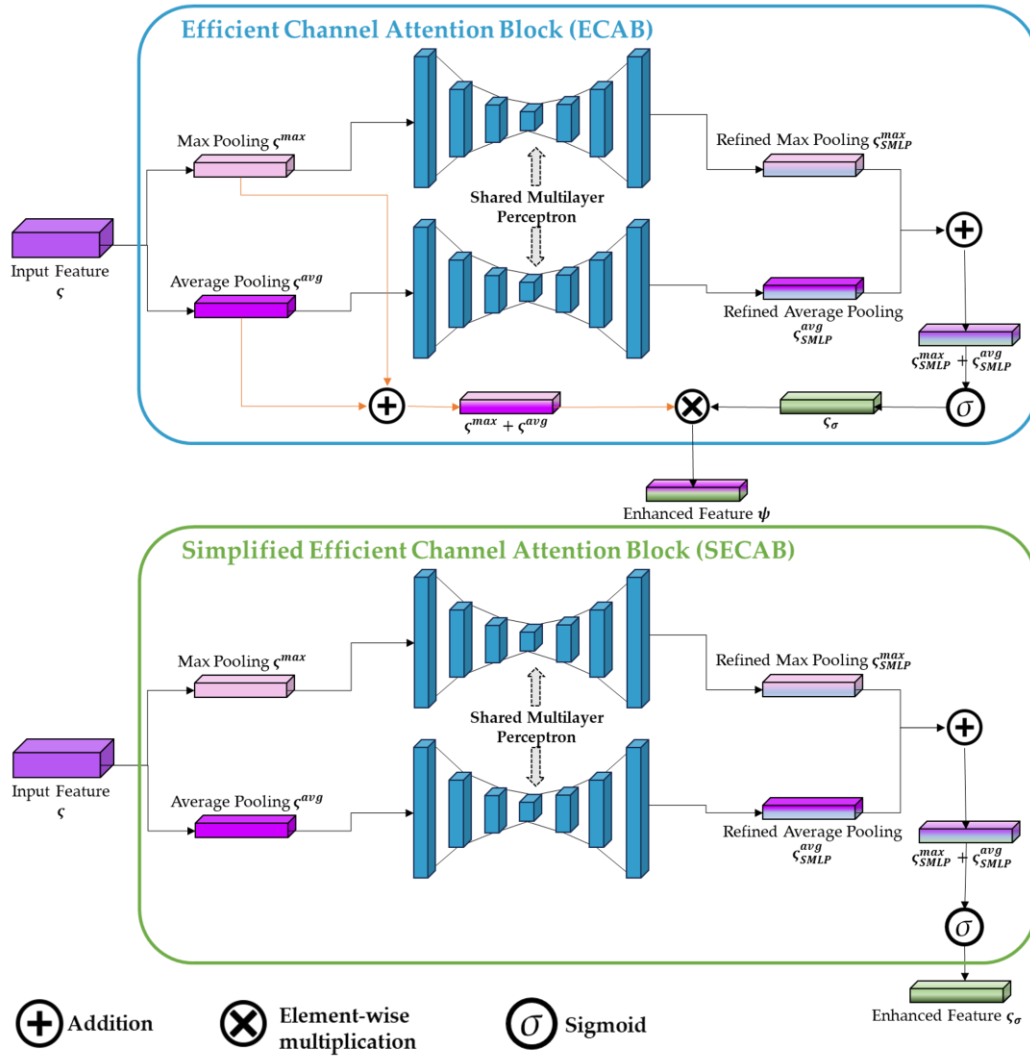
Figure 9. Comparison of ECAB and SECAB. The structure of our SECAB is similar to ECAB [11] but simpler, the module only takes the Shared Multilayer Perceptron into account. It has odd h hidden layers, where the first $\frac{h-1}{2}$ layers are reduced in size with the reduction rate r, and the last $\frac{h-1}{2}$ layers will be expanded with the same rate r.

We utilize both average-pooling and max-pooling features simultaneously to aggregate spatial information. These operations squeeze the spatial dimensions of the input to 1x1 using adaptive pooling, which aggregates global spatial information. **Error! Reference source not found.** shows the design of SECAB. Given an intermediate input feature map $\varsigma \in \mathbb{R}^{C \times W \times H}$, where C, W, H denote the number channel, width, and height respectively. After performing max-pooling and average-pooling, then fit the outputs ς^{max}, ς^{avg} into a Shared Multilayer Perceptron (SMP), we can obtain refined feature as ς_{SMLP}^{max} and ς_{SMLP}^{avg}. The SMP has multiple hidden layers with reduction rate r and the same expansion rate, activation function ReLU. Sigmoid activation squashes the sum to the range between 0 and 1, producing a channel-wise attention mask. The enhanced attention map $\varsigma_\sigma \in \mathbb{R}^{C \times 1 \times 1}$ is calculated as:

$$\varsigma_\sigma = \sigma(\varsigma_{SMLP}^{max} + \varsigma_{SMLP}^{avg}), \tag{15}$$

3.3.4. Greedy K-Means++

In the K-means clustering problem, we are given a set of points $N_T \subseteq \mathbb{R}^{dim}$ in a di-dimensional space, a specified number of clusters $M_{T,j}$. The objective is to identify a set of $M_{T,j}$ centroids $C = \{c_1, c_2, \dots, c_{M_{T,j}}\} \subseteq \mathbb{R}^{dim}$ that minimizes the total sum of squared distances from each point in N_T to

its nearest centroid. Specifically, if we define the cost of a point x with respect to a set of centers C as $\mathcal{D}(x, C) := min_{c \in C} \|x - c\|^2$, the goal is to find C such that $|C| = M_{T,j}$ and the cost $\mathcal{D}(N_T, C) := \sum_{x \in N_T} \mathcal{D}(x, C)$ is minimized. In practice, a simple way to initialize centroids is to select a random subset of N_T with the size $M_{T,j}$. However, this random approach does not guarantee any approximation bounds and can perform poorly in certain cases, such as when $M_{T,j}$ well-separated clusters exist along a single line. Arthur and Vassilvitskii [19] propose a probabilistic seeding method known as K-means++, which improves centroid initialization by favoring points that are far from already-selected centroids, while still maintaining a degree of randomness. Empirical results demonstrate that K-means++ consistently outperforms random seeding on real-world datasets [19].

The Greedy K-means++ algorithm [77] refines this process further by eliminating randomness and selecting centroids deterministically to explicitly maximize the spatial spread. The algorithm operates as follows. At each step, it samples ℓ candidate $c_{i+1}^1, c_{i+1}^2, \ldots, c_{i+1}^\ell$ from a distribution constructed based on the current centroid configuration. Then, for each candidate c_{i+1}^j, the algorithm calculates the new cost $\mathcal{D}(X, C \cup \{c_{i+1}^j\})$ that would result from adding this candidate to the set of centroids. Next, the candidate center that minimizes this cost is selected as the next centroid. In our implementation, ℓ is typically set to $2 + \log(M_{T,j})$. By systematically evaluating multiple candidates at each step, Greedy K-means++ ensures better centroid initialization and improved clustering outcomes (Algorithm 3).

Algorithm 3: Greedy K-means++ seeding

Input: The number of images in target dataset N_T;
  The number of the clusters $M_{T,j}$;
  The number of candidate centers ℓ.
Output: The set of centers C.
Initialization: Uniformly independently sample $c_1^1, c_1^2, \ldots, c_1^\ell \in N_T$.
  1: Let $c_1 = argmin_{c \in \{c_1^1, c_1^2, \ldots, c_1^\ell\}} \mathcal{D}(X, c)$ and set $C_1 = \{c_1\}$.
  2: **for** $i \leftarrow 1, 2, \ldots, M_{T,j} - 1$ **do**
  3:    Sample $c_{i+1}^1, c_{i+1}^2, \ldots, c_{i+1}^\ell \in N_T$ independently;
  4:    Sample x with probability $\frac{\mathcal{D}(x, C_i)}{\mathcal{D}(X, C_i)}$;
  5:    Let $c_{i+1} = argmin_{c \in \{c_i^1, c_i^2, \ldots, c_i^\ell\}} \mathcal{D}(X, C_i \cup \{c\})$;
  6:    Set $C_{i+1} = C_i \cup \{c_{i+1}\}$.
  7: **return** $C := C_{M_{T,j}}$

3.3.5. Detailed Implementation

The training process lasts 80 epochs, with each epoch consisting of 400 iterations. A fixed learning rate of 0.00035 for Person ReID task (0.00007 for Vehicle ReID task) is maintained throughout, and the Adam optimizer is employed with a weight decay of 0.0005 to ensure stable convergence. Clustering operations utilize the K-means algorithm with Greedy K-means++ initialization, where the maximum number of iterations is capped at 100, striking a balance between computational efficiency and solution accuracy. The mini-batch size is set to 512, allowing for efficient centroid updates without processing the entire dataset. An early stopping criterion is applied, terminating clustering if no improvement in inertia is observed over 50 consecutive mini-batches. To address the issue of empty clusters, a reassignment ratio of 0.05 is used, ensuring toughness in dynamic data distributions. For centroid initialization, 1,500 data points are used for global features, while 900 data points are allocated for both top and bottom local features.

In the temporal ensemble regularization process, the momentum parameter (η) is set to 0.999. To balance the contributions of the various components in the loss function, we assign weights as follows: $\alpha = 1, \lambda = 1, \gamma = 0.5,$ and $\delta = 0.5$. For the Ensemble Fusion++ module, a reduction ratio and expansion rate (r) of 4 are utilized, along with 5 hidden layers (h) for both ECAB and SECAB components.

The data adapter resizes input images to 128×256 for the Person ReID task and 256×256 for the Vehicle ReID task. Edge padding of 10 pixels is applied before randomly cropping the images to their respective dimensions (128×256 or 256×256). Data augmentation strategies include random horizontal flipping, global grayscale transformation, local grayscale transformation, and random erasing, applied with probabilities of 0.5, 0.05, 0.4, and 0.5, respectively. These steps ensure a robust and diverse training dataset to improve generalization performance.

