International Journal of Computer Vision
https://doi.org/10.1007/s11263-024-02124-5

Exert Diversity and Mitigate Bias: Domain Generalizable Person Re-identification with a Comprehensive Benchmark

Bingyu Hu[1] · **Jiawei Liu**[1] · **Yufei Zheng**[1] · **Kecheng Zheng**[1] · **Zheng-Jun Zha**[1]

Received: 15 October 2023 / Accepted: 14 May 2024

Abstract

Person re-identification (ReID), aiming at retrieving persons of the same identity across non-overlapping cameras, holds immense practical significance for security and surveillance applications. In pursuit of a more general and practical solution, recent research attention has gradually shifted from the traditional single-domain ReID to the domain generalizable person re-identification (DG-ReID). However, the DG-ReID landscape lacks a meticulously designed and all-encompassing benchmark to provide a common ground for competing approaches. To this end, in this paper, we first delve into the intricate challenges of DG-ReID and introduce a comprehensive and large-scale benchmark with enhanced distributional variety and shifts to facilitate the research progress. Furthermore, in response to the highlighted challenges, a novel DG-ReID framework based on diverse feature space learning with domain factorization is proposed to effectively learn rich domain-adaptive discriminative features through the two designed blocks with fairly limited additional cost in both memory and computation. Firstly, the feature diversification block promotes a diverse feature space capable of learning domain-specific characteristics under the rich distributional variety. Secondly, the domain-adaptive shielding block applies channel-wise shielding operations based on subspace-based domain factorization in order to prevent the model from prediction bias caused by distributional shifts. Our extensive experiments demonstrate the effectiveness of the proposed framework, surpassing the performance of current state-of-the-art methods under various evaluation protocols.

Keywords Person re-identification · Domain generalization · Benchmark establish · Feature diversification · Subspace learning

1 Introduction

Person re-identification (dubbed as ReID) has drawn extensive research attention in recent years, which aims at retrieving persons of the same identity across non-overlapping cameras. Along with the success of the deep learning technique, a large amount of sophisticated ReID methods (Liu et al., 2016; Su et al., 2017; Yin et al., 2020; Zhang et al., 2020c; Ye et al., 2021; Zhu et al., 2021) have been proposed and achieved promising performances under the assumption that the training set and testing set are collected from the same domain. However, this ideal hypothesis is hardly satisfied in real applications owing to the limitation of data collection and intricacy of the scenarios. Consequently, recent efforts have been devoted to the domain generalizable person re-identification (dubbed as DG-ReID), which aims at training models using multiple source domains to enable effective generalization to unseen target domains without requiring model updates (Song et al., 2019; Jin et al., 2020; Zhao et al., 2021).

In the purpose of driving research progress and fostering innovation, it is necessary to provide a common ground for fair and rational comparisons among methods by developing well-designed benchmarks (Li et al., 2021; Zhong et

Communicated by Zhun Zhong.

✉ Jiawei Liu
jwliu6@ustc.edu.cn

Bingyu Hu
hby0728@mail.ustc.edu.cn

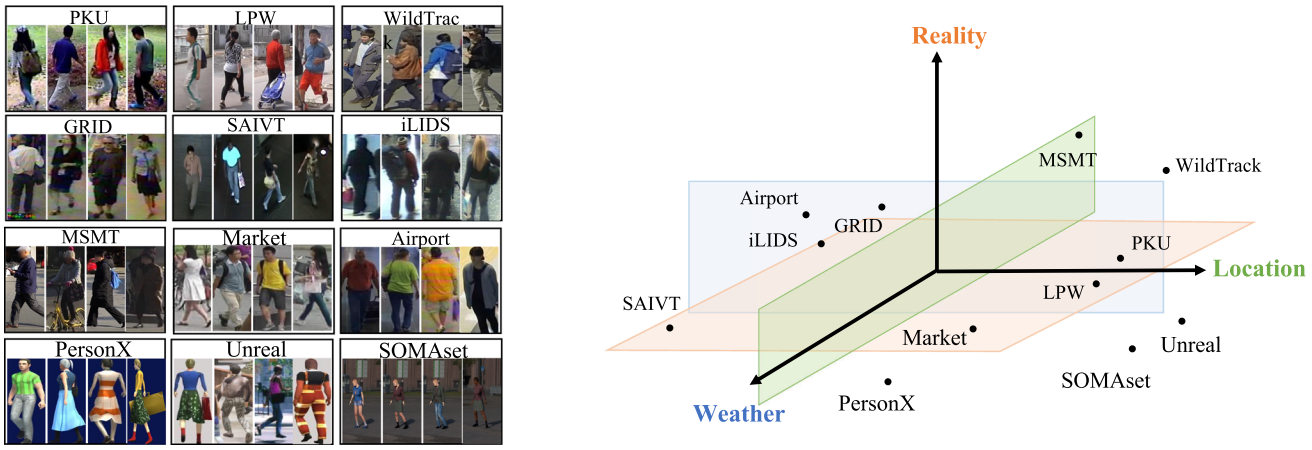
Yufei Zheng
zyf2001@mail.ustc.edu.cn

Kecheng Zheng
zkcys001@mail.ustc.edu.cn

Zheng-Jun Zha
zhazj@ustc.edu.cn

[1] University of Science and Technology of China, Hefei, China

Published online: 03 June 2024



(a) Diverse collected datasets

(b) Collected datasets characterized with three properties

Fig. 1 **a** To provide different distributions for emulating different domains, the proposed benchmark EDVARS collects twelve diverse datasets captured in various environments. **b** To simulate different distribution shifts with these datasets, we first characterize the collected datasets with three representative properties, ie., location, weather, reality, and design evaluation protocols across the two separations on each property

al., 2023). Different from the traditional DG paradigm, DG-ReID is a open-set image retrieval task that has different label spaces between source and target domains. Therefore, benchmarks for DG-ReID are methodologically constructed with two basic components: (1) Collecting ReID datasets that emulate various domains to provide different distributions. (2) Designing evaluation protocols that specify which datasets will act as source domains for training or target domains for testing to simulate distributional shifts in real scenarios. To this end, an advanced benchmark should contain a sufficient variety of distributions and deliver rational distributional shifts specified by the evaluation protocols to comprehensively evaluate DG-ReID methods.

However, existing benchmarks either exhibit a shortage of diverse domains within their composing datasets or prove difficult to simulate significant distribution shifts due to overly simplistic evaluation protocols (He et al., 2021). Specifically, the first DG-ReID benchmark (Song et al., 2019), referred as Benchmark-1 in the following, predominantly utilizes datasets collected in a single campus scene as source domains, with target domains too small-scale to fully assess the model's generalization ability (Dai et al., 2021; Xu et al., 2022). To tackle these limitations, the second benchmark, referred as Benchmark-2, is proposed, which is composed of the same collection of four medium-scale source datasets, but conducts a leave-one-out protocol that selects one domain from four large datasets for testing and all the remaining domains for training (Dai et al., 2021; Zhao et al., 2021). Nevertheless, it is still not ideal for generalization but more like a compromise to the limited domains, since the model is tested only on one specific distribution instead of multiple unseen distributions after training (Zhang et al., 2023). Additionally, ongoing inconsistencies in experimental settings have arisen due to ethical issues with some datasets being retracted from previous benchmarks. This underscores the need for a fair and common ground for competing DG-ReID approaches to drive research progress and innovation.

In order to evaluate DG-ReID methods more comprehensively by covering the mentioned shortage in existing benchmarks, we construct a new benchmark with **E**nhanced **D**istributional **VAR**iety and **S**hift (dubbed as EDVARS), with two key improvements. (1) *Diverse collected datasets*. As presented in Fig. 1a, we collected a total of twelve existing datasets, covering three quantities levels, eight collected scenes and even three synthetic datasets. The abundant diversity supports flexible assignments for training and testing, controllable degree of distributional shifts, and extensive evaluation on multiple target domains. (2) *Rational evaluation protocols*. We first characterize the collected datasets with three representative properties, ie., location, weather and reality, as shown in Fig. 1b. For each property, we group two disjoint sets of datasets having complementary traits on this property (e.g., 'indoor' versus 'outdoor') and design a pair of evaluation protocols across the two separations. In this way, we build up the artificial distinction between source domains and target domains to better deliver practical challenges where a trained model may encounter any possible test data. A detailed comparison between the previous benchmarks and our EDVARS is shown in Table 1.

When evaluating on the comprehensive benchmark, current DG-ReID models are either challenged by the great distributional variety to learn the domain-invariant features

Table 1 Comparison between previous benchmarks and the proposed benchmark EDVARS

Benchmark	# Images	# Identities	# Cameras	# Scenes	# Synthetic datasets	# Evaluation protocols	Elaborate distribution shifts
Benchmark-1	89,300	19,424	19	4	0	1	No
Benchmark-2	204,531	19,903	24	3	0	3	No
EDVARS	**763,728**	**22,983**	**98**	**8**	**3**	**5**	**Yes**

Best statistics are highlighted in bold

or trapped in the domain-level composition caused by the significant distribution shifts. In detail, existing DG-ReID methods can be mainly divided into two categories: single model methods (Choi et al., 2021; Jia et al., 2019; Zhuang et al., 2020; Jin et al., 2020; Choi et al., 2021) and ensemble learning based methods (Xu et al., 2022; Dai et al., 2021). The former collects all source domain data and trains a single model on them to extract the shared domain-invariant representations, which is limited when confronted with diverse distributions. Besides, these methods discard the diverse domain-specific characteristics, which can provide discriminative and meaningful information, thus resulting in unsatisfactory generalization capability (Dai et al., 2021; Xu et al., 2022). The latter train domain-specific models (e.g., branches, classifiers, or experts) for each source domain and combine these domain-specific models to enhance the generalization ability. However, the coarse-grained domain-level composition limits the flexibility in modeling target domains, and the per-domain network as well as the meta-learned combination strategies increase unscalable memory and computation cost proportional to the number of source domains. Therefore, it still remains challenging for ReID models to deal with both distributional variety and distributional shift.

In correspondence to the unique challenges identified by EDVARS, we propose a novel **D**iverse **F**eature space learning with **D**omain **F**actorization approach (dubbed as DF^2) for a more effective and efficient DG-ReID system, which belongs to the single model based methods but is capable of capturing domain-adaptive feature by exploiting domain-specific characteristics. It improves the existing approaches mentioned above from the following two perspectives. First of all, considering that diverse distributional variety brings diverse domain-related information, DF^2 promotes the diversity of the common feature space to cover as much discriminative information as possible for all domains. Second, to make the feature adaptive to heterogeneous domain characteristics caused by distributional shift, DF^2 models each domain with underlying factors and builds the relationship between the feature discriminativeness with factors. In this way, DF^2 can accurately perceive the inconsistency between the target domain and the source domains on the same factors, so as to extract domain-adaptive discriminative representations.