4. Results

In this section, we present experimental results, comparing our method against state-of-the-art (SOTA) techniques on widely-used datasets for the task of Unsupervised Domain Adaptation (UDA) for Object ReID.

4.1. Dataset Description

We evaluate the effectiveness of our proposal by conducting evaluations on three benchmark datasets: Market-1501 [78], CUHK03 [79], and MSMT17 [30] for Person ReID and two benchmark datasets: Veri-776 [1] and VehicleID [72] for Vehicle ReID.

Market-1501 [78] contains 32,668 images of 1,501 individuals captured from six different camera views. The training set includes 12,936 images representing 751 identities, while the testing set comprises 3,368 query images and 19,732 gallery images, covering the remaining 750 identities.

CUHK03 [79] features 14,097 images of 1,467 unique individuals, recorded by 6 campus cameras, with each identity captured by 2 cameras. The dataset provides two types of annotations: manually labeled bounding boxes and those generated by an automatic detector. For both training and testing, we utilize the manually annotated bounding boxes. Additionally, we follow a more rigorous testing protocol proposed in [80], which splits the dataset into 767 identities (7,365 images) for training and 700 identities for testing, with 5,332 images in the gallery and 1,400 images in the query set.

MSMT17 [30] is a large-scale dataset comprising 126,441 bounding boxes of 4,101 identities, recorded by 12 outdoor and 3 indoor cameras (15 cameras total) during three periods of the day (morning, afternoon, and noon) over 4 different days. The training set includes 32,621 images featuring 1,041 identities, while the testing set contains 93,820 images representing 3,060 identities. The testing set is further divided into 11,659 query images and 82,161 gallery images. Especially, MSMT17 is significantly larger in scale than both Market-1501 and CUHK03.

VeRi-776 [1] was collected from 20 real-world surveillance cameras in an urban area under diverse conditions, such as orientations, illuminations, and occlusions. It comprises over 50,000 images of 776 vehicles, and approximately 9000 trajectories. The dataset provides a variety of labels, including identity annotations, vehicle attributes, and spatiotemporal information. It is divided into two subsets for training and testing: the training set contains 37,778 images of 576 vehicles, while the test set consists of 11,579 images of the remaining 200 vehicles.

VehicleID [72] contains vehicle images captured by real-world cameras during the daytime. Each subject in the dataset has numerous images taken from the front and back, with some images annotated with model information to aid vehicle identification. The training set comprises 110,178 images of 13,134 vehicles. The test set is divided into three sections: Test800, with 6,532 query images and 800 gallery images of 800 vehicles; Test1600, with 11,385 query images and 1,600 gallery images of 1,600 vehicles; and Test2400, with 17,638 query images and 2,400 gallery images of 2,400 subjects. Following the evaluation protocol of the authors [72], each testing subset divides the query, and gallery sets by randomly selecting one image per subject for the query subset, while the remaining images for each subject form the gallery subset.

The comprehensive overview of the datasets utilized in this document is presented in Table 2.

Table 2. Details of datasets used in this manuscript.

Category	Dataset	Cameras	Training Set (ID/Image)	Test Set (ID/Image)	
				Gallery	Query
Person ReID	Market-1501	6	751/12,936	750/19,732	750/3,368
	CUHK03	2	767/7,365	700/5,332	700/1,400
	MSMT17	15	1,401/32,621	3,060/82,161	3,060/11,659
Vehicle ReID	VeRi-776	20	576/37,778	200/11,579	200/1,678
	VehicleID	-	13,134/110,178	Test800: 800/800	Test800: 800/6,532
				Test1600: 1,600/1,600	Test1600: 1,600/11,395
				Test2400: 2,400/2,400	Test2400: 2,400/17,638

4.2. Evaluation Metrics

For the cross-domain Object ReID task, we utilize Rank-k accuracy (where $k \in \{1, 5, 10\}$ and mean average precision (mAP) to evaluate overall performance on test images.

Rank Ratio Accuracy (Rank-k): The ranking process involves comparing the features extracted from a query object image i with all images in the gallery. This comparison results in a list of images sorted in descending order of similarity, with the most similar images appearing at the top. According to the ground truth of the selected dataset, the position within this sorted list where an image corresponds to the same object as the query image determines its rank. The Rank-k metric reflects the algorithm's accuracy in correctly identifying object images within the top k ranks among the retrieved results for each query:

$$Rank\text{-}k = \frac{\sum_{i=1}^{M} gt(i, k)}{M} \tag{16}$$

Here, M represents the total number of probe images queried from the gallery, and $gt(i, k)$ is a binary function:

$$gt(i, k) = \begin{cases} 1 \text{ if there are positive samples i within the top n ranking results} \\ 0 \text{ otherwise} \end{cases}. \tag{17}$$

Mean Average Precision (mAP): In object ReID, where models produce a ranked list of images, it is crucial to consider the position of each image within the list. For each probe image, the average precision (AP) is calculated as follows:

$$AP = \frac{\sum_{j=1}^{N} p(j) \times gt(j)}{N} \tag{18}$$

where N is the total number of images in the gallery set. The values $p(j)$ and $gt(j)$ represent the precision at the j-th position in the ranking list and a binary function, respectively. If the probe matches the j-th element, then $gt(j) = 1$; otherwise, $gt(j) = 0$. The mean average precision (mAP) across all probe images is then computed using the AP values:

$$mAP = \frac{\sum_{i=1}^{M} AP(i)}{M} \tag{19}$$

Here, M denotes the total number of probe images queried, and $AP(i)$ is the average precision calculated for each probe image i.

4.3. Benchmark on Person ReID

Our study begins by comparing CORE-ReID V2 with state-of-the-art (SOTA) methods on two domain adaptation tasks: Market → CUHK and CUHK → Market (Table 2). We then expand the evaluation to include two additional tasks: Market → MSMT and CUHK → MSMT (Table 3). In these comparisons, "Baseline" refers to the CORE-ReID method developed in our previous work, while CORE-ReID V2 represents the framework proposed in this paper. Additionally, CORE-ReID V2 Tiny

is a lightweight version utilizing the smaller ResNet18 backbone. The evaluation metrics include mAP (%) and rank (R) at k accuracy (%).

Table 3. Experimental results of the proposed CORE-ReID V2 framework and SOTA methods (Acc %) on Market-1501 and CUHK03 datasets. **Bold values** represent the best results while <u>Underline values</u> indicate the second-best performance.

Method	Reference	Market → CUHK				CUHK → Market			
		mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10
SNR [a] [81]	CVPR 2020	17.5	17.1	-	-	52.4	77.8	-	-
UDAR [14]	PR 2020	20.9	20.3	-	-	56.6	77.1	-	-
QAConv50 [a] [82]	ECCV 2020	32.9	33.3	-	-	66.5	85.0	-	-
M³L [a] [83]	CVPR 2021	35.7	36.5	-	-	62.4	82.7	-	-
MetaBIN [a] [84]	CVPR 2021	43.0	43.1	-	-	67.2	84.5	-	-
DFH-Baseline [85]	CVPR 2022	10.2	11.2	-	-	13.2	31.1	-	-
DFH [a] [85]	CVPR 2022	27.2	30.5	-	-	31.3	56.5	-	-
META [a] [86]	ECCV 2022	47.1	46.2	-	-	76.5	90.5	-	-
ACL [a] [87]	ECCV 2022	49.4	50.1	-	-	76.8	90.6	-	-
RCFA [88]	Electronics 2023	17.7	18.5	33.6	43.4	34.5	63.3	78.8	83.9
CRS [89]	JSJTU 2023	-	-	-	-	65.3	82.5	93.0	95.9
MTI [90]	JVCIR 2024	16.3	16.2	-	-	-	-	-	-
PAOA+ [a] [91]	WACV 2024	50.3	50.9	-	-	77.9	91.4	-	-
Baseline (CORE-ReID) [11]	Software 2024	<u>62.9</u>	<u>61.0</u>	<u>79.6</u>	<u>87.2</u>	<u>83.6</u>	<u>93.6</u>	<u>97.3</u>	<u>98.7</u>
CORE-ReID V2 Tiny (ResNet18)	Ours	33.0	31.9	48.9	59.1	60.3	83.4	91.8	94.7
CORE-ReID V2	Ours	**66.4**	**66.9**	**83.4**	**88.9**	**84.5**	**93.9**	**97.6**	**98.7**

Table 4. Experimental results of the proposed CORE-ReID framework and SOTA methods (Acc %) from Market-1501 and CUHK03 source datasets to target domain MSMT17 dataset. **Bold values** represent the best results while <u>Underline values</u> indicate the second-best performance. [a] denotes the method uses multiple source datasets, [b] indicates the implementation is based on the author's code.

Method	Reference	Market → MSMT				CUHK → MSMT			
		mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10
NRMT [92]	ECCV 2020	19.8	43.7	56.5	62.2	-	-	-	-
DG-Net++ [71]	ECCV 2020	22.1	48.4	-	-	-	-	-	-
MMT [15]	ICLR 2020	22.9	52.5	-	-	13.5 [b]	30.9 [b]	44.4 [b]	51.1 [b]
UDAR [14]	PR 2020	12.0	30.5	-	-	11.3	29.6	-	-
Dual-Refinement [93]	ArXiv 2020	25.1	53.3	66.1	71.5	-	-	-	-
SNR [a] [81]	CVPR 2020	-	-	-	-	7.7	22.0	-	-
QAConv50 [a] [82]	ECCV 2020	-	-	-	-	17.6	46.6	-	-
M³L [a] [83]	CVPR 2021	-	-	-	-	17.4	38.6	-	-
MetaBIN [a] [84]	CVPR 2021	-	-	-	-	18.8	41.2	-	-
RDSBN [94]	CVPR 2021	30.9	61.2	73.1	77.4	-	-	-	-
ClonedPerson [95]	CVPR 2022	14.6	41.0	-	-	13.4	42.3	-	-
META [a] [86]	ECCV 2022	-	-	-	-	24.4	52.1	-	-
ACL [a] [87]	ECCV 2022	-	-	-	-	21.7	47.3	-	-
CLM-Net [96]	NCA 2022	29.0	56.6	69.0	74.3	-	-	-	-
CRS [89]	JSJTU 2023	22.9	43.6	56.3	62.7	22.2	42.5	55.7	62.4
HDNet [97]	IJMLC 2023	25.9	53.4	66.4	72.1	-	-	-	-
DDNet [98]	AI 2023	28.5	59.3	72.1	76.8	-	-	-	-
CaCL [99]	ICCV 2023	36.5	66.6	75.3	80.1	-	-	-	-
PAOA+ [a] [91]	WACV 2024	-	-	-	-	26.0	52.8	-	-
OUDA [100]	WACV 2024	20.2	46.1	-	-	-	-	-	-

Method	Reference								
M-BDA [101]	VCIR 2024	26.7	51.4	64.3	68.7	-	-	-	-
UMDA [102]	VCIR 2024	32.7	62.4	72.7	78.4	-	-	-	-
Baseline (CORE-ReID) [11]	Software 2024	41.9	69.5	80.3	84.4	40.4	67.3	79.0	83.1
CORE-ReID V2 Tiny (ResNet18)	Ours	35.8	64.7	76.6	80.8	18.8	44.2	57.1	62.3
CORE-ReID V2	Ours	**44.1**	**71.3**	**82.4**	**86.0**	**40.7**	**68.7**	**79.7**	**83.4**

The results highlight that CORE-ReID V2 significantly outperforms existing SOTA methods, demonstrating the effectiveness of our approach. By incorporating the Ensemble Fusion++ component with ECAB and proposed SECAB, CORE-ReID V2 achieves substantial improvements over the original CORE-ReID. Notably, CORE-ReID V2 surpasses PAOA+ by large margins, achieving mAP improvements of 16.1% and 6.6% on the Market → CUHK and CUHK → Market tasks, respectively, even though PAOA+ utilizes additional training data. Additionally, our framework delivers significant enhancements over CACL and PAOA+, achieving mAP gains of 7.6% and 14.7% mAP on Market → MSMT and CUHK → MSMT tasks, respectively.

4.4. Benchmark on Vehicle ReID

We evaluate CORE-ReID V2 against state-of-the-art methods on VehicleID → VeRi-776 (Table 5) and VeRi-776→ VehicleID (Table 6) tasks. "Baseline" refers to the implementation based on CORE-ReID [11] with Ensemble Fusion component, while CORE-ReID V2 is the proposed algorithm. CORE-ReID V2 Tiny is a lightweight variant using ResNet18. Metrics include mAP (%) and rank (R) at k accuracy (%).

Table 5. Experimental results of CORE-ReID V2 framework and SOTA methods on VehicleID → VeRi-776. **Bold values** represent the best results while <u>Underline values</u> indicate the second-best performance.

Method	Reference	VehicleID → VeRi-776			
		mAP	R-1	R-5	R-10
FACT [1]	ECCV 2016	18.75	52.21	72.88	-
PUL [43]	ACM 2018	17.06	55.24	67.34	-
SPGAN [22]	CVPR 2018	16.4	57.4	70.0	75.6
VR-PROUD [103]	PR 2019	22.75	55.78	70.02	-
ECN [104]	CVPR 2019	20.06	57.41	70.53	-
MMT [15]	ICLR 2020	35.3	74.6	82.6	-
SPCL [45]	NIPS 2020	38.9	80.4	86.8	-
PAL [105]	IJCAI 2020	42.04	68.17	79.91	-
UDAR [14]	PR 2020	35.80	76.90	85.80	<u>89.00</u>
ML [106]	ICME 2021	36.90	77.80	85.50	-
PLM [107]	Sci.China 2022	47.37	77.59	87.00	-
VDAF [108]	MTA 2023	24.86	46.32	55.17	-
CSP+FCD [109]	Elec 2023	45.60	74.30	83.70	-
MGR-GCL [5]	ArXiv 2024	48.73	<u>79.29</u>	87.95	-
MATNet+DMDU [110]	ArXiv 2024	<u>49.25</u>	79.13	<u>88.97</u>	-
Baseline	Ours	47.70	78.12	86.23	88.14
CORE-ReID V2 Tiny (ResNet18)	Ours	40.17	73.00	81.41	85.40
CORE-ReID V2	Ours	**49.50**	**80.15**	**89.05**	**90.29**

Table 6. Experimental results of the proposed CORE-ReID V2 framework and SOTA methods on VeRi-776 → VehicleID. **Bold values** represent the best results while Underline values indicate the second-best performance.

Method	Reference	VeRi-776 → VehicleID Test800				VeRi-776 → VehicleID Test1600				VeRi-776 → VehicleID Test2400			
		mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10
FACT [1]	ECCV 2016	-	49.53	67.96	-	-	44.63	64.19	-	-	39.91	60.49	-
Mixed Diff+CCL [72]	CVPR 2016	-	49.00	73.50	-	-	42.80	66.80	-	-	38.20	61.60	-
PUL [43]	ACM 2018	43.90	40.03	56.03	-	37.68	33.83	49.72	-	34.71	30.90	47.18	-
PAL [105]	IJCAI 2020	53.50	50.25	64.91	-	48.05	44.25	60.95	-	45.14	41.08	59.12	-
UDAR [14]	PR 2020	59.60	54.00	66.10	72.01	55.30	48.10	64.10	70.20	52.90	45.20	62.60	69.14
ML [106]	ICME 2021	61.60	54.80	69.20	-	48.70	40.30	57.70	-	45.00	36.50	54.10	-
PLM [107]	Sci.China 2022	54.85	51.23	67.11	-	49.41	45.40	63.37	-	46.00	41.73	60.94	-
CSP+FCD [109]	Elec 2023	51.90	54.40	67.40	-	46.50	52.70	65.60	-	42.70	45.90	60.30	-
VDAF [108]	MTA 2023	-	-	-	-	-	47.03	64.86	-	-	43.69	61.76	-
MGR-GCL [5]	ArXiv 2024	55.24	52.38	75.29	-	50.56	45.88	67.65	-	47.59	42.83	64.36	-
DMDU [110]	TITS 2024	61.83	55.61	68.25	-	56.73	53.28	63.56	-	53.97	47.59	61.85	-
Baseline	Ours	64.28	56.16	74.55	81.15	60.02	51.84	71.62	78.08	56.15	47.85	66.89	75.27
CORE-ReID V2 Tiny (ResNet18)	Ours	63.87	55.18	73.43	81.11	59.69	50.05	70.88	77.75	55.14	45.99	65.07	73.54
CORE-ReID V2	Ours	**67.04**	**58.32**	**76.51**	**84.32**	**63.02**	**53.49**	**74.36**	**81.85**	**57.99**	**48.62**	**68.30**	**77.11**

Across both transfer scenarios (VeRi-776 → VehicleID and VehicleID → VeRi-776), the proposed CORE-ReID V2 framework demonstrates superior performance compared to existing state-of-the-art (SOTA) methods. Specifically, our evaluation includes supervised approaches such as FACT and Mixed Diff+CCL, as well as unsupervised Person ReID methods PUL and UDAR. Additionally, we incorporate leading SOTA techniques, including PAL, MMT, SPCL PLM, CSP+FCD, and DMDU, for a comprehensively comparative analysis. For the VeRi-776 → VehicleID adaptation task, CORE-ReID V2 achieves 67.0%, 63.02%, and 57.99% mAP across three test modes, surpassing the DMDU method by 5.21%, 6.29%, and 4.02% in each case. In the VehicleID → VeRi-776 transfer scenario, MGR-GCL and MATNet+DMDU attain 48.73% and 49.25% mAP, respectively. CORE-ReID V2 outperforms all competing methods, achieving a mAP of 49.50% and Rank-1 accuracy of 80.15%, setting its position as a new SOTA approach.