In detail, to promote the diversity of the feature space, the Feature Diversification Block (dubbed as FDB) is designed, which drives the independence and complementarity between feature channels so as to learn more diverse useful information. Specifically, we propose the Instance-Batch Whitening (dubbed as IBW) operation to conduct channel decorrelation with merely small computational overhead. Besides, a diversity loss is designed to encourage different channels focusing on diverse spatial locations. To prevent the model from extracting inconsistent domain-related information due to characteristic discrepancy, we establish a systematic bias shielding mechanism by proposing the Domain-adaptive Shielding Block (dubbed as DSB). Instead of dealing with domains individually at a coarse-grained level, we break them into underlying domain-aware factors with subspace-based learning technique. Each source domain is factorized into a low-dimensional subspace in the common feature space, which is spanned by orthogonal basis vectors trained via sophisticated losses. The projection distances of a feature map on the bases reflect how each channel reacts on specific underlying domain-aware factors. Along this line, we design a channel-wise shielding strategy to disable those channels that have inconsistent activation. The proposed FDB and DSB together make up all the innovations of our DF^2 framework. Compared to previous methods (Dai et al., 2021; Choi et al., 2021; Xu et al., 2022), DF^2 keeps a good balance of generalization ability and model complexity because it does not need to train multiple expert models or conduct meta-learning strategy, reducing both memory and computation cost to a large extent. Extensive experiments have demonstrated the effectiveness of the proposed method. Our contributions can be summarized as follows:

- To facilitate the progressive research for DG-ReID, we propose a large-scale benchmark named EDVARS, which highlights the challenges from the distributional variety and the distributional shifts in DG-ReID with diverse collected datasets and rational evaluation protocols.
- We propose a novel Diverse Feature space learning with Domain Factorization approach (DF^2) to learn a well-generalized ReID model with low additional cost on memory and computation.

- We design a Feature Diversification Block (FDB) to drive the independence and complementarity between feature channels for promoting more diverse and discriminative feature learning.
- We design a Domain-adaptive Shielding Block (DSB), which factorizes each domain as a subspace in the common feature space and achieves channel-wise inconsistency shielding based on the projection on each subspace.
- Extensive experiments demonstrate the effectiveness of our framework, which surpasses state-of-the-art methods under various evaluation protocols.

2 Related Work

2.1 Benchmarks for DG-ReID

Benchmarks drive research progress by providing a common ground for competing approaches, fostering innovation and enabling fair comparisons among methods. Song et al. (2019) proposed the first large-scale DG-ReID benchmark that using existing large-scale ReID datasets, ie., CUHK02 (Li & Wang, 2013), CUHK03 (Li et al., 2014), Market-1501 (Zheng et al., 2015a), DukeMTMC-ReID (Zheng et al., 2017) and CUHK-SYSU (Xiao et al., 2016) to form the source domains, and the smaller ones, ie., VIPeR (Gray & Tao, 2008), PRID (Hirzer et al., 2011), GRID (Loy et al., 2013) and i-LIDS (Cai et al., 2010) as target domains. Later, arguing that the image quality of the small-scale Re-ID datasets is quite poor, Zhao et al. (2021) and Dai et al. (2021) proposed a new benchmark, which conducted the leave-one-out setting on four large-scale datasets, ie., CUHK03, Market-1501, DukeMTMC-ReID and MSMT17 (Wei et al., 2018). However, DukeMTMC-ReID has been withdrawn by its creators due to its privacy issues, which is widely used in the previous benchmarks. Thus Xu et al. (2022) and Zhang et al. (2022) revised the benchmarks by deleting it in the first benchmark with four source datasets left and replacing it with CUHK-SYSU in the second benchmark. As the CUHK-SYSU dataset only contains 1 camera, it is not used for testing. Most of the datasets in existing benchmarks are collected on the campus, which are not enough to cover the complexity in reality. Besides, the rationale behind designed evaluation protocols is not clear, lacking comprehensive evaluations of DG-ReID methods.

2.2 Methods for DG-ReID

Generalization capability to unseen domains is crucial for ReID models when deploying to practical applications (Wang et al., 2018; Liu et al., 2019a; Huang et al., 2021). To address this problem, several tailored methods (Jin et al., 2020; Choi et al., 2021; Zhao et al., 2021; Jia et al., 2019; Zhuang et al., 2020; Zhao et al., 2021) have been proposed. To deal with the domain discrepancy, batch normalization (BN) (Ioffe & Szegedy, 2015) and instance normalization (IN) (Ulyanov et al., 2016) have been widely explored to improve generalization capability by statistical feature regularization. For example, Jin et al. (2020) proposed a style normalization and restitution module, which utilizes the IN layers to filter out style variations and compensates for the identity-relevant features discarded by IN layers. Zhuang et al. (2020) proposed camera-based batch normalization (CBN) to force the images of all cameras to fall onto the same subspace and to shrink the distribution gap between any camera pair. However, they have ignored that individual domains' discriminative characteristics are able to provide complementary information for better generalization on target domains in the open-set DG-ReID task. Consequently, recent methods (Dai et al., 2021; Jiao et al., 2022; Zhang et al., 2022; Xu et al., 2022) focus more on modeling domain-specific features. Typically, these methods are usually implemented by designing a specific network for each domain and then aggregating them through modeling the relevance between seen source domains and unseen target domains by a domain-aware voting network or adapter. For example, Dai et al. (2021) proposed to train an expert for each source domain, and designed a voting network for integrating multiple experts. Xu et al. (2022) designed specific BN layers for each domain and aggregated them by calculating the similarity between the IN/BN statistics of different domains. However, the representation capacity is bounded by the number of source domains, thus constraining its capacity for generalization, particularly when confronted with substantial distribution shifts. Besides, due to the lack of target training data, these methods are optimized by the meta-learning strategy, which requires twice forward and backward passes for each update step, resulting in expensive computation and doubled training time.

In addition, recent works begin to investigate the generalization ability of the vision transformer (Dosovitskiy et al., 2021). TransMatcher (Liao & Shao, 2021) employs hard attention to cross-matching similarity computing, which is more efficient for image matching. However, it still uses CNN as the main feature extractor, and the role of the transformer is mainly reflected in image matching. PAT (Ni et al., 2023) is the first to investigate the generalization ability of pure transformer in DG-ReID, which designs a proxy task to mine local visual information shared by different IDs. Besides, the latest transformer-based foundation model (Chen et al., 2023) utilizes large-scale unsupervised training, which serves as a stronger pre-trained model for ReID methods. Consequently, in addition to adopting the ResNet-50 pre-trained on ImageNet-1K as the backbone, we also includes these recent pre-trained models to facilitate future research efforts.

3 Proposed Benchmark

In the field of computer vision, benchmarks play a critical role in promoting the advance of research (Deng et al., 2009). On the one hand, it is necessary to provide a common ground for comparing the performance of various proposed methods. On the other hand, there is a close relationship between the construction process of the benchmark and the goal of the studied problem. Specifically, a benchmark should directly reflect the most attractive challenges we expect the evaluated methods to overcome in real-world applications, so that the evaluated methods can be reliably put into practice. For example, miniImageNet (Vinyals et al., 2016) was purpose-built from the full ImageNet (Deng et al., 2009) for promoting the image classification models to overcome the challenges of real-life few-shot scenarios. Therefore, rationally designed benchmarks can actually drive research progress and foster innovation. In this section, we first give a thorough discussion on the construction principles of benchmarks in the DG-ReID task. Then, we give a detailed description of the proposed benchmark.

3.1 Construction Principles

A general principle to construct benchmarks is to simulate the paradigm of real-world applications as closely as possible. For the DG-ReID task, there are two key points. First, there must exist a variety of available domains, each of which has a unique data distribution and label space. Second, the training and testing of models are typically set upon two disjoint groups of domains, namely source domains and target domains, respectively. Consequently, DG-ReID benchmarks should meet the following requirements: (1) *Diverse distributional variety.* A large group of existing ReID datasets are collected to act as different domains, which should contain diverse distribution to cover typical potential scenarios. (2) *Rational evaluation protocols.* Evaluation protocols are designed to specify which datasets will act as source domains during training or target domains for testing. The simulated distribution shifts should meet the needs of various real-world applications where a trained model may encounter any possible test data.

3.2 The EDVARS Benchmark

To promote the advance of DG-ReID research, we propose a new benchmark with **E**nhanced **D**istributional **VAR**iety and **S**hift, namely **EDVARS** for short. The key point is that, in the purpose of having more elaborate control over the data distribution, we need to characterize each domain (ie., dataset) via some properties so that we can artificially judge whether the data distribution between different domains varies significantly. For ReID datasets, it is the properties of the data collection environment that greatly influence the underlying data distribution. Specifically, we select three representative properties that generally exist in current ReID datasets and have an explicit influence on data distribution.

- **Location.** Existing ReID datasets are captured in different locations, including airport (Cai et al., 2010), campus (Zheng et al., 2015a), buildings (Bialkowski et al., 2012), etc., which in turn influence the data distribution. On the whole, we divide them by whether the shooting location is indoor or outdoor, where the main variations depend on the lighting conditions. As shown in Fig. 2a, we give the comparison between some identities all wearing black clothes in the outdoor datasets (above) and indoor datasets (below). We can find that it is harder to isolate individuals from the background in the indoor environments.

Fig. 2 Comparisons between the images taken in different environments. **a** The images above are taken in the outdoor locations and the below are taken in the indoor locations. **b** The images above are taken in Winter and the below are in Summer. **c** The above are synthetic images and the below are real images

- **Weather.** After carefully analyzing the existing datasets, we found that they were all collected in a certain period of season, resulting in significant differences in the clothing of pedestrians photographed due to different weather conditions. As illustrated in Fig. 2b, pedestrians in winter wear heavier clothing, which results in more obstruction of the pedestrian's posture as well as rougher outlines. Besides, in winter, people often wear hats, scarves and other accessories, which are discriminative clues for person re-identification (Liu et al., 2019b; Lin et al., 2019; Zhang et al., 2020).
- **Reality.** Since data annotation in ReID is difficult and costly, there is a series of research on training models with images of synthetic identities instead of real people (Sun & Zheng, 2019; Zhang et al., 2021). Consequently, it is worth taking the reality property of datasets into consideration, which obviously influences the overall style of the samples as shown in Fig. 2c.