4.5. Ablation Study

To validate our approach, we employ Grad-CAM [111] to visualize feature maps at the global feature level. Key features for each person and vehicle are highlighted using heatmaps, where color intensity indicates importance - blue represents less significant regions, while red denotes the most crucial areas for Object Re-identification. As illustrated in Figures 10 and 11, the essential features are concentrated on the target person's body and the vehicle's structure. Furthermore, the heatmaps exhibit similar distributions between the original and flipped images, reinforcing the performance of our method. This consistency aligns with the accuracy results reported above, further validating the effectiveness of CORE-ReID V2.
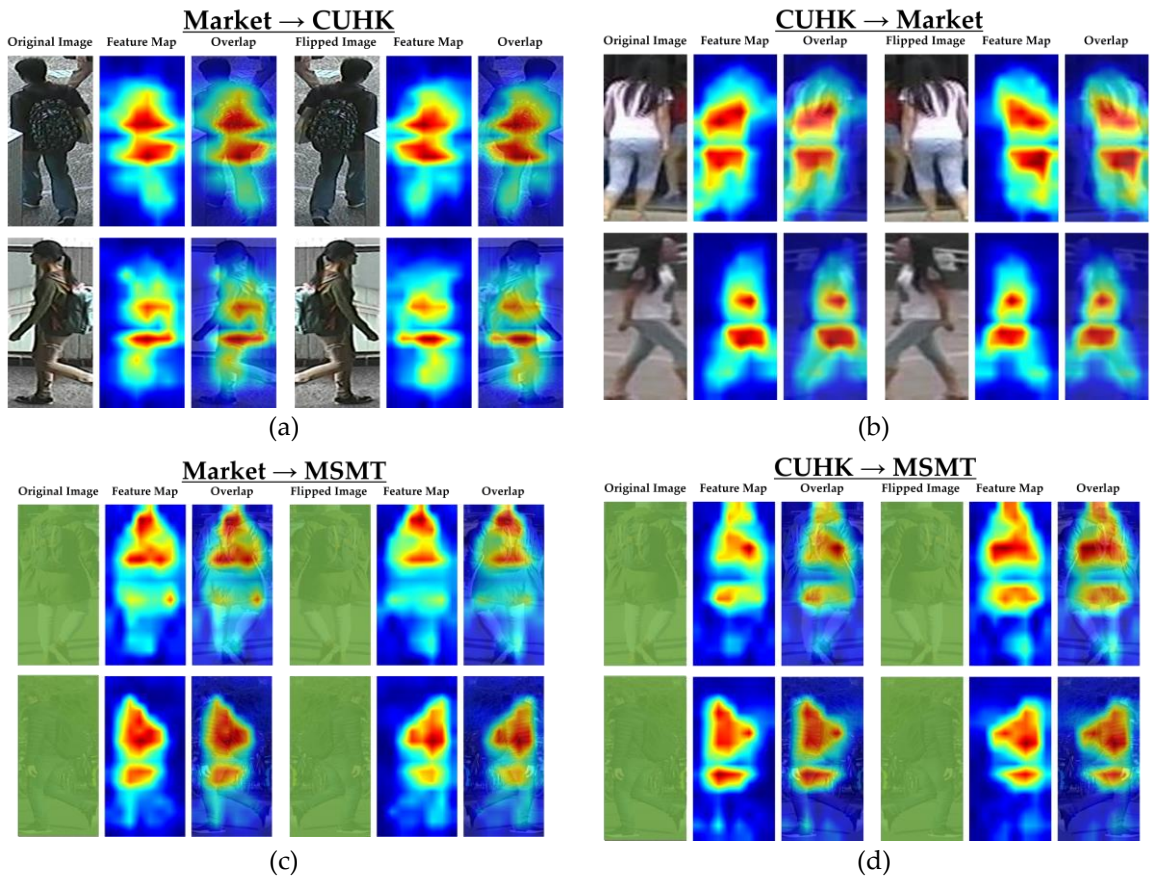
Figure 10. Feature maps visualization using Grad-CAM [111]. (a), (b), (c), and (d) illustrate the feature maps of those pairs on Market→CUHK, CUHK→Market, Market→MSMT, and CUHK→MSMT, respectively.
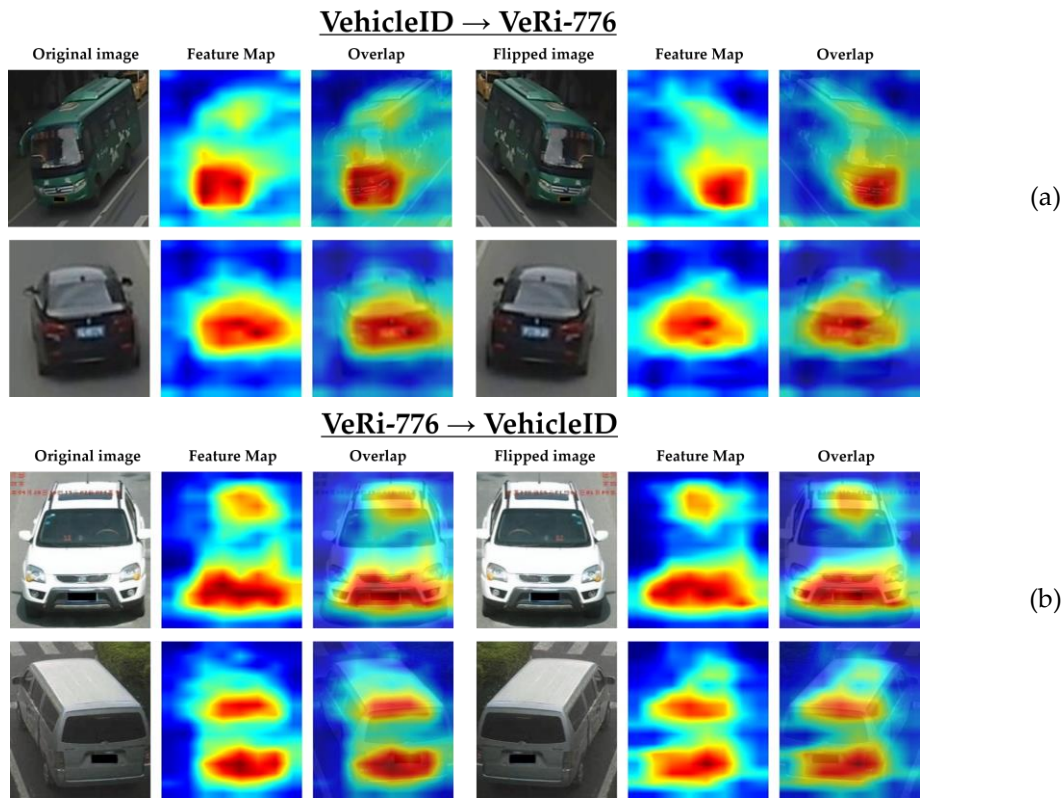

Figure 11. Feature maps visualization using Grad-CAM [111]. (a), (b) illustrate the feature maps of those pairs on VehicleID→ VeRi-776 and VeRi-776 → VehicleID respectively.

In the Market → MSMT and CUHK → MSMT transfer scenarios, the Market → MSMT model demonstrates a slightly superior ability to extract important features. The heatmaps reveal a more concentrated distribution in the middle and lower body regions for both the original and flipped images. This observation may explain the higher accuracy achieved by the Market → MSMT model compared to the CUHK → MSMT model, as reported in Table 4.

K-means Clustering Settings: we utilize the K-Means clustering approach to generate pseudo-labels for the target domain, with parameters varying across different datasets. As shown in Table 7, our framework achieves optimal performance on Market → CUHK, CUHK → Market, Market → MSMT, CUHK → MSMT, VehicleID → VeRi-776, and VeRi-776 → VehicleID Small with cluster settings of 900, 900, 2000, 2000, 500 and 700, respectively.

Table 7. Experimental results on different settings of number of pseudo identities in K-means clustering algorithm for both Person and Vehicle ReID tasks. **Bold values** represent the best results.

Person ReID	Market → CUHK				CUHK → Market			
Number of Clusters	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10
Ours ($M_{T,j} = 500$)	44.4	43.2	65.3	76.4	69.4	86.8	94.9	96.7
Ours ($M_{T,j} = 700$)	57.8	59.1	76.1	83.6	81.7	92.7	97.1	98.1
Ours ($M_{T,j} = 900$)	**66.4**	**66.9**	**83.4**	**88.9**	**84.5**	**93.9**	**97.6**	**98.7**
Person ReID	Market → MSMT				CUHK → MSMT			
Number of Clusters	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10
Ours ($M_{T,j} = 2000$)	**44.1**	**71.3**	**82.4**	**86.0**	**40.68**	**68.66**	**79.74**	**83.36**
Ours ($M_{T,j} = 2500$)	41.1	68.9	80.5	84.2	38.91	67.26	78.97	82.80
Ours ($M_{T,j} = 3000$)	38.9	67.2	79.0	83.2	35.8	64.7	76.6	80.8
Vechile ReID	VehicleID → VeRi-776				VeRi-776 → VehicleID Small			
Number of Clusters	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10
Ours ($M_{T,j} = 500$)	49.50	**80.15**	**89.05**	**90.29**	66.60	58.20	75.90	83.70
Ours ($M_{T,j} = 700$)	**49.63**	79.14	86.65	89.69	**67.04**	**58.32**	**76.51**	**84.32**
Ours ($M_{T,j} = 900$)	48.61	79.02	86.29	89.15	66.70	57.50	77.60	84.20

Figure 12 shows that the performance of our approach varies depending on the dataset pairs and the clustering parameter values ($M_{T,j}$) utilized.
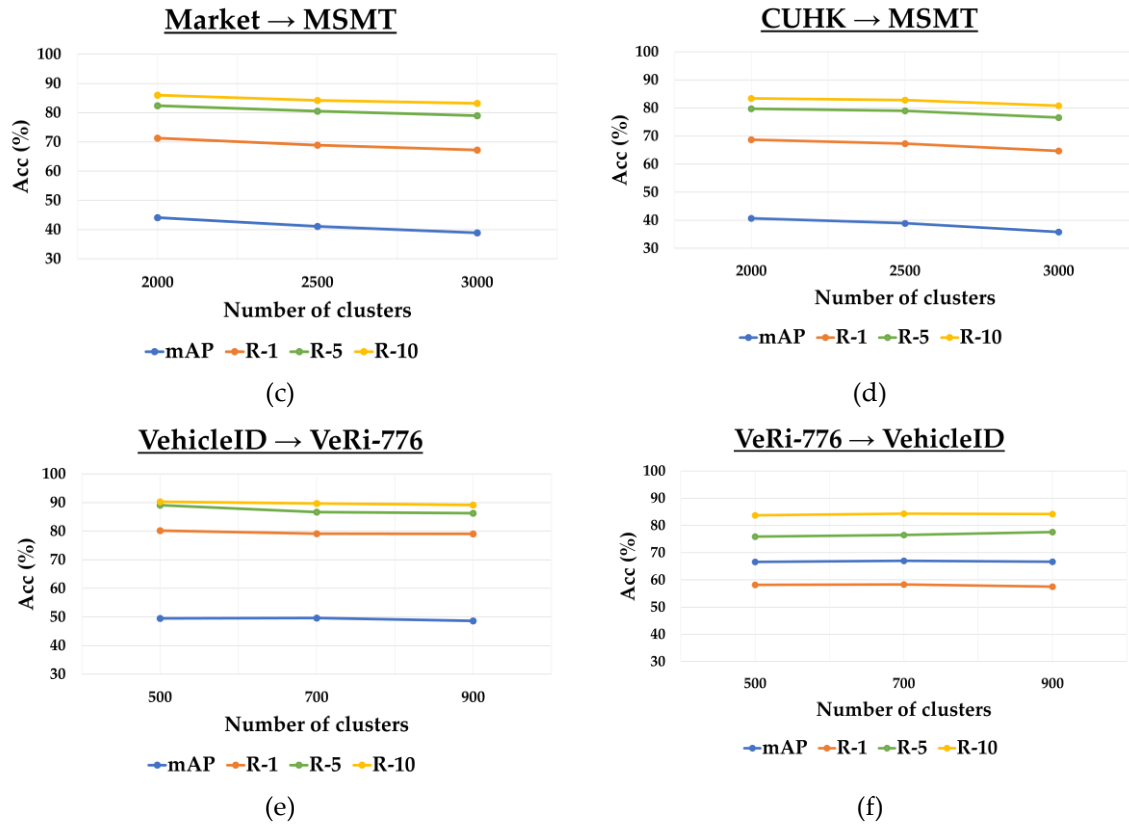

(a)   (b)
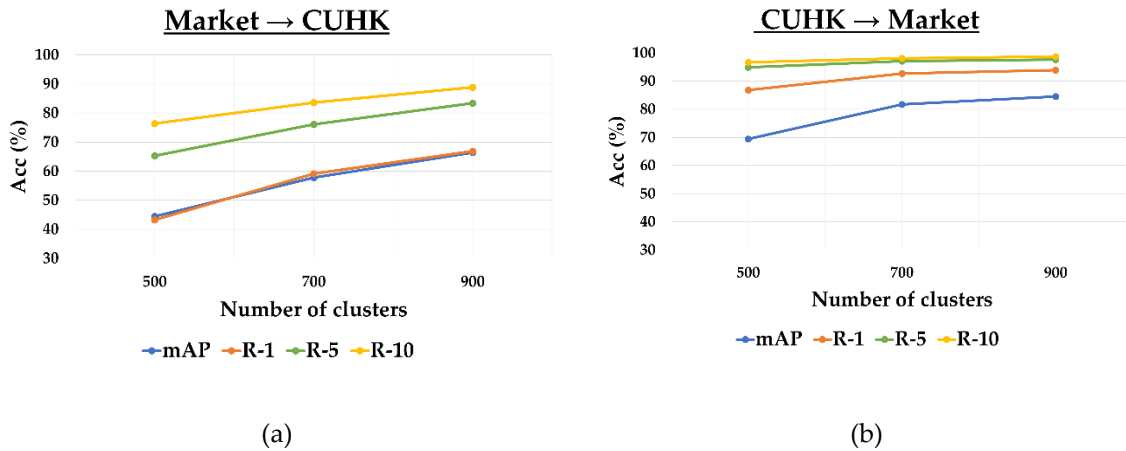
Figure 12. Impact of clustering parameter $M_{T,j}$. Results on (a) Market → CUHK, (b) CUHK → Market, (c) Market → MSMT, (d) CUHK → MSMT, (e) VehicleID → VeRi-776, and (f) VeRi-776 → VehicleID Small.

Greedy K-means++ Initialization: we enhance clustering performance by employing the greedy K-means++ initialization strategy, which optimally balances randomness and centroid selection to improve cluster quality. This approach not only strengthens feature learning but also ensures a more stable pseudo-label generation, addressing challenges associated with obscure learning. Table 8 presents the experimental results comparing greedy K-means++ initialization with a random approach.

Table 8. Experimental results using the greedy K-means++ initialization and random approach. The clustering parameter values $M_{T,j}$ are carried out from the study of K-means clustering settings. **Bold values** represent better results.

Person ReID	Market → CUHK ($M_{T,j} = 900$)				CUHK → Market ($M_{T,j} = 900$)			
Method	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10
Ours (Random)	63.6	63.8	80.9	87.8	83.8	93.6	97.4	98.6
Ours (Greedy Initialization)	**66.4**	**66.9**	**83.4**	**88.9**	**84.5**	**93.9**	**97.6**	**98.7**
Person ReID	Market → MSMT ($M_{T,j} = 2000$)				CUHK → MSMT ($M_{T,j} = 2000$)			
Method	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10
Ours (Random)	42.2	69.7	80.2	84.9	40.5	67.6	78.8	83.1
Ours (Greedy Initialization)	**44.1**	**71.3**	**82.4**	**86.0**	**40.7**	**68.7**	**79.7**	**83.4**
Vehicle ReID	VehicleID → VeRi-776 ($M_{T,j} = 500$)				VeRi-776 → VehicleID Small ($M_{T,j} = 700$)			
Method	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10
Ours (Random)	47.72	78.23	86.56	88.26	65.79	56.14	75.95	83.56
Ours (Greedy Initialization)	**49.50**	**80.15**	**89.05**	**90.29**	**67.04**	**58.32**	**76.51**	**84.32**

SECAB Configuration: We use SECAB to enhance global features, leveraging attention mechanisms to emphasize crucial features while suppressing unnecessary ones. To validate the effectiveness of SECAB, we conduct an experiment by removing it from our network, as shown in Table 9.

Table 9. Experimental results validating the effectiveness of SECAB in our proposed framework. The clustering parameter values ($M_{T,j}$) are derived from the study of K-means clustering settings. **Bold values** represent the better results.

Person ReID	Market → CUHK ($M_{T,j} = 900$)				CUHK → Market ($M_{T,j} = 900$)			
Method	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10
Ours (without SECAB)	65.0	65.1	82.6	87.6	83.9	93.7	97.4	98.6
Ours (with SECAB)	**66.4**	**66.9**	**83.4**	**88.9**	**84.5**	**93.9**	**97.6**	**98.7**
Person ReID	Market → MSMT ($M_{T,j} = 2000$)				CUHK → MSMT ($M_{T,j} = 2000$)			
Method	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10
Ours (without SECAB)	43.2	70.3	81.8	85.2	40.5	68.0	79.2	83.1
Ours (with SECAB)	**44.1**	**71.3**	**82.4**	**86.0**	**40.7**	**68.7**	**79.7**	**83.4**
Vehicle ReID	VehicleID → VeRi-776 ($M_{T,j} = 500$)				VeRi-776 → VehicleID Small ($M_{T,j} = 700$)			
Method	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10
Ours (without SECAB)	48.03	78.92	87.61	88.93	65.14	57.02	75.56	82.97
Ours (with SECAB)	**49.50**	**80.15**	**89.05**	**90.29**	**67.04**	**58.32**	**76.51**	**84.32**

Backbone Settings: we assess the performance of both complex backbones (ResNet50, ResNet101, and ResNet152) and lightweight backbones (ResNet18 and ResNet34) for unsupervised domain adaptation in Object ReID. Through extensive experiments and analysis, we gain insights into the impact of backbone architecture on overall performance, as well as its computational efficiency and suitability for resource-constrained environments, as shown in Table 10. Among the tested configurations, ResNet101 delivers the best performance in Market → CUHK, CUHK → Market, VehicleID → VeRi-776, and VeRi-776 → VehicleID scenarios. All experiments were conducted on two machines, each equipped with dual Quadro RTX 8000 GPUs.

Table 10. Experimental results on different settings of ResNet backbones in Market → CUHK, CUHK → Market scenarios. **Bold values** represent the best results.

Person ReID	Market → CUHK ($M_{T,j} = 900$)				CUHK → Market ($M_{T,j} = 900$)			
Method	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10
Ours (ResNet18)	33.0	31.9	48.9	59.1	60.3	83.4	91.8	94.7
Ours (ResNet34)	38.8	38.4	55.9	64.7	64.4	85.9	93.7	95.4
Ours (ResNet50)	64.9	64.1	81.3	87.9	83.7	93.8	97.6	98.5
Ours (ResNet101)	**66.4**	**66.9**	**83.4**	**88.9**	**84.5**	**93.9**	**97.6**	**98.7**
Ours (ResNet152)	65.2	65.1	82.1	87.9	83.5	93.2	97.5	98.1
Vehicle ReID	VehicleID → VeRi-776 ($M_{T,j} = 500$)				VeRi-776 → VehicleID Small ($M_{T,j} = 700$)			
Method	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10
Ours (ResNet18)	40.17	73.00	81.41	85.40	63.87	55.18	73.43	81.11
Ours (ResNet34)	46.62	75.92	83.73	87.49	63.80	54.80	73.60	80.30
Ours (ResNet50)	48.11	78.84	86.71	89.81	67.02	58.30	**77.00**	83.90
Ours (ResNet101)	**49.50**	**80.15**	**89.05**	**90.29**	**67.04**	**58.32**	76.51	**84.32**
Ours (ResNet152)	48.07	78.26	86.73	89.96	66.97	58.23	76.49	83.86

Figure 13 demonstrates that lightweight backbones, such as ResNet18 and ResNet34, are also supported, highlighting the adaptability of our framework to various backbone architectures.
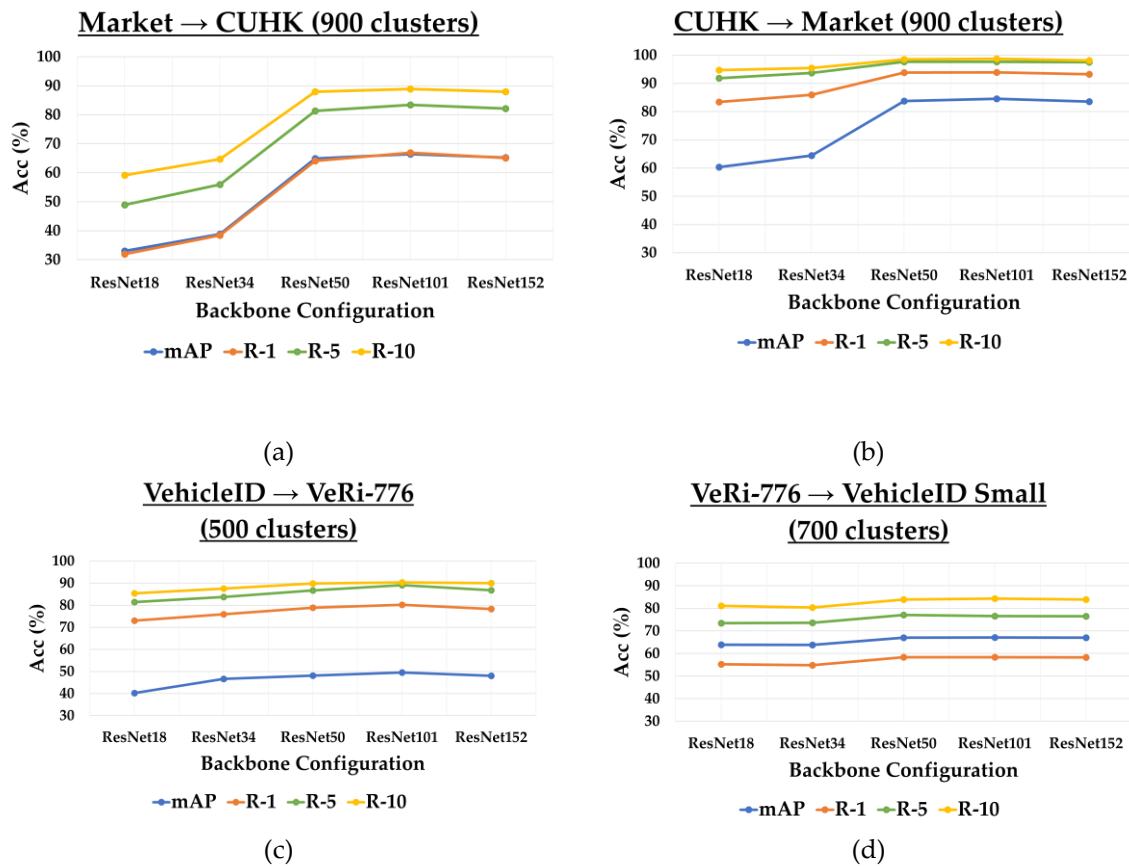
Figure 13. Impact of the backbone settings. Results on (a) Market → CUHK, (b) CUHK → Market, (c) VehicleID → VeRi-776, and (d) VeRi-776 → VehicleID Small show that ResNet101 backbone gives the best overall results.