Along this line, our improvements towards existing benchmarks are two-folds. To enhance distributional variety, we adopt more diverse datasets with various collection environment properties, ie., location, weather and reality. To make proper distributional drift, we design several evaluation protocols in which the properties of source domains have deliberate distinction from those of target domains. The rest of this section gives a thorough description towards the proposed EDVARS benchmark, and more analysis and discussion will be presented in Sect. 5.

3.2.1 Diverse Collected Datasets

For the purpose of providing sufficient distributional variety, EDVARS is comprised of 12 diverse datasets as shown in Table 2, which are carefully collected with the following rules. First of all, to minimize the impact of dataset size on generalization evaluation, EDVARS encompass datasets across three scales for each property: small-scale (<10K), medium-scale (≈ 30K), and large-scale (>100K) image quantities. Furthermore, to cover the variety of reality, EDVARS makes the first incorporation of synthetic datasets, ie., SOMAset (Barbosa et al., 2018), Unreal (Zhang et al., 2021) and PersonX (Sun & Zheng, 2019), whose scales are of the same order of magnitude as the real-world dataset, ie., MSMT17 (Wei et al., 2018). Then, EDVARS includes a range of real-world datasets captured in various locations and periods of seasons. The main rules to select these datasets are listed below.

- Collecting places are distinct and diverse. For example, PRID (Hirzer et al., 2011) and VIPeR (Gray & Tao, 2008) were usually utilized in the previous benchmarks, while they were not selected by the proposed benchmark because of their unknown collection places.
- Collection locations are as diverse as possible. Unlike previous benchmarks (Dai et al., 2021; Zhao et al., 2021), which predominantly featured campus scenes, the selected datasets encompass a wide range of scenes including campuses, streets, open areas, airports, underground stations, etc.
- The camera information is clear and complete. For instance, CUHK-SYSU (Xiao et al., 2016) were not included, because its images were taken from two sources, i.e., street and movie, which makes it difficult to divide as real or virtual dataset. It is precisely because of this that it does not have clear camera information, which prevents it from dividing the training and testing set (Zhang et al., 2022; Xu et al., 2022).
- Images are normal and complete. Aerial imagery datasets, eg., PRAI (Zhang et al., 2020) and partial datasets, eg., PartialReID (Zheng et al., 2015b) and OccludedReID (Miao et al., 2019) were not included, which will be more useful for studying specific challenged problems in future works, such as generalizing from general images to aerial images.
- Datasets have not been retracted due to the privacy issues, such as DukeMTMC-ReID (Zheng et al., 2017).

In summary, the proposed benchmark incorporates diverse dataset properties by location, weather and reality, offering adequate distributional variety and enabling flexible evaluation protocol design in the following section.

3.2.2 Elaborate Evaluation Protocols

Despite the fact that the widely adopted evaluation methods in DG-ReID effectively show the generalization ability of models to the unseen target domain, they fail to sufficiently simulate real scenarios in application. For example, the most popular evaluation method, namely leave-one-out evaluation conducted in Benchmark-2 (Dai et al., 2021; Xu et al., 2022), tests models on a single target domain for each training process, while in real applications, a trained model is required to be reliable under any possible scenarios with various data distributions. The compromise on the limitation of domain numbers in current benchmarks can be addressed by EDVARS with sufficient domains. The superiority supports designing more realistic evaluation metrics to test models' generalization ability comprehensively. In addition, little attention has been paid to making a deliberate distinction between the designated source domains and target domains which is likely to appear in real applications.

To overcome these limitations, we have devised three kinds of distinct evaluation protocols, each aligned with the key properties of the datasets discussed above: loca-

Table 2 Collected datasets in the proposed benchmark for DG-ReID

	Datasets	Collected scenes	# Identities	# Cameras	# Total Images
Small-scale	iLIDS (Cai et al., 2010)	Airport arrival hall	119	2	476
	GRID (Loy et al., 2013)	Underground station	1025	8	1275
	PKU (Ma et al., 2016)	Campus	114	2	1824
	SAIVT (Bialkowski et al., 2012)	Buildings	152	8	7150
Medium-scale	Market-1501 (Zheng et al., 2015a)	Campus	1501	6	29,419
	LPW (Song et al., 2018)	Street	2731	4	30,678
	WildTrack (Chavdarova et al., 2018)	An open area	313	7	33,979
	Airport (Gou et al., 2018)	Airport	9651	6	39,902
Large-scale	SOMAset (Barbosa et al., 2018)	Synthetic	50	–	100,000
	Unreal (Zhang et al., 2021)	Synthetic	1960	34	119,128
	MSMT17 (Wei et al., 2018)	Indoor and outdoor	4101	15	126,441
	PersonX (Sun & Zheng, 2019)	Synthetic	1266	6	273,456

tion, weather, and reality. Each of them simulates the specific real-world generalization situation and provides a more comprehensive assessment of model performance together.

- **Generalization across location.** We first separate the real-world datasets into two opposite groups, ie., Indoor (GRID, SAIVT, iLIDS, Airport) and Outdoor (PKU, Market1501, LPW, WildTrack). Then we designate one group as source domains and the other as target domains (e.g., *Indoor*→*Outdoor*), vice versa (e.g., *Outdoor*→*Indoor*).
- **Generalization across weather.** We select the real-world datasets that have obvious seasonal attributes and divide them into two group, ie., Summer (SAIVT, Market1501) and Winter (WildTrack, MSMT17). Next we designate one group as source domains and the other as target domains (e.g., *Summer*→*Winter*), and vice versa (e.g., *Winter*→*Summer*)
- **Generalization across reality.** Last but no least, we build the protocols to evaluate the generalization ability from synthetic to real-world datasets. To be more comprehensive, the selected target domains cover small-scale, medium-scale and large-scale datasets simultaneously.

In summary, we design five evaluation protocols covering three types of distribution drifts in the proposed EDVARS benchmark for comprehensively evaluating the competing approaches. The detailed training/testing information is presented in Table 3. In this way, we can not only artificially construct distribution shifts between source domains and target domains, but also observe the impact of specific dataset properties on model generalization to some extent, which will be detailed in Sect. 5. As far as we know, it is the first time to explore specific challenges posed by different distribution shifts.

4 Proposed Methodology

The proposed benchmark sets up a reasonable target for designing DG-ReID methods. Especially, the highlighted challenges lie in how to learn effective domain-specific characteristics under diverse distributional variety while preventing from extracting inconsistent domain-related information due to unforeseen distributional shifts. Moreover, the solution should be efficient and scalable as the number of domains grows. To achieve these goals, we propose a novel Diverse Feature space learning with Domain Factorization approach (DF^2) as depicted in Fig. 3. The overall innovation of DF^2 involves two types of designed blocks, ie., the Feature Diversification Block (FDB) and the Domain-adaptive Shielding Block (DSB). Firstly, the FDBs, plugged among the backbone blocks, aim at diversifying the feature space and promoting rich representation capacity for covering as much discriminative information as possible for all domains. Secondly, the DSB factorizes each source domain into a low-dimensional subspace spanned by learned basis vectors, which are supposed to represent underlying factors embedded with information about domain-specific characteristics. Then, given a fully diversified feature map, its projection distances on the factorized bases indicate how each channel reacts to specific underlying domain-aware factors. Finally, DSB outputs a channel-wise mask through a designed domain-adaptive strategy to shield the channels having inconsistent activation between the target domain and the source domains on the factors. In the rest of this section, we will introduce the components of DF^2 framework in detail, ie., the Feature Diversification Block and Domain-adaptive Shielding Block.

Table 3 Details of the proposed evaluation protocols

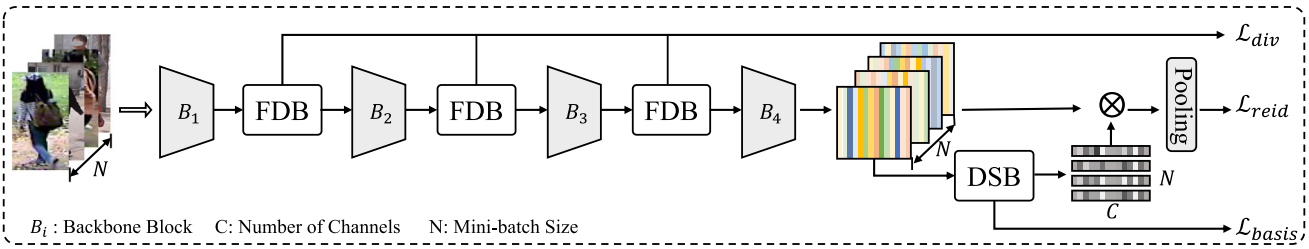
	Source			Target				
	Dataset	# Train ID	# Train images	Dataset	# Pr. IDs	# Ga. IDs	# Pr. imgs	# Ga. Imgs
Generalization across location								
Protocol-1	iLIDS	119	368	PKU	57	57	57	855
	SAIVT	152	7150	LPW	756	756	3013	4277
	GRID	1025	1275	WildTrack	157	152	157	15,862
	Airport	9652	39,902	Market1501	750	751	3368	15,913
Protocol-2	PKU	114	1824	iLIDS	60	60	60	60
	Market1501	1501	29,419	GRID	125	900	125	900
	LPW	2731	30,678	SAIVT	76	76	76	3525
	WildTrack	313	33,979	Airport	1003	1382	2264	6400
Generalization across weather								
Protocol-3	SAIVT	152	7150	WildTrack	157	152	157	15,862
	Market1501	1501	29,419	MSMT17	3060	3060	11,659	82,161
Protocol-4	WildTrack	313	33,979	SAIVT	76	76	76	3525
	MSMT17	4101	126,441	Market1501	750	751	3368	15,913
Generalization across reality								
Protocol-5	SOMAset	50	100,000	iLIDS	60	60	60	60
	Unreal	1960	119,128	GRID	125	900	125	900
	PersonX	1266	273,456	Market1501	750	751	3368	15,913
				MSMT17	3060	3060	11,659	82,161

4.1 Feature Diversification Block

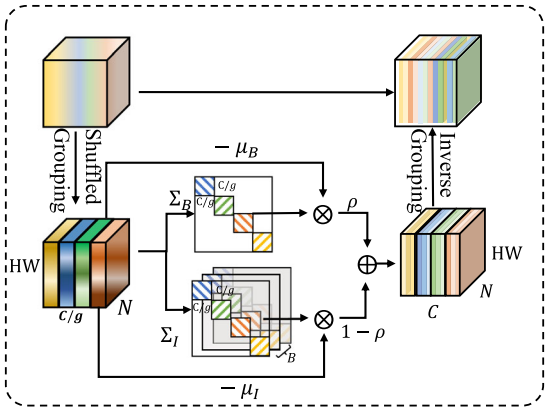
We first introduce the design of Feature Diversification Block (FDB), which aims at maximizing the utilization of the model's representation capacity and promoting a diverse feature space for learning various domain-specific characteristics. Our motivation lies in the observation that existing single model based methods are prone to output highly correlated features when forced to be compatible with multiple source domains. The highly correlated features focus on the small local areas, such as the most frequent shapes or colors in the training data. However, when the limited areas are not discriminative characteristics for the target domain, the highly correlated features may result in incorrect predictions. As shown in Fig. 4, the training images are from three source domains that all carry a 'backpack'. We visualize the feature map extracted by the strong baseline MetaBIN (Choi et al., 2021) and find that the model only focuses on the 'backpack'. However, when testing on the target domain, if 'backpack' is not a discriminative characteristic, the prediction will be incorrect. In a nutshell, the high correlation between feature channels is prone to lead the model to rely on limited knowledge and impede its generalization ability. Moreover, the correlation also inhibits the channels' potential to learn more domain-aware factors. Consequently, we design the Feature Diversification Block (FDB) to promote feature decorrelation and encourage the model to capture diverse factors, avoiding being biased to limited knowledge. The core design of FDB is the Instance-Batch Whitening (IBW) together with a diversity loss.