5. Conclusions

In this paper, we introduced CORE-ReID V2, an enhanced framework designed to address limitations in its predecessor, CORE-ReID, while extending its applicability to Object ReID tasks. By incorporating the novel Ensemble Fusion++ module, which adaptively enhances both local and global features, and utilizing advanced clustering techniques such as greedy KMeans++ initialization, CORE-ReID V2 achieves superior performance in Unsupervised Domain Adaptation (UDA) for Person and Vehicle ReID tasks. Furthermore, support for lightweight backbones like ResNet18 and ResNet34 makes the framework suitable for real-time and resource-constrained applications. Experimental results on widely used UDA Person ReID and Vehicle ReID datasets demonstrate that CORE-ReID V2 outperforms state-of-the-art methods, showcasing its strength and adaptability. These contributions not only push the boundaries of UDA-based Object ReID but also provide a solid foundation for further exploration in this domain.

Despite its advancements, CORE-ReID V2 has several limitations that warrant attention. The scalability of the framework to large-scale datasets with millions of instances remains unexplored, which could introduce challenges in both performance and efficiency. While the framework performs well on benchmark datasets, its scalability to larger and more diverse datasets with millions of instances remains unexplored. Furthermore, the framework's primary focus on Person and Vehicle ReID tasks also limits its exploration of broader Object ReID applications, such as animal or product identification. Moreover, its reliance on the quality of pseudo-labels makes it vulnerable to performance degradation in noisy or highly complex scenarios.

Future work will address these limitations by incorporating distributed training techniques and more efficient clustering algorithms to enhance scalability for large datasets. Extending the framework's application to a broader range of Object ReID tasks, such as animal, product, or scene-specific ReID, by incorporating domain-specific priors and augmentation strategies. Exploring

advanced techniques, such as contrastive learning and adversarial regularization, to mitigate the impact of noisy pseudo-labels and improve model performance. Hopefully, these directions aim to unlock the full potential of CORE-ReID V2 and inspire future advancements in UDA for Object ReID.

Abbreviations

The following abbreviations are used in this manuscript:

ECAB	Efficient Channel Attention Block
BMFN	Bidirectional Mean Feature Normalization
CBAM	Convolutional Block Attention Module
CNN	Convolutional Neural Network
CORE-ReID	Comprehensive Optimization and Refinement through Ensemble fusion in Domain Adaptation for Person Re-identification
HHL	Hetero-Homogeneous Learning
MMFA	Multi-task Mid-level Feature Alignment
MMT	Mutual Mean-Teaching
Object ReID	Object Re-identification
SECAB	Simplified Efficient Channel Attention Block
SOTA	State-Of-The-Art
SSG	Self-Similarity Grouping
UDA	Unsupervised Domain Adaptation
UNRN	Uncertainty-Guided Noise-Resilient Network

References

1. Liu, X.; Liu, W.; Mei, T.; Ma, H. A Deep Learning-Based Approach to Progressive Vehicle Re-identification for Urban Surveillance. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 2016; pp. 869–884.

2. Liu, X.; Liu, W.; Mei, T.; Ma, H. PROVID: Progressive and Multimodal Vehicle Reidentification for Large-Scale Urban Surveillance. *IEEE Transactions on Multimedia* **2018**, *20*, 645-658, doi:10.1109/TMM.2017.2751966.

3. Wang, Z.; He, L.; Tu, X.; Zhao, J.; Gao, X.; Shen, S. Robust Video-Based Person Re-Identification by Hierarchical Mining. *IEEE Transactions on Circuits and Systems for Video Technology 32*, 8179 - 8191, doi:10.1109/TCSVT.2021.3076097.

4. Zheng, L.; Yang, Y.; Hauptmann, A.G. Person Re-identification: Past, Present and Future. *arXiv* **2016**, *arXiv:1610.02984*.

5. Chang, Z.; Zheng, S. Revisiting Multi-Granularity Representation via Group Contrastive Learning for Unsupervised Vehicle Re-identification. *ArXiv* **2024**, *abs/2410.21667*, doi:10.48550/arXiv.2410.21667.

6. Luo, H.; Jiang, W.; Gu, Y.; Liu, F.; Liao, X.; Lai, S. A Strong Baseline and Batch Normalization Neck for Deep Person Re-Identification. *IEEE Transactions on Multimedia 22*, 2597 - 2609, doi:10.1109/TMM.2019.2958756.

7. Sharma, C.; Kapil, S.R.; Chapman, D. Person Re-Identification with a Locally Aware Transformer. *ArXiv* **2021**, *abs/2106.03720*, doi:10.48550/arXiv.2106.03720.

8. Chen, W.; Xu, X.; Jia, J.; Luo, H.; Wang, Y.; Wang, F. Beyond Appearance: A Semantic Controllable Self-Supervised Learning Framework for Human-Centric Visual Tasks. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17-24 June 2023, 2023; pp. 15050-15061.

9. Almeida, E.; Silva, B.; Batista, J. Strength in Diversity: Multi-Branch Representation Learning for Vehicle Re-Identification. In Proceedings of the 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC), Bilbao, Spain, 24-28 September 2023; pp. 4690-4696.

10. Li, J.; Gong, X. Prototypical Contrastive Learning-based CLIP Fine-tuning for Object Re-identification. *arXiv* **2023**, doi:10.48550/arXiv.2310.17218.

11. Nguyen, T.Q.; Prima, O.D.A.; Hotta, K. CORE-ReID: Comprehensive Optimization and Refinement through Ensemble Fusion in Domain Adaptation for Person Re-Identification. *Software* **2024**, *3*, 227-249, doi:10.3390/software3020012.

12. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), as Vegas, NV, USA, 27-30 June 2016; pp. 770-778.

13. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jegou, H. Training Data-efficient Image Transformers & Distillation Through Attention. In Proceedings of the 38th International Conference on Machine Learning, Virtual, 2021; pp. 10347 - 10357.

14. Song, L.; Wang, C.; Zhang, L.; Du, B.; Zhang, Q.; Huang, C.; Wang, X. Unsupervised Domain Adaptive Re-identification: Theory and Practice. *Pattern Recognition* **2020**, *102*, 11, doi:10.1016/j.patcog.2019.107173.

15. Ge, Y.; Chen, D.; Li, H. Mutual Mean-Teaching: Pseudo Label Refinery for Unsupervised Domain Adaptation on Person Re-identification. *ArXiv* **2020**, doi:10.48550/arXiv.2001.01526.

16. Fu, Y.; Wei, Y.; Wang, G.; Zhou, Y.; Shi, H.; Uiuc, U. Self-Similarity Grouping: A Simple Unsupervised Cross Domain Adaptation Approach for Person Re-Identification. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 27 October 2019 - 02 November 2019; pp. 6111-6120.

17. Ding, J.; Zhou, X. Learning Feature Fusion for Unsupervised Domain Adaptive Person Re-identification. In Proceedings of the 2022 26th International Conference on Pattern Recognition (ICPR), Montreal, QC, Canada, 21–25 August 2022; pp. 2613-2619.

18. Zhou, R.; Wang, Q.; Cao, L.; Xu, J.; Zhu, X.; Xiong, X.; Zhang, H.; Zhong, Y. Dual-Level Viewpoint-Learning for Cross-Domain Vehicle Re-Identification. *Electronics* **2024**, *13*, doi:10.3390/electronics13101823.

19. Arthur, D.; Vassilvitskii, S. K-means++: The Advantages of Careful Seeding. In Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, New Orleans, Louisiana, 2007; pp. 1027 - 1035.

20. 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), Montreal, QC, Canada, 10-17 October 2021; pp. 15013-15022.

21. Chen, Y.; Zhu, X.; Gong, S. Instance-Guided Context Rendering for Cross-Domain Person Re-Identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 27 October - 02 November 2019; pp. 232-242.

22. Deng, W.; Zheng, L.; Ye, Q.; Kang, G.; Yang, Y.; Jiao, J. Image-Image Domain Adaptation with Preserved Self-Similarity and Domain-Dissimilarity for Person Re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18-23 June 2018; pp. 994-1003.

23. Zhang, X.; Ge, Y.; Qiao, Y.; Li, H. Refining Pseudo Labels with Clustering Consensus over Generations for Unsupervised Object Re-identification. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20-25 June 2021; pp. 3435-3444.

24. Saenko, K.; Kulis, B.; Fritz, M.; Darrell, T. Adapting Visual Category Models to New Domains. In Proceedings of the Computer Vision – ECCV 2010, Berlin, Heidelberg, 5-11 September, 2010; pp. 213–226.

25. Gong, B.; Shi, Y.; Sha, F.; Grauman, K. Geodesic Flow Kernel for Unsupervised Domain Adaptation. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16-21 June 2012; pp. 2066-2073.

26. Fernando, B.; Habrard, A.; Sebban, M.; Tuytelaars, T. Unsupervised Visual Domain Adaptation Using Subspace Alignment. In Proceedings of the 2013 IEEE International Conference on Computer Vision, Sydney, NSW, Australia, 01-08 December 2013; pp. 2960-2967.

27. Sun, B.; Feng, J.; Saenko, K. Return of Frustratingly Easy Domain Adaptation. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, Phoenix, Arizona, 12 February 2016; pp. 2058 - 2065.

28. Zhou, L.; Ye, M.; Li, X.; Zhu, C.; Liu, Y.; Li, X. Disentanglement Then Reconstruction: Unsupervised Domain Adaptation by Twice Distribution Alignments. *Expert Systems with Applications* **2024**, *237*, 121498, doi:10.1016/j.eswa.2023.121498.

29. Bousmalis, K.; Silberman, N.; Dohan, D.; Erhan, D.; Krishnan, D. Unsupervised Pixel–Level Domain Adaptation with Generative Adversarial Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21-26 July 2017; pp. 95-104.

30. Wei, L.; Zhang, S.; Gao, W.; Tian, Q. Person Transfer GAN to Bridge Domain Gap for Person Re-Identification. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18-23 June 2018; pp. 79-88.

31. Huang, Y.; Wu, Q.; Xu, J.; Zhong, Y. SBSGAN: Suppression of Inter-Domain Background Shift for Person Re-Identification. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 27 October 2019 - 02 November 2019; pp. 9526-9535.

32. Liu, M.-Y.; Tuzel, O. Coupled Generative Adversarial Networks. In Proceedings of the 30th International Conference on Neural Information Processing Systems, Barcelona, Spain; pp. 469 - 477.

33. Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; March, M.; Lempitsky, V. Domain-Adversarial Training of Neural Networks. *Journal of Machine Learning Research* **2016**, *17*, 1-35.

34. Motiian, S.; Piccirilli, M.; Adjeroh, D.A.; Doretto, G. Unified Deep Supervised Domain Adaptation and Generalization. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22-29 October 2017; pp. 5716-5726.

35. Tang, Y.; Yang, X.; Wang, N.; Song, B.; Gao, X. CGAN-TM: A Novel Domain-to-Domain Transferring Method for Person Re-Identification. *IEEE Transactions on Image Processing* **2020**, *29*, 5641-5651, doi:10.1109/TIP.2020.2985545.

36. Dai, Y.; Liu, J.; Sun, Y.; Tong, Z.; Zhang, C.; Duan, L.-Y. IDM: An Intermediate Domain Module for Domain Adaptive Person Re-ID. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10-17 October 2021; pp. 11844-11854.

37. Long, M.; Cao, Y.; Wang, J.; Jordan, M. Learning Transferable Features with Deep Adaptation Networks. In Proceedings of the 32nd International Conference on Machine Learning, 2015; pp. 97-105.

38. Hoffman, J.; Tzeng, E.; Park, T.; Zhu, J.-Y.; Isola, P.; Saenko, K.; Efros, A.; Darrell, T. CyCADA: Cycle-Consistent Adversarial Domain Adaptation. In Proceedings of the 35th International Conference on Machine Learning, Stockholmsmässan, Stockholm Sweden, 10-15 July 2018; pp. 1989-1998.

39. Gretton, A.; Borgwardt, K.M.; Rasch, M.J.; Schölkopf, B.; Smola, A. A Kernel Two-Sample Test. *Journal of Machine Learning Research* **2012**, *13*, 723–773.

40. Gao, R.; Liu, F.; Zhang, J.; Han, B.; Liu, T.; Niu, G.; Sugiyama, M. Maximum Mean Discrepancy Test is Aware of Adversarial Attacks. *arXiv* **2020**, doi:arXiv.2010.11415.

41. Zhang, W.; Ouyang, W.; Li, W.; Xu, D. Collaborative and Adversarial Network for Unsupervised Domain Adaptation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18-23 June 2018; pp. 3801-3809.

42. Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M. Domain-Adversarial Neural Networks. *arXiv* **2014**, doi:10.48550/arXiv.1412.4446.

43. Fan, H.; Yan, L.Z.C.; Yang, Y. Unsupervised Person Re-identification: Clustering and Fine-tuning. *ACM Trans. Multimedia Comput. Commun. Appl* **2018**, *14*, 1 - 10, doi:10.1145/3243316.

44. Lin, Y.; Dong, X.; Zheng, L.; Yan, Y.; Yang, Y. A Bottom-Up Clustering Approach to Unsupervised Person Re-Identification. In Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, Honolulu, Hawaii, USA, 2019; pp. 8738-8745.

45. Ge, Y.; Zhu, F.; Chen, D.; Zhao, R.; Li, H. Self-paced Contrastive Learning with Hybrid Memory for Domain Adaptive Object Re-ID. In Proceedings of the 34th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 06 December 2020; pp. 11309 - 11321.

46. Zheng, K.; Lan, C.; Zeng, W.; Zhang, Z.; Zha, Z.-J. Exploiting Sample Uncertainty for Domain Adaptive Person Re-Identification. In Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI), Virtual Event, February 2-9, 2021.

47. Akbarian, S.; Seyyed-Kalantari, L.; Khalvati, F.; Dolatabadi, E. Evaluating Knowledge Transfer in Neural Network for Medical Images. *IEEE Access* **2023**, *11*, 85812 - 85821, doi:10.1109/ACCESS.2023.3283216.

48. Chen, s.; Wang, Y.; Shi, Y.; Yan, K.; Geng, M.; Tian, Y. Deep Transfer Learning for Person Re-Identification. In Proceedings of the 2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM), Xi'an, China, 13-16 September 2018.

49. Zhao, G.; Zhang, X.; Tang, H.; Shen, J.; Qian, X. Domain-Oriented Knowledge Transfer for Cross-Domain Recommendation. *IEEE Transactions on Multimedia* **2024**, *26*, 9539 - 9550, doi:10.1109/TMM.2024.3394686.

50. Wang, Q.; Qian, X.; Li, B.; Fu, Y.; Xue, X. Image-Text-Image Knowledge Transferring for Lifelong Person Re-Identification with Hybrid Clothing States. *arXiv* **2024**, doi:10.48550/arXiv.2405.16600.

51. Passalis, N.; Tefas, A. Learning Deep Representations with Probabilistic Knowledge Transfer. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8-14 September 2018; pp. 268-284.

52. Laine, S.; Aila, T. Temporal Ensembling for Semi-Supervised Learning. In Proceedings of the International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017.

53. Zhang, Y.; Xiang, T.; Hospedales, T.M.; Lu, H. Deep Mutual Learning. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18-23 June 2018; pp. 4320-4328.

54. Zhai, Y.; Ye, Q.; Lu, S.; Jia, M.; Ji, R.; Tian, Y. Multiple Expert Brainstorming for Domain Adaptive Person Re-identification. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 594–611.

55. Tian, Y.; Zhang, W.; Zhang, Q.; Lu, G.; Wu, X. Selective Multi-Convolutional Region Feature Extraction based Iterative Discrimination CNN for Fine-Grained Vehicle Model Recognition. In Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20-24 August 2018; pp. 3279-3284.

56. Lu, L.; Yi, Y.; Huang, F.; Wang, K.; Wang, Q. Integrating Local CNN and Global CNN for Script Identification in Natural Scene Images. *IEEE Access* **2019** *7*, 52669 - 52679, doi:10.1109/ACCESS.2019.2911964.

57. He, Z.; Nan, F.; Li, X.; Lee, S.-J.; Yang, Y. Traffic Sign Recognition by Combining Global and Local Features Based on Semi-supervised Classification. *IET Intelligent Transport Systems* **2019**, *14*, 323 - 330, doi:10.1049/iet-its.2019.0409.

58. Suh, S.; Lukowicz, P.; Lee, Y.O. Fusion of Global-Local Features for Image Quality Inspection of Shipping Label. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10-15 January 2021; pp. 2643-2649.

59. Cong, R.; Yang, N.; Li, C.; Fu, H.; Zhao, Y.; Huang, Q. Global-and-Local Collaborative Learning for Co-Salient Object Detection. *IEEE Transactions on Cybernetics* *53*, 1920-1931, doi:10.1109/TCYB.2022.3169431.

60. Li, T.; Zhang, Z.; Zhu, M.; Cui, Z.; Wei, D. Combining Transformer Global and Local Feature Extraction for Object Detection. *Complex & Intelligent Systems* **2024**, 4897–4920, doi:10.1007/s40747-024-01409-z.

61. Wei, J.; Wang, S.; Huang, Q. F3Net: Fusion, Feedback and Focus for Salient Object Detection. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, New York, USA, 7-12 February 2020; pp. 12321–12328.

62. Zou, Z.; Chen, K.; Shi, Z.; Guo, Y.; Ye, J. Object Detection in 20 Years: A Survey. *Proceedings of the IEEE* **2023**, *111*, 257 - 276, doi:10.1109/JPROC.2023.3238524.

63. Zhuge, M.; Fan, D.-P.; Liu, N.; Zhang, D.; Xu, D.; Shao, L. Salient Object Detection via Integrity Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2023**, *45*, 3738-3752.

64. Yang, X.; Li, S.; Chen, Z.; Chanussot, J.; Jia, X.; Zhang, B.; Li, B.; Chen, P. An Attention-Fused Network for Semantic Segmentation of Very-High-Resolution Remote Sensing Imagery. *ISPRS Journal of Photogrammetry and Remote Sensing* **2021**, *177*, 238-262, doi:10.1016/j.isprsjprs.2021.05.004.

65. Tian, Q.; Zhao, F.; Zhang, Z.; Qu, H. GLFFNet: A Global and Local Features Fusion Network with Biencoder for Remote Sensing Image Segmentation. *Applied Sciences* **2023**, *13*, doi:10.3390/app13158725.

66. Zhou, X.; Zhou, L.; Zhang, H.; Ji, W.; Zhou, B. Local–Global Multiscale Fusion Network for Semantic Segmentation of Buildings in SAR Imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2024**, *17*, 7410 - 7421, doi:10.1109/JSTARS.2024.3379403.