Feature channels decorrelation. Whitening transformation is a well-adopted technique that aims to transform the input features to have zero means and unit variances, and remove the correlation between channels (Huang et al., 2018; Siarohin et al., 2018; Pan et al., 2019; Cho et al., 2021). Motivated by the success of a mixture of batch normalization (BN) and instance normalization (IN) in the DG-ReID field, we propose a novel Instance-Batch Whitening (IBW) module, which combines the Batch Whitening (BW) and Instance Whitening (IW) with a learnable balancing parameter to decorrelate the feature channels based on the statistics of channel dependency with respect to both the mini-batch and each sample. Formally, let $\boldsymbol{X} \in \mathbb{R}^{C\times NHW}$ be the data matrix of a mini-batch, where N, C, H, W indicate the mini-batch size, number of channels, height, and width, respectively. The n-th sample in the mini-batch is denoted as $\boldsymbol{X}_n \in \mathbb{R}^{C\times HW}$, where $n \in \{1, 2, \ldots, N\}$. The designed IBW module can be formulated as follows:

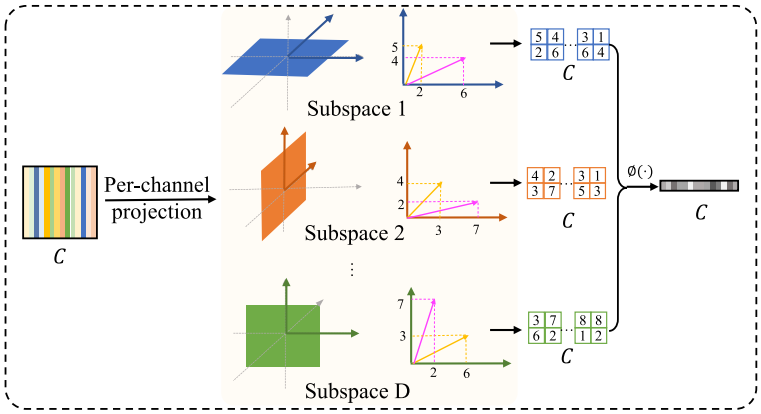
$$IBW(\boldsymbol{X}_n) = \rho BW(\boldsymbol{X}_n) + (1-\rho) IW(\boldsymbol{X}_n), \tag{1}$$



(a) Overall Flowchart of the Proposed Diverse Feature Space Learning with Domain Factorization Approach.



(b) Feature Diversification Block (FDB)



(c) Domain-adaptive Shielding Block (DSB)

Fig. 3 **a** Overall flowchart of the proposed diverse feature space learning with domain factorization framework. The proposed framework consists of the **b** Feature Diversification Block (FDB) and **c** Domain-adaptive Shielding Block (DSB), which are plugged between the backbone blocks and after the last block, respectively. The proposed FDB conducts feature channel decorrelation with a diversity loss to enhance the model representation capacity. And the proposed DSB builds a subspace for each domain and projects the diverse feature map into each subspace to obtain the channel-wise shielding weights, which in turn are multiplied on the original feature map

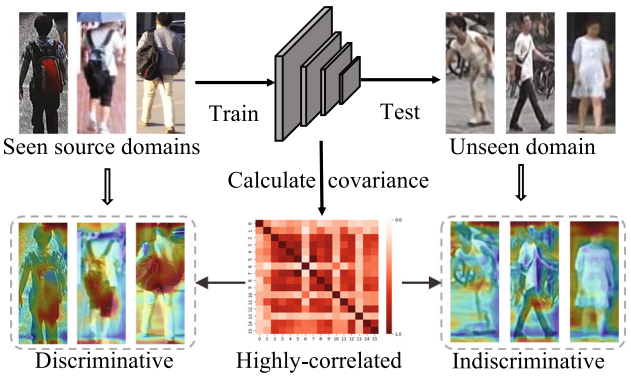


Fig. 4 Existing single-model based methods tend to lean highly-correlated features that focus on the most frequent local areas, such as the 'backpack' shown in the figure, resulting in limited predictions

where:

$$BW(\boldsymbol{X}_n) = \boldsymbol{\Sigma}_B^{-\frac{1}{2}} \cdot (\boldsymbol{X} - \boldsymbol{\mu}_B \cdot \mathbf{1}^T), \tag{2}$$

$$IW(\boldsymbol{X}_n) = \boldsymbol{\Sigma}_I^{-\frac{1}{2}} \cdot (\boldsymbol{X} - \boldsymbol{\mu}_I \cdot \mathbf{1}^T), \tag{3}$$

where ρ is a learnable balancing parameter, $\mathbf{1}$ represents a column vector of ones, $\boldsymbol{\Sigma}$ and μ are the covariance matrix and mean of the input mini-batch/sample, respectively, ie.,

$$\boldsymbol{\mu}_B = \frac{1}{NHW}\boldsymbol{X} \cdot \mathbf{1}, \tag{4}$$

$$\boldsymbol{\Sigma}_B = \frac{1}{NHW}(\boldsymbol{X} - \boldsymbol{\mu}_B \cdot \mathbf{1}^T)(\boldsymbol{X} - \boldsymbol{\mu}_B \cdot \mathbf{1}^T)^T + \epsilon \boldsymbol{I}, \tag{5}$$

$$\boldsymbol{\mu}_I = \frac{1}{HW}\boldsymbol{X}_n \cdot \mathbf{1}, \tag{6}$$

$$\boldsymbol{\Sigma}_I = \frac{1}{HW}(\boldsymbol{X}_n - \boldsymbol{\mu}_I \cdot \mathbf{1}^T)(\boldsymbol{X}_n - \boldsymbol{\mu}_I \cdot \mathbf{1}^T)^T + \epsilon \boldsymbol{I}, \tag{7}$$

where $\epsilon > 0$ is a small positive number to prevent singular covariance. The decorrelation requires computing $\boldsymbol{\Sigma}^{-\frac{1}{2}}$, which usually employs eigen-decomposition or SVD involving heavy computations (Huang et al., 2018; Siarohin et al., 2018). To reduce the complexity and ease the inefficiency problem, we conduct two strategies from the following two aspects.

On the one hand, we leverage approximate calculation to accelerate the computation procedure. Specifically, following the previous methods (Huang et al., 2019; Pan et al., 2019), we use Newton's Iteration (Bini et al., 2005) to obtain

$\mathbf{\Sigma}^{-\frac{1}{2}}$. Given a covariance matrix $\mathbf{\Sigma}$, Newton's Iteration calculates $\mathbf{\Sigma}^{-\frac{1}{2}}$ by following the iterations:

$$\mathbf{\Sigma}^{-\frac{1}{2}} = \frac{\mathbf{\Sigma}_T}{\sqrt{tr(\mathbf{\Sigma})}}, \tag{8}$$

where T is the iteration number and set as 3 in our work, and $\mathbf{\Sigma}_T$ is calculated iteratively as:

$$\begin{cases} \mathbf{\Sigma}_0 &= \boldsymbol{I}, \\ \mathbf{\Sigma}_k &= \frac{1}{2}(3\mathbf{\Sigma}_{k-1} - \mathbf{\Sigma}_{k-1}^3\mathbf{\Sigma}), k = 1, 2, \ldots, T. \end{cases} \tag{9}$$

Note that the convergence of Eq. (9) is guaranteed if $||\boldsymbol{I} - \mathbf{\Sigma}|| < 1$ (Bini et al., 2005). To satisfy the condition, $\mathbf{\Sigma}$ is normalized by $\mathbf{\Sigma} = \mathbf{\Sigma}/tr(\mathbf{\Sigma})$ (Huang et al., 2019).

On the other hand, we attempt to reduce the overall computation task by grouping operations. Similar to the previous methods (Huang et al., 2018; Cho et al., 2019), we divide the feature channels into groups and perform IBW within each group. Particularly, we add a shuffle operation to reduce the impact of group division. In detail, the channels of the input feature are first randomly shuffled and then evenly grouped into several average groups along the channel dimension. Next the proposed instance-batch whitening operation is performed within each group of feature. After that, the inverse grouping operation and inverse shuffling operation are conducted to restore channel arrangement. To sum up, the total process of the feature channels decorrelation is described as follows:

$$\textit{Shuffled Grouping: } \{\boldsymbol{X}_n^i\}_{i=1}^G = \mathcal{G}(\mathcal{S}(\boldsymbol{X}_n)), \tag{10}$$

$$\textit{Instance-Batch Whitening: } \tilde{\boldsymbol{X}}_n^i = IBW(\boldsymbol{X}_n^i), \tag{11}$$

$$\textit{Inverse Grouping: } \tilde{\boldsymbol{X}}_n = \mathcal{S}^{-1}(\mathcal{G}^{-1}(\{\tilde{\boldsymbol{X}}_n^i\}_{i=1}^G)), \tag{12}$$

where $\mathcal{S} : \mathbb{R}^{C\times HW} \rightarrow \mathbb{R}^{C\times HW}$ denotes the randomly shuffling operation. $\mathcal{G} : \mathbb{R}^{C\times HW} \rightarrow \mathbb{R}^{G\times g\times HW}$ denotes the grouping operation, where $C = Gg$, G is the number of groups and g is the number of channels in each group, i.e., group size. $\boldsymbol{X}_n^i \in \mathbb{R}^{g\times HW}$ is the i-th group of feature. $\mathcal{G}^{-1}$ and $\mathcal{S}^{-1}$ are the inverse grouping and shuffling operations.

Diversity loss. Along with the feature channels decorrelation module mentioned above, we further propose a diversity loss for encouraging different channels focusing on diverse spatial locations. First of all, we conduct the spatial softmax to normalize each channel:

$$SM_c(\tilde{\boldsymbol{X}}_n) = \frac{exp(\tilde{\boldsymbol{X}}_n^{hw})}{\sum_{hw=1}^{HW} exp(\tilde{\boldsymbol{X}}_n^{hw})}, \tag{13}$$

where $c \in \{1, 2, \ldots, C\}$. By fixing the maximum sum of all feature maps within each channel as 1, we establish a consistent and controlled distribution of attention across spatial locations. This method enhances the magnitudes of selected pixels while suppressing the influence of pixels in other locations, effectively transforming each channel into an 'expert' for the selected pixel. Subsequently, we calculate the activation degree of each pixel by averaging its Top-K representative channels:

$$AD(\tilde{\boldsymbol{X}}_n) = \frac{1}{K}\sum_{k=1}^{K} \underset{c=1,2,\ldots,C}{Top\text{-}k} (SM_c(\tilde{\boldsymbol{X}}_n)) \tag{14}$$

Finally, the diversity loss is designed to to maximize the average magnitude, a strategic step aimed at encouraging a more diverse focus on spatial locations. This approach enables us to optimize the attention and magnitudes assigned to pixels, ultimately contributing to forcing different channels to pay their most salient attention to different spatial locations:

$$\mathcal{L}_{div} = -\lambda_{div}\frac{1}{HW}\sum_{hw=1}^{HW} AD(\tilde{\boldsymbol{X}}_n), \tag{15}$$

where λ_{div} is the hyper-parameter to balance the loss.

4.2 Domain-Adaptive Shielding Block

After feature diversification, a common feature space capable of extracting rich information from all source domains has been constructed so that we can utilize domain-specific characteristics for prediction. However, due to the existence of distributional shifts in DG-ReID, directly utilizing all the domain-specific characteristics is prone to suffer from inconsistency between source domains and target domains. To tackle this issue, we design the Domain-adaptive Shielding Block (DSB) comprised of a *project layer* and a *shield layer*, which can be formalized as follows.

Input. Let $\boldsymbol{X} \in \mathbb{R}^{C\times HW}$ be a sample of a mini-batch, where C, H, W indicate the number of channels, height, and width, respectively. For convenience, we denote $\boldsymbol{X}[c] \in \mathbb{R}^{HW}$ as the c-th channel in the feature map, $c \in \{1, \ldots, C\}$.

Project layer. The project layer consists of D subspaces, each of which is spanned by M basis vectors, ie., $\boldsymbol{B} \in \mathbb{R}^{D\times M\times L}$, where D equals to the number of source domains, the hyperparameter M indicates the number of hidden factors, and length of basis vectors L is the same as the length of channels HW. Given $\boldsymbol{X}$ as input, the project layer calculates the channel-wise projection distance to each subspace, ie., $\boldsymbol{P} \in \mathbb{R}^{C\times D\times M}$ as follows:

$$\boldsymbol{P}[c, d, m] = \boldsymbol{X}[c]^T \cdot \boldsymbol{B}[d, m], \tag{16}$$

where $\boldsymbol{P}[c, d, m]$ indicates the projection distance of the c-th channel on the m-th basis of the d-th subspace.

Shield layer. The shield layer takes the projection results as input and generates channel-wise weights by a designed domain-adaptive strategy, ie., $\boldsymbol{W} = \{w_c\}_{c=1}^{C} \in \mathbb{R}^C$.

$$\boldsymbol{W} = \phi(\boldsymbol{P}), \tag{17}$$

where the weight generation function $\phi(\cdot)$ will be thoroughly discussed later.

Output. Finally, the output feature map of DSB, $\hat{\boldsymbol{X}}$, is obtained by re-weighting the original feature map $\boldsymbol{X}$, along the channel dimension, ie.,

$$\hat{\boldsymbol{X}}[c] = w_c \boldsymbol{X}[c], c \in \{1, 2, \ldots, C\}. \tag{18}$$

In the following, we will describe (1) how to construct the domain subspace (ie., how to train the basis vectors $\boldsymbol{B}$) and (2) how to obtain the channel-wise weights to achieve domain-adaptive shielding (ie., how to design the weight generation function $\phi(\cdot)$).

4.2.1 Subspace Construction

The whole procudure of subspace constrcution is shown in (Fig. 5). When learning the basis vectors for domain-specific subspaces, several key points should be satisfied. Firstly, the basis vectors within the same subspace should be orthogonal to guarantee non-overlapping semantics. Furthermore, different subspaces are supposed to be well separated in the common feature space to distinguish each other. Finally and most importantly, we target at building relationship between the basis vectors and the underlying factors reflecting domain-specific characteristics. To achieve these goals, the following losses are designed to supervise the learning of basis.

Orthogonality within each subspace. The basis for each subspace should focus on different aspects and cover non-overlapping clues. Consequently, the orthogonality loss is designed to push the basis within each subspace apart from each other:

$$L_{orth} = \sum_{d=1}^{D} \left\| \boldsymbol{B}[d]\boldsymbol{B}[d]^T - \boldsymbol{I}_M \right\|_F^2, \tag{19}$$

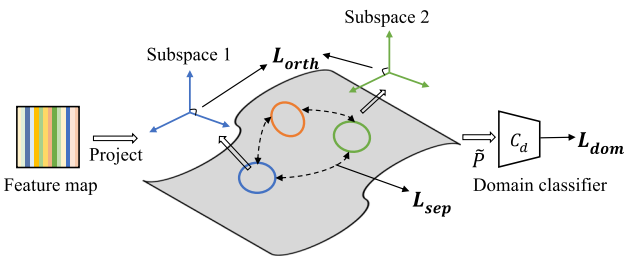


Fig. 5 The procedure of subspace construction

where $\boldsymbol{B}[d] \in \mathbb{R}^{M \times L}$ denotes the basis matrix of the d-th subspace, $\|\cdot\|_F^2$ represents Frobenius norm and $\boldsymbol{I}_M$ is an $M \times M$ identity matrix.

Separation between subspaces. To make sure different domains distinctive, their corresponding subspaces should be far away from each other. Therefore, we need a way of measuring the distance between subspaces. According to the basic Riemannian geometry (Edelman et al., 1998), the set of M-dimensional linear subspaces of the space $\mathbb{R}^C$ ($0 < M \leq C$) makes up a Grassmann manifold, which is a $M(C - M)$ compact Riemannian manifold. In this sense, each domain-specific subspace can be taken as a point on the Grassmann manifold in the common feature space. As a consequence, the distance between subspaces can be naturally quantified with the popular Projection Metric (Harandi et al., 2013) which can be formulated as:

$$d\left(\boldsymbol{B}[d_i], B[d_j]\right) = \frac{1}{\sqrt{2}} \left\| \boldsymbol{B}[d_i]^T \boldsymbol{B}[d_i] - \boldsymbol{B}[d_j]^T \boldsymbol{B}[d_j] \right\|_F. \tag{20}$$

Finally, the metric can be used to form the following separation loss to push subspaces away from each other:

$$L_{sep} = \frac{-1}{\sqrt{2}} \sum_{d_i=1}^{D-1} \sum_{d_j=d_i+1}^{D} \left\| \boldsymbol{B}[d_i]^T \boldsymbol{B}[d_i] - \boldsymbol{B}[d_j]^T \boldsymbol{B}[d_j] \right\|_F. \tag{21}$$

Domain identification. To make the basis embed the underlying factors for each corresponding domain, we utilize the domain classification loss to supervise the projection distance on each subspace. Firstly, we utilize the MaxPool operation along the channel dimension to obtain the global projection distance on each subspace:

$$\tilde{\boldsymbol{P}} = \text{MaxPool}(\boldsymbol{P}) \in \mathbb{R}^{D \times M} \tag{22}$$

Then a linear domain classifier $\boldsymbol{C}_{dom} : \mathbb{R}^{D \times M} \rightarrow \mathbb{R}^D$ is designed for domain label prediction, which is implemented with a weight matrix $\boldsymbol{G} \in \mathbb{R}^{(D \times M) \times D}$. The domain classification loss can be formulated as:

$$\begin{aligned} L_{dom} = &- \sum_{i=1}^{D} \boldsymbol{y}[i] \log \left(\frac{\exp(\tilde{\boldsymbol{P}}\boldsymbol{G}[i])}{\sum_{i=1}^{D} \exp(\tilde{\boldsymbol{P}}\boldsymbol{G}[i])} \right) \\ &+ (1 - \boldsymbol{y}[i]) \log \left(1 - \frac{\exp(\tilde{\boldsymbol{P}}\boldsymbol{G}[i])}{\sum_{i=1}^{D} \exp(\tilde{\boldsymbol{P}}\boldsymbol{G}[i])} \right), \end{aligned} \tag{23}$$

where $\tilde{\boldsymbol{P}}\boldsymbol{G} \in \mathbb{R}^D$ is the classification logits and $\boldsymbol{y} \in \{0, 1\}^D$ denotes the one-hot representation for the domain label.

Monotonicity constraint. Instead of training the weights of the domain classifier, $\boldsymbol{G}$, we fix the weights when learning the basis vectors. In specific, for $j \in \{1, 2, \ldots, DM\}$ and $d \in \{1, 2, \ldots, D\}$ $\boldsymbol{G}[j, d] = 1$ if $\lfloor j/D \rfloor = d$, and otherwise $\boldsymbol{G}[j, d] = -\mu$ where $\mu \in (0, \infty)$ is the penalty

parameter set as 0.5. The key point is to guarantee monotonic semantic for the projection distance $\boldsymbol{P}$. To be more concrete, consider the projection on the j-th basis of subspace d, ie., $\boldsymbol{P}[d, j]$. With the classifier frozen as above, $\boldsymbol{P}[d, j]$ has a strict monotonic positive correlation with the classification logit of domain d, and a strict monotonic negative correlation with the other domains. Such monotonicity is critical in the shield layer which will be discussed later. Besides, freezing classifier can also reduce the difficulty of optimization.

Summarily, the total objection function for learning the subspace basis is:

$$L_{basis} = \lambda_{orth} L_{orth} + \lambda_{sep} L_{sep} + \lambda_{dom} L_{dom}, \tag{24}$$

where λ_{orth}, λ_{sep} and λ_{dom} are the hyper-parameters to balance the losses.

4.2.2 Channel-Wise Shielding

Having factorized the source domains with a group of domain-aware basis vectors, we now discuss how to apply domain-adaptive feature shielding to prevent the model from being biased on target domains due to distributional shifts. Recall that given the feature map of a sample $\boldsymbol{X} \in \mathbb{R}^{C\times HW}$, we project it onto all the basis vectors in the project layer and obtain $\boldsymbol{P} \in \mathbb{R}^{D\times C\times M}$ (Eq. (16)), based on which we aim to acquire channel-wise weights $\boldsymbol{W}$ (Eq. (17)) to mask the output feature map (Eq. (18)).

In general, the weight generation function $\phi(\cdot)$ is designed based on the following assumptions: (1) *Channel independence*. The channel decorrelation in FDB inspires an independent weight generation for each channel based on its own projection distances, which significantly reduces the complexity. (2) *Projection monotonicity*. As discussed above, the fixed domain identification classifier during subspace construction guarantees monotonic semantic for projection distances, such that larger elements in $\boldsymbol{P}$ directly indicate larger channel activation on specific domain-aware factors. (3) *Factor sharing*. With a tractable distributional shift, the target domain is supposed to have similar characteristics with the source domains, which further results in partly shared underlying factors between the target domain and source domains. The channels that effectively react on the co-occurring factors can be treated as domain-adaptive discriminative feature. On the contrary, low activation on all the underlying factors indicates inconsistency inside the channel between the target domain and the source domains. Based on this observation, we can selectively shield the inconsistent channels with a heuristic channel-wise mask formulated as follows:

$$w_c = \phi_c(\boldsymbol{P}[c]) = \underset{\substack{d=1,\dots,D\\ m=1,\dots,M}}{GM^p}\ \sigma(\boldsymbol{P}[c, d, m]). \tag{25}$$

$\phi_c(\boldsymbol{P}[c])$ denotes the c-th independent component of $\phi(\cdot)$ induced by channel independence. $\sigma(\cdot)$ is an activation function mapping the original projection distance to (0, 1) utilizing projection monotonicity. We use the Sigmoid function for simplicity. $GM^p(\cdot)$ denotes the generalized mean operation, ie.,

$$GM^p(x_1, \dots, x_n) = \left(\frac{1}{n} \cdot \sum_{i=1}^{n} x_i^p\right)^{\frac{1}{p}}, \tag{26}$$

which prefers the channels having high activation on some part of shared underlying factors. Note that the hyper-parameter p plays a role in the tolerance degree for filtering out a channel. For example, when $p \to \infty$, $GM^p(\cdot)$ degrades to $\max(\cdot)$, which means that the channel will be filtered only if it does not have any activation on all the factors.

4.3 Overall Objective Function

In a nutshell, the overall objective function of the propose method can be formulated as follows:

$$\mathcal{L} = \mathcal{L}_{reid} + \mathcal{L}_{div} + \mathcal{L}_{basis}, \tag{27}$$

where $\mathcal{L}_{reid}$ consists of the cross-entropy loss $\mathcal{L}_{ce}$, triplet loss $\mathcal{L}_{tri}$ and center loss $\mathcal{L}_{cntr}$ following the standard formulation in previous works (Ye et al., 2021; Zhang et al., 2022).

5 Experiments

5.1 Experimental Setup

Implementation Details. Following the most of current DG-ReID methods, we adopt the wily-used ResNet-50 (He et al., 2016) pre-trained on ImageNet-1K (Deng et al., 2009) as the backbone. Moreover, since the vision transformer presents great potential in various vision task, we further conduct transformer-based experiments to make thorough comparisons. Following the state-of-the-art transformer-based methods (Ni et al., 2023; Chen et al., 2023), vanilla vision transformer (Dosovitskiy et al., 2021) and Swim-transformer (Liu et al., 2021) are both included in our extensive experiments. The main results in Tables 4, 5, 6, 7 and 8 are based on the ViT-B/16 and Swim-B. When adopting the ResNet-50 as the backbone, the last residual layer's stride size is set to 1 following the previous method (Luo et al., 2019). Images are resized to 256 × 128, and the training batch size of each domain is set to 64. For data augmentation, we use random flipping, random cropping and color jittering. We optimize the model using the SGD optimizer with a momentum of 0.9 and weight decay of 5e−4 for 60 epochs,

and the warmup strategy is used in the first 10 epochs. The initial learning rate is set to 4e−2, which is cosine decayed to 4e−5 at the final iteration. The selected channel number K in the designed diversity loss is set as 5 and the tolerance degree p in the channel-wise shielding block is set as 3 as default. Besides, the group size g and the number of bases for each subspace are both set as 16 for the reported results. The baseline is based on the vanilla ResNet50 and trained without any designed blocks under the same training setting as ours. We conduct all the experiments with PyTorch on four 1080Ti GPUs. Datasets and codes will be available at: https://github.com/Xiaofu233/DG-ReID.

Compared DG-ReID methods. The compared methods can be categoried into three categories based on the backbone, ie., resnet-based, vit-based and swim-transformer-based.

For the first category, with the ResNet-50 as the backbone, we re-implement state-of-the-art DG-ReID methods with their public codes for fair comparisons. These methods cover various techniques that are popular in DG-ReID, including the normalization, like SNR (Jin et al., 2020), DualNorm (Jia et al., 2019), CBN (Zhuang et al., 2020), meta-learning, such as M3L (Zhao et al., 2021), MetaBIN (Choi et al., 2021), IL (Tan et al., 2023), distribution alignment, like DDAN (Chen et al., 2021), image matching, like QAConv (Liao & Shao, 2020), ensemble learning, like META (Xu et al., 2022) and dynamic network, like ACL (Zhang et al., 2022), etc..

For the second category, we compare to the recent transformer-based methods, ie., TransMatcher (Liao & Shao, 2021) and PAT (Ni et al., 2023), where TransMatcher use the ResNet-50 together with ViT-B and PAT use the pure ViT-B as the backbone. We utilize the same ViT-B as our backbone for fair comparison.

For the third category, recent pre-training model, ie., SOLIDER (Chen et al., 2023) that conducts large-scale unsupervised learning with larger pre-training datasets based on the swim transformer (Liu et al., 2021), ie., LUPerson (Fu et al., 2021), is included. In detailed, we first evaluate its zero-shot performance, denoted as SOIDER-ZS, where the pre-trained model is testing on the target domains directly. Next, we evaluate its fintuning performance, where the pre-trained model is fine-tuning on the source domains and then testing on the target domains. For fair comparison, the proposed model loads the pre-trained weights and is fine-tuned in the same way to show the effectiveness of the proposed modules.

Evaluation Metrics. We employ the standard metrics in the literature, ie., the Cumulative Matching Characteristics (CMC) at Rank-1, Rank-5, Rank-10 and the mean Average Precision (mAP), to evaluate the performance.

5.2 Results on Proposed Benchmark

Generalization across location. In order to evaluate the generalization ability across different locations of the existing methods and the proposed method, we conduct the experiments under the protocol-1 and protocol-2. The results are shown in Tables 4 and 5, respectively.

Among the ResNet-based methods, we can find that ACL (Zhang et al., 2022) reveals superiority among previous approaches under this type of distribution shift, which designs a static branch to learn domain-invariant features and a dynamic branch to learn domain-specific features. It achieves 52.3% on Rank-1 accuracy and 35.4% mAP on the average, proving the effectiveness of building a common space for domain-invariant and domain-specific feature. Compared to it, the proposed DF^2 focuses on capturing the discriminative feature for each domain in one single branch, promoting the feature diversity and adaptability simultaneously. It performs best compared with all the baselines, outperforming ACL by 4.2% on Rank-1 accuracy and 0.4% mAP, respectively. Besides, the performance of all approaches generalizing from outdoor to indoor datasets are shown in Table 5. We can find similar results as the protocol-1, where the DF^2 still performs best and ACL follows.

Among the ViT-based methods, the proposed model outperforms the recent state-of-the-art method, ie., PAT (Ni et al., 2023) by a large margin. For example, under the protocol-1, the proposed model achieves 41.4% and 28.4% on the average Rank-1 accuracy and mAP, surpassing PAT by 3.7% and 2.9%. Moreover, it becomes clear that under the protocol-1, where the training datasets are relatively limited in size, the performance of ViT-based methods significantly trails behind that of ResNet-based approaches. Additionally, the proposed model, employing ResNet as its backbone, demonstrates superior performance in comparison to the TransMatcher (Liao & Shao, 2021), which integrates ResNet and ViT within its architecture. This performance gap can plausibly be ascribed to the inherent characteristics of ViTs, which typically demand larger training data to effectively learn visual representations. Consequently, when subjected to smaller datasets, they are more prone to overfitting.

Last but not the least, by comparing the performance of SOLIDER (Chen et al., 2023) and ours, we can find a significant performance increase of our DF^2, which demonstrate the effectiveness of the deisnged methods. Furthermore, the results of SOLIDER-ZS are not as satisfactory as anticipated, indicating that solely employing large-scale unsupervised pre-training may not readily address domain gaps, although it remains a promising direction. In conclusion, these findings underscore the efficacy of our DF^2, positioning it as a robust choice for applications that require seamless generalization across outdoor and indoor scenarios.

Table 4 Performance (%) comparison with the state-of-the-art methods under Protocol-1

Method	Backbone	Target: PKU		Target: LPW		Target: WildTrack		Target: Market		Average	
		Rank-1	*m*AP	Rank-1	*m*AP	Rank-1	*m*AP	Rank-1	*m*AP	Rank-1	*m*AP
DualNorm (Jia et al., 2019)	ResNet	19.3	17.7	18.9	12.3	15.9	3.8	36.2	14.2	22.6	12.0
DDAN (Chen et al., 2021)		21.4	15.8	30.6	19.3	15.2	3.2	45.7	18.6	28.2	14.2
SNR (Jin et al., 2020)		32.1	25.8	24.4	17.0	19.3	4.6	41.3	17.6	29.3	16.2
IL (Tan et al., 2023)		21.1	17.4	38.1	27.4	13.0	2.7	61.6	33.2	33.5	20.2
M^3L (Zhao et al., 2021)		35.1	25.8	34.8	25.4	9.6	2.8	63.3	34.1	35.7	22.0
MetaBIN (Choi et al., 2021)		35.1	25.2	**53.3**	39.2	23.5	7.1	68.2	39.2	45.0	27.7
META (Xu et al., 2022)		50.9	36.0	45.8	32.4	19.3	6.2	65.1	34.5	45.3	27.3
CBN (Zhuang et al., 2020)		63.2	45.4	43.5	29.7	24.2	7.6	63.3	33.5	48.5	29.1
$QAConv_{50}$ (Liao & Shao, 2020)		66.7	40.3	51.2	36.0	25.5	6.6	71.0	38.3	53.6	30.3
ACL (Zhang et al., 2022)		55.4	48.1	52.1	**39.1**	26.9	**8.7**	**74.9**	**45.9**	52.3	35.4
Ours		**71.4**	**52.2**	52.4	37.7	**27.6**	8.3	74.4	44.8	**56.5**	**35.8**
TransMatcher (Liao & Shao, 2021)	Res+ViT	61.4	39.6	46.3	31.5	28.3	6.8	67.2	34.2	50.8	28.0
PAT (Ni et al., 2023)	ViT	21.1	22.1	47.2	35.0	16.6	6.8	63.6	37.1	37.1	25.3
Ours		26.3	24.9	53.7	40.8	19.3	7.3	66.3	40.0	41.4	28.4
SOLIDER-ZS (Chen et al., 2023)	Swim	35.1	22.7	16.1	9.7	17.9	3.8	32.5	11.0	25.4	11.8
SOLIDER (Chen et al., 2023)		40.4	34.6	66.4	55.3	29.0	11.8	80.8	61.1	54.1	40.7
Ours		40.4	35.4	67.8	56.9	33.1	12.5	84.4	64.4	56.4	42.3

The ResNet-based best results are highlighted in bold, and the ViT-based best results are underlined.

Table 5 Performance (%) comparison with the state-of-the-art methods under Protocol-2

Method	Backbone	Target: iLIDS		Target: GRID		Target: SAIVT		Target: Airport		Average	
		Rank-1	*m*AP	Rank-1	*m*AP	Rank-1	*m*AP	Rank-1	*m*AP	Rank-1	*m*AP
SNR (Jin et al., 2020)	ResNet	51.7	59.8	19.2	25.6	26.7	14.1	7.0	6.1	26.2	26.4
DualNorm (Jia et al., 2019)		56.7	63.7	25.6	34.7	41.3	20.4	15.8	11.9	34.9	32.7
IL (Tan et al., 2023)		58.3	68.8	29.6	39.7	42.7	24.9	22.0	17.6	38.2	37.8
CBN (Zhuang et al., 2020)		63.3	70.6	35.2	44.3	50.7	25.0	18.5	13.6	41.9	38.4
M^3L (Zhao et al., 2021)		58.3	69.5	32.0	44.6	56.0	28.9	23.5	17.8	42.5	40.2
DDAN (Chen et al., 2021)		61.7	70.9	42.4	52.5	48.0	23.2	21.2	15.9	43.3	40.6
MetaBIN (Choi et al., 2021)		65.0	74.5	44.0	53.7	57.3	30.3	28.2	21.2	48.6	44.9
META (Xu et al., 2022)		65.0	74.4	38.4	49.5	57.3	33.3	30.5	25.1	47.8	45.5
$QAConv_{50}$ (Liao & Shao, 2020)		66.7	75.9	44.0	51.6	57.3	30.7	**33.4**	24.2	50.3	45.6
ACL (Zhang et al., 2022)		65.0	75.7	43.2	52.7	57.9	35.9	30.8	24.0	49.2	47.1
Ours		**70.0**	**76.7**	**44.8**	**55.0**	**69.3**	**42.0**	33.2	**27.3**	**54.3**	**49.9**
TransMatcher (Liao & Shao, 2021)	Res+ViT	75.0	82.6	52.0	60.0	65.3	34.4	36.2	27.3	57.1	51.0
PAT (Ni et al., 2023)	ViT	66.7	74.1	40.0	52.5	64.0	42.2	31.8	25.6	50.6	48.6
Ours		70.0	78.0	44.0	55.3	72.0	48.4	33.3	27.7	54.8	52.4
SOLIDER-ZS (Chen et al., 2023)	Swim	45.0	55.7	17.6	24.7	42.7	15.6	9.8	6.8	28.8	25.7
SOLIDER (Chen et al., 2023)		76.7	84.6	56.0	67.3	74.7	50.4	49.8	44.1	64.3	61.6
Ours		83.3	88.9	59.2	69.9	74.7	54.0	50.4	45.0	66.9	64.5

The ResNet-based best results are highlighted in bold, and the ViT-based best results are underlined.

Table 6 Performance (%) comparison with the state-of-the-art methods under Protocol-3

Method	Backbone	Target: WildTrack		Target: MSMT		Average	
		Rank-1	*m*AP	Rank-1	*m*AP	Rank-1	*m*AP
IL (Tan et al., 2023)	ResNet	13.0	3.2	17.3	5.8	15.2	4.5
DDAN (Chen et al., 2021)		16.5	4.1	16.5	5.2	16.5	4.6
SNR (Jin et al., 2020)		20.7	5.1	14.8	4.7	17.7	4.9
M^3L (Zhao et al., 2021)		17.1	4.0	21.2	7.1	19.6	5.6
DualNorm (Jia et al., 2019)		24.8	7.0	19.0	6.3	21.9	6.7
ACL (Zhang et al., 2022)		24.1	7.1	27.0	9.2	25.5	8.1
$QAConv_{50}$ (Liao & Shao, 2020)		23.5	6.5	40.7	12.0	32.1	9.3
META (Xu et al., 2022)		29.7	8.6	32.8	12.0	31.2	10.3
MetaBIN (Choi et al., 2021)		28.3	8.1	35.3	13.5	31.8	10.8
CBN (Zhuang et al., 2020)		**30.3**	**11.9**	44.4	17.9	37.4	14.9
Ours		29.0	11.5	**48.1**	**20.7**	**38.5**	**16.3**
TransMatcher (Liao & Shao, 2021)	Res+ViT	26.2	8.5	43.6	15.6	34.9	12.0
PAT (Ni et al., 2023)	ViT	24.1	10.3	41.9	17.9	33.0	14.1
Ours		24.8	10.7	41.6	18.1	33.2	14.4
SOLIDER-ZS (Chen et al., 2023)	Swim	17.9	3.8	15.4	3.4	16.7	3.6
SOLIDER (Chen et al., 2023)		44.1	19.8	49.7	23.3	46.9	21.6
Ours		45.2	19.9	50.0	24.0	47.6	22.0

The ResNet-based best results are highlighted in bold, and the ViT-based best results are underlined.

Generalization across weather. Generalizing across weather presents a formidable task due to substantial variations in appearance differences, like clothes, attributes, shapes, etc.. The performance under this type of distribution shift of the previous methods and ours are shown in Tables 6 and 7, respectively.

Among the ResNet-based methods, we can find that CBN (Zhuang et al., 2020) outperforms others to some extent which confirms the camera information as the important factor to bridge the data distribution across different domains. In detail, stead of modeling the distributions of different domains, CBN estimates the raw distribution of each camera from a more nuanced perspective, which benefits to exploring subtle appearance differences. Compared to it, we forward a step to learn more diverse underlying factors that influence the distribution of a domain. As s result, the proposed DF^2 achieves 38.5% under the protocol-3 and 62.9% under the protocol-4 on Rank-1 accuracy, surpassing CBN by 1.1% and 2.6%, respectively.

Among the ViT-based methods, the proposed model demonstrates a substantial performance advantage over the state-of-the-art method, ie., PAT, as evidenced by its achievements under protocol-3 and protocol-4. Specifically, the proposed model attains an average Rank-1 accuracy and mAP of 33.2% and 14.4%, respectively, outperforming PAT by 0.2% and 0.3% under the protocol-3. Besides, under the protocol-4, the proposed model exhibits a marked improvement, achieving a 3.6% and 2.6% increase in Rank-1 accuracy and mAP. Moreover, the proposed model is superior than TransMatcher with less parameters, especially under the protocol-4. In details, the proposed model surpasses TransMatcher by 2.4% and 4.8% on mAP under the protocol-3 and protocol-4, respectively. Finally, the similar results as the generalization scene across location can be found in the generalization scene across weather, where the proposed DF^2 outperformed SOLIDER to a certain degree, which demonstrates the great adaptability of the proposed method. This adaptability is crucial for applications where robust performance across changing seasons is paramount, such as an all-year round identification system.

Generalization across reality. Comparing our proposed DF^2 to existing techniques in the context of generalizing from synthetic to real datasets reveals a notable breakthrough and effective solutions to inherent challenges. Generalizing from synthetic domains to real domains presents substantial difficulties due to differences in image style. As shown in Table 8, most existing methods reveal suboptimal performance when confronted with real-world data. MetaBIN (Choi et al., 2021) is superior than others in the context of this type of distribution shift, which utilizes the batch-instance normalization with meta learning to prevent the model from overfitting to the given source styles effectively. In particular, DDAN (Chen et al., 2021), that conducts domain-wise adversarial feature learning scheme to align domains with minimal distributional shift, achieves surprising performance. Since there is a clear difference between the synthetic datasets and

Table 7 Performance (%) comparison with the state-of-the-art methods under Protocol-4

Method	Backbone	Target: Market		Target: SAIVT		Average	
		Rank-1	*m*AP	Rank-1	*m*AP	Rank-1	*m*AP
SNR (Jin et al., 2020)	ResNet	50.7	24.8	24.0	11.1	37.3	18.0
$QAConv_{50}$ (Liao & Shao, 2020)		68.0	23.8	41.7	15.6	54.9	19.7
DDAN (Chen et al., 2021)		56.9	29.1	34.7	12.2	45.8	20.6
IL (Tan et al., 2023)		55.0	27.8	29.3	14.3	42.2	21.1
DualNorm (Jia et al., 2019)		60.0	30.8	40.0	16.7	50.0	23.7
M^3L (Zhao et al., 2021)		61.3	32.4	46.7	20.7	54.0	26.6
MetaBIN (Choi et al., 2021)		69.0	39.5	40.0	20.2	54.5	29.9
META (Xu et al., 2022)		68.8	39.0	41.3	21.4	55.1	30.2
ACL (Zhang et al., 2022)		73.7	47.2	46.7	20.1	60.2	33.7
CBN (Zhuang et al., 2020)		**76.6**	**47.2**	44.0	20.8	60.3	**34.0**
Ours		69.7	40.0	**56.0**	**25.6**	**62.9**	32.8
TransMatcher (Liao & Shao, 2021)	Res+ViT	79.6	51.2	62.7	29.8	71.1	40.5
PAT (Ni et al., 2023)	ViT	69.7	43.3	70.7	42.0	70.2	42.7
Ours		76.0	44.6	71.6	45.9	73.8	45.3
SOLIDER-ZS (Chen et al., 2023)	Swim	52.3	11.0	42.7	15.6	37.6	13.3
SOLIDER (Chen et al., 2023)		87.4	68.0	70.7	49.7	79.1	58.9
Ours		87.9	68.4	71.6	50.8	79.8	59.6

The ResNet-based best results are highlighted in bold, and the ViT-based best results are underlined.

the real-world datasets in the overall image style, such a feature alignment using domain-wise adversarial learning can effectively reduce the influence of the style variants. In contrast, the proposed DF^2 employs feature diversification to promote the model for capturing diverse discriminative clues and utilizes domain-adaptive shielding to explore the underlying relations between source and target domains, enabling the model to adapt to different styles. By addressing the domain shifts intelligently, our approach consistently outperforms other pure ResNet-based or ViT-based methods, providing more accurate and applicable results in practical scenarios. Specially, TransMatcher achieves wonderful performance under this distribution shift, which reveals the effectiveness of the image matching and metric learning. However, its distinct combination of ResNet and ViT results in a more complex architecture, which may pose challenges for deployment in real-world applications due to its substantial computational requirements.

5.3 Results on Previous Benchmarks

In order to demonstrate the robustness of the proposed method, we further conduct the experiments on the previous benchmarks. Following the recent researches (Song et al., 2019; Zhang et al., 2022; Xu et al., 2022), the detailed evaluation protocols of the previous benchmarks are shown in Table 9. Benchmark-1 utilized Market1501 (M), CUHK-SYSU (CS), CUHK02 (C2) and CUHK03 (C3) as the source domains, while VIPeR, PRID, GRID and i-LIDS as target domains. Benchmark-2 used CUHK03, Market-1501, CUHK-SYSU and MSMT17 (MT) and conducted the leave-one-out strategy to design the evaluation protocols. 'Com' denotes adopting both training and testing sets of the datasets for model training. As shown in Tables 10 and 11, we report the experimental results of the proposed method on the previous open benchmarks, respectively.

First of all, the proposed method obtains the best average performance on both previous benchmarks, which proves the robustness of DF^2. Specifically, on benchmark-1, the mean performance of the proposed DF^2 outperforms ACL (Zhang et al., 2022) by 1.2% on Rank-1 accuracy and 1.2% on mAP, respectively. Besides, under the two protocols on benchmark-2, the results in Table 11 show that DF^2 outperforms the performances of previous methods by a large margin. To be specific, among the ResNet-based methods, the proposed DF^2 improves the second-best ACL by 1.5% Rank-1 accuracy, 1.2% mAP and 3.1% Rank-1 accuracy, 3.1% mAP on average under the two protocols, respectively. Besides, the proposed DF^2 surpasses TransMatcher (Liao & Shao, 2021) which uses more heavy backbone on all open benchmarks. Moreover, the proposed DF^2 still outperforms the state-of-the-art ViT-based method, i.e., PAT (Ni et al., 2023), and the similar results as the comparison to the PAT can be found in the comparison to the SOLIDER (Chen et al., 2023), where the proposed DF^2 surpasses SOLIDER to a certain degree with the same pre-trained weights. These evaluations

Table 8 Performance (%) comparison with the state-of-the-art methods under Protocol-5

Method	Backbone	Target: iLIDS		Target: GRID		Target: Market		Target: MSMT		Average	
		Rank-1	*m*AP	Rank-1	*m*AP	Rank-1	*m*AP	Rank-1	mAP	Rank-1	*m*AP
DualNorm (Jia et al., 2019)	ResNet	31.7	44.5	13.6	17.2	53.5	25.8	17.7	5.5	29.1	23.2
IL (Tan et al., 2023)		55.0	63.2	12.8	19.8	55.7	29.2	22.6	7.5	36.5	29.9
SNR (Jin et al., 2020)		51.7	63.2	17.6	25.9	52.9	27.6	19.4	6.2	35.4	30.7
META (Xu et al., 2022)		60.0	68.0	18.4	27.1	57.0	31.5	24.1	7.9	39.9	33.6
$QAConv_{50}$ (Liao & Shao, 2020)		**63.3**	**70.2**	22.4	21.0	60.5	30.6	34.7	9.9	45.2	35.4
CBN (Zhuang et al., 2020)		60.0	66.1	21.6	31.5	67.5	40.1	24.6	8.4	43.4	36.5
M^3L (Zhao et al., 2021)		61.7	69.8	20.0	27.8	63.4	36.7	29.6	10.5	43.7	36.2
ACL (Zhang et al., 2022)		58.3	68.3	25.6	31.6	65.4	40.6	22.4	7.3	42.9	36.9
DDAN (Chen et al., 2021)		58.3	68.5	**29.6**	**37.5**	62.8	35.5	21.0	6.4	42.9	37.0
MetaBIN (Choi et al., 2021)		58.3	68.1	25.6	33.0	66.0	41.7	29.2	10.7	44.8	38.4
Ours		60.0	68.1	29.2	36.4	**69.2**	**43.5**	**36.8**	**13.7**	**49.3**	**40.8**
TransMatcher (Liao & Shao, 2021)	Res+ViT	73.3	80.1	30.4	40.6	75.7	51.6	45.5	16.3	56.2	47.1
PAT (Ni et al., 2023)	ViT	48.3	61.3	9.6	14.6	46.7	12.4	12.2	3.7	29.2	23.0
Ours		56.7	69.4	24.8	36.3	71.1	45.4	36.0	13.4	47.2	41.1
SOLIDER-ZS (Chen et al., 2023)	Swim	45.0	55.7	17.6	24.7	32.5	11.0	15.4	3.4	27.6	23.7
SOLIDER (Chen et al., 2023)		63.3	73.0	28.0	37.4	77.1	53.7	38.0	15.4	51.6	44.9
Ours		73.3	78.7	27.2	35.0	77.7	54.8	37.5	14.7	53.9	45.8

The ResNet-based best results are highlighted in bold, and the ViT-based best results are underlined.

Table 9 Details of the previous benchmarks

Setting	Training data	Testing data
Benchmark-1	Com-(M+C2+C3+CS)	PRID, GRID, VIPeR, iLIDs
Benchmark-2	CS+C3+MT	M
	M+CS+MT	C3
	M+CS+C3	MT
	Com-(CS+C3+MT)	M
	Com-(M+CS+MT)	C3
	Com-(M+CS+C3)	MT

Table 10 Performance (%) comparison with the state-of-the-art methods on the previous benchmark-1

Method	Backbone	Target: VIPeR		Target: PRID		Target: GRID		Target: iLIDS		Average	
		Rank-1	*m*AP	Rank-1	*m*AP	Rank-1	*m*AP	Rank-1	*m*AP	Rank-1	*m*AP
SNR (Jin et al., 2020)	ResNet	26.6	36.4	18.0	26.3	15.2	23.7	48.3	56.1	27.0	35.6
DualNorm (Jia et al., 2019)		32.9	43.7	49.0	59.0	23.2	33.7	58.3	66.5	40.9	50.7
DDAN (Chen et al., 2021)		41.1	52.4	44.0	54.0	39.2	49.2	63.3	70.8	46.9	56.6
CBN (Zhuang et al., 2020)		46.5	55.2	47.0	57.0	44.0	52.0	71.7	76.1	52.3	60.1
IL (Tan et al., 2023)		60.4	68.1	65.0	73.7	32.8	44.5	62.0	71.6	55.1	64.5
M^3L (Zhao et al., 2021)		60.8	68.2	55.0	65.3	40.0	50.5	65.0	74.3	55.2	64.6
$QAConv_{50}$ (Liao & Shao, 2020)		57.0	66.3	52.3	62.2	48.6	57.4	75.0	81.9	58.2	67.0
MetaBIN (Choi et al., 2021)		55.9	64.3	61.2	70.8	50.2	57.9	74.7	82.7	60.5	68.9
META (Xu et al., 2022)		61.5	68.4	61.9	71.7	52.4	60.1	79.2	83.5	63.8	70.9
ACL (Zhang et al., 2022)		66.4	75.1	63.0	73.4	55.2	**65.7**	**81.8**	**86.5**	66.6	75.2
Ours		**66.5**	**75.2**	**71.0**	**78.1**	**55.4**	65.0	78.3	85.7	**67.8**	**76.0**
TransMatcher (Liao & Shao, 2021)	Res+ViT	53.5	63.3	64.0	73.0	48.8	56.7	71.7	79.2	59.5	68.0
PAT (Ni et al., 2023)	ViT	58.5	67.7	48.0	59.7	40.8	52.6	63.3	74.8	52.7	63.7
Ours		58.9	68.1	58.0	68.3	44.0	54.5	64.3	75.7	56.3	66.7
SOLIDER-ZS (Chen et al., 2023)	Swim	12.3	20.3	12.0	19.5	17.6	24.7	45.0	55.7	21.7	30.1
SOLIDER (Chen et al., 2023)		74.1	80.6	76.0	83.5	71.2	80.1	86.7	91.1	77.0	83.8
Ours		75.1	81.2	77.0	83.2	72.8	80.7	87.7	92.0	78.2	84.3

The ResNet-based best results are highlighted in bold, and the ViT-based best results are underlined.
MS denotes the multiple source domains, including Market1501, CUHK-SYSU (Xiao et al., 2016), CUHK02 (Li & Wang, 2013) and CUHK03 (Li et al., 2014). The best results under MS setting are highlighted in bold

using open benchmarks, beyond pre-defined protocols, can demonstrate the transparent and universal effectiveness of the proposed method.

Secondly, we can observe that the state-of-the-art methods on previous benchmarks are not guaranteed to perform well on the new benchmark, where there are more challenged and specific distribution shifts. For instance, META (Xu et al., 2022) and ACL (Zhang et al., 2022) perform well on the previous benchmarks while their performance on the new benchmark are not satisfactory under some circumstances. META is based on the model ensemble learning, which is bounded by the quantity of the source domains. ACL maintains a common space for both domain-invariant and domain-specific features but ignores the representation capacity of the learned feature, which limits the generalization ability when faced with significant distribution shifts. Moreover, the performance of ViT-based models doesn't meet expectations, underscoring the necessity for further exploration to unlock the full potential of ViT. We hope the extensive experiments can provide valuable insights for researchers to advance the development of this field.

5.4 Ablation Study

Effectiveness of designed components. To demonstrate the effectiveness of our custom-designed Feature Diversification Block (FDB) and Domain-adaptive Shielding Block (DSB), we incorporate them separately into the baseline model and conduct a performance comparison, as detailed in Table 12. This evaluation pertains to the model's generalization capabilities across outdoor and indoor domains under protocol-1. When we refer to '*w/* DSB', we mean the model equipped solely with the DSB block, whereas '*w/* FDB' signifies the model equipped solely with the FDB block. Notably, our

Table 11 Performance (%) comparison with the state-of-the-art methods under the previous benchmark-2

	Method	Backbone	MT+CS+C3→ M		M+MT+CS→ C