67. Minaee, S.; Boykov, Y.; Porikli, F.; Plaza, A.; Kehtarnavaz, N.; Terzopoulos, D. Image Segmentation Using Deep Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2022**, *44*, 3523 - 3542, doi:10.1109/TPAMI.2021.3059968.

68. Ju, Y.; Jia, S.; Cai, J.; Guan, H.; Lyu, S. GLFF: Global and Local Feature Fusion for AI-synthesized Image Detection. *IEEE Transactions on Multimedia* **2024**, *26*, 4073 - 4085, doi:10.1109/TMM.2023.3313503.

69. Tarvainen, A.; Valpola, H. Mean Teachers Are Better Role Models: Weight-averaged Consistency Targets Improve Semi-supervised Deep Learning Results. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA, 4 - 9 December 2017; pp. 1195–1204.

70. Polyak, B.T.; Juditsky, A. Acceleration of Stochastic Approximation by Averaging. *SIAM Journal on Control and Optimization* **1992**, *30*, 838-855, doi:10.1137/0330046.

71. Zou, Y.; Yang, X.; Yu, Z.; Kumar, B.V.K.V.; Kautz, J. Joint Disentangling and Adaptation for Cross-Domain Person Re-Identification. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23-28 August, 2020; pp. 87 - 104.

72. Liu, H.; Tian, Y.; Wang, Y.; Pang, L.; Huang, T. Deep Relative Distance Learning: Tell the Difference between Similar Vehicles. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27-30 June 2016; pp. 2167-2175.

73. Gong, Y.; Huang, L.; Chen, L. Eliminate Deviation with Deviation for Data Augmentation and a General Multi-modal Data Learning Method. *arXiv* **2021**, doi:arXiv:2101.08533.

74. Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7-9 May, 2015.

75. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA, 4-9 December, 2017; pp. 6000–6010.

76. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the Computer Vision – ECCV 2018, Munich, Germany, 8–14 September, 2018; pp. 3–19.

77. Grunau, C.; Özüdoğru, A.A.; Rozhoň, V.; Tětek, J. A Nearly Tight Analysis of Greedy K-means++. In *2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*; 2023; pp. 1012-1070.

78. Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; Tian, Q. Scalable Person Re-identification: A Benchmark. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 07-13 December, 2015; pp. 1116-1124.

79. Li, W.; Zhao, R.; Xiao, T.; Wang, X. DeepReID: Deep Filter Pairing Neural Network for Person Re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23-28 June 2014; pp. 152-159.

80. Zhong, Z.; Zheng, L.; Cao, D.; Li, S. Re-ranking Person Re-identification with k-Reciprocal Encoding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21-26 July 2017; pp. 3652-3661.

81. Jin, X.; Lan, C.; Zeng, W.; Chen, Z.; Zhang, L. Style Normalization and Restitution for Generalizable Person Re-Identification. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13-19 June 2020; pp. 3140-3149.

82. Liao, S.; Shao, L. Interpretable and Generalizable Person Re-identification with Query-Adaptive Convolution and Temporal Lifting. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020.

83. Zhao, Y.; Zhong, Z.; Yang, F.; Luo, Z.; Lin, Y.; Li, S.; Sebe, N. Learning to Generalize Unseen Domains via Memory-based Multi-Source Meta-Learning for Person Re-Identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021.

84. Choi, S.; Kim, T.; Jeong, M.; Park, H.; Kim, C. Meta Batch-Instance Normalization for Generalizable Person Re-Identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021.

85. Yang, F.; Zhong, Z.; Luo, Z.; Li, S.; Sebe, N. Federated and Generalized Person Re-identification through Domain and Feature Hallucinating. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022.

86. Xu, B.; Liang, J.; He, L.; Sun, Z. Mimic Embedding via Adaptive Aggregation: Learning Generalizable Person Re-identification. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022; pp. 372–388.

87. Zhang, P.; Dou, H.; Yu, Y.; Li, X. Adaptive Cross-Domain Learning for Generalizable Person Re-Identification. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022.

88. Li, H.; Wang, Y.; Zhu, L.; Wang, W.; Yin, K.; Li, Y.; Yin, G. Weakly Supervised Cross-Domain Person Re-Identification Algorithm Based on Small Sample Learning. *Electronics* **2023**, *12*, doi:10.3390/electronics12194186.

89. Mao, Y.; Huafeng, L.; Yafei, Z. Unsupervised Domain Adaptation for Cross-Regional Scenes Person Re-identification. *J. Shanghai Jiao Tong Univ* **2023**, doi:10.16183/j.cnki.jsjtu.2023.635.

90. Sun, J.; Li, Y.; Chen, L.; Chen, H.; Peng, W. Multiple Integration Model for Single-source Domain Generalizable Person Re-identification. *J. Vis. Commun. Image Represent* **2024**, *98*, 104037, doi:10.1016/j.jvcir.2023.104037.

91. Li, Q.; Gong, S. Mitigate Domain Shift by Primary-Auxiliary Objectives Association for Generalizing Person ReID. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2024; pp. 393–402.

92. Zhao, F.; Liao, S.; Xie, G.-S.; Zhao, J.; Zhang, K.; Shao, L. Unsupervised Domain Adaptation with Noise Resistible Mutual-Training for Person Re-identification. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 526-544.

93. Dai, Y.; Liu, J.; Bai, Y.; Tong, Z.; Duan, L.-y. Dual-Refinement: Joint Label and Feature Refinement for Unsupervised Domain Adaptive Person Re-Identification. *IEEE Transactions on Image Processing* **2020**, *30*, 7815-7829.

94. Bai, Z.; Wang, Z.; Wang, J.; Hu, D.; Ding, E. Unsupervised Multi-Source Domain Adaptation for Person Re-Identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 12914–12923.

95. Wang, Y.; Liang, X.; Liao, S. Cloning Outfits from Real-World Images to 3D Characters for Generalizable Person Re-Identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 4890-4899.

96. Tay, C.-P.; Yap, K.-H. Collaborative learning mutual network for domain adaptation in person re-identification. *Neural Comput. Appl.* **2022**, *34*, 12211–12222, doi:10.1007/s00521-022-07108-5.

97. Zhou, H.; Kong, J.; Jiang, M.; Liu, T. Heterogeneous dual network with feature consistency for domain adaptation person re-identification. *Int. J. Mach. Learn. Cyber* **2023**, *14*, 1951–1965, doi:10.1007/s13042-022-01739-9.

98. Yun, X.; Wang, Q.; Cheng, X.; Song, K.; Sun, Y. Discrepant mutual learning fusion network for unsupervised domain adaptation on person re-identification. *Appl. Intell* **2023**, *53*, 2951–2966, doi:10.1007/s10489-022-03532-1.

99. Lee, G.; Lee, S.; Kim, D.; Shin, Y.; Yoon, Y.; Ham, B. Camera-Driven Representation Learning for Unsupervised Domain Adaptive Person Re-identification. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 1453–11462.

100. Rami, H.; Giraldo, J.H.; Winckler, N.; Lathuilière, S. Source-Guided Similarity Preservation for Online Person Re-Identification. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3-8 January 2024; pp. 1700-1709.

101. Zhang, B.; Wu, D.; Lu, X.; Li, Y.; Gu, Y.; Li, J.; Wang, J. A domain generalized person re-identification algorithm based on meta-bond domain alignment. *J. Vis. Commun. Image Represent* **2024**, *98*, 2913–2933, doi:10.1016/j.jvcir.2024.104054.

102. Tian, Q.; Cheng, Y.; He, S.; Sun, J. Unsupervised multi-source domain adaptation for person re-identification via feature fusion and pseudo-label refinement. *Comput. Electr. Eng* **2024**, *113*, 109029, doi:10.1016/j.compeleceng.2023.109029.

103. Bashir, R.M.S.; Shahzad, M.; Fraz, M.M. VR-PROUD: Vehicle Re-identification using Progressive Unsupervised Deep Architecture. *Pattern Recognition* **2019**, *90*, 52-65, doi:10.1016/j.patcog.2019.01.008.

104. 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 (CVPR), Long Beach, CA, USA, 15-20 June 2019; pp. 598-607.

105. Peng, J.; Wang, Y.; Wang, H.; Zhang, Z.; Fu, X.; Wang, M. Unsupervised Vehicle Re-identification with Progressive Adaptation. In Proceedings of the International Joint Conference on Artificial Intelligence, Yokohama, Japan, 7 - 15 January, 2021; pp. 913-919.

106. Wang, H.; Peng, J.; Jiang, G.; Fu, X. Learning Multiple Semantic Knowledge For Cross-Domain Unsupervised Vehicle Re-Identification. In Proceedings of the IEEE International Conference on Multimedia and Expo, Shenzhen, China, 05-09 July 2021; pp. 1-6.

107. Wang, Y.; Peng, J.; Wang, H.; Wang, M. Progressive learning with multi-scale attention network for cross-domain vehicle re-identification. *Sci. China Inf. Sci* **2022**, *65*, doi:0.1007/s11432-021-3383-y.

108. Zhang, F.; Zhang, L.; Zhang, H.; Ma, Y. Image-to-image domain adaptation for vehicle re-identification. *Multimed Tools Appl* **2023**, *82*, 40559–40584, doi:10.1007/s11042-023-14839-7.

109. Zhan, G.; Wang, Q.; Min, W.; Han, Q.; Zhao, H.; Wei, Z. Unsupervised Vehicle Re-Identification Based on Cross-Style Semi-Supervised Pre-Training and Feature Cross-Division. *Electronics* **2023**, *12*, 2931, doi:10.3390/electronics12132931.

110. Zhang, X.; Ling, Y.; Li, K.; Zhou, W.S.Z. Multimodality Adaptive Transformer and Mutual Learning for Unsupervised Domain Adaptation Vehicle Re-Identification. *Trans. Intell. Transport. Sys* **2024**, *25*, 20215 - 20226, doi:10.1109/TITS.2024.345533.

111. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22-29 October 2017; pp. 618-626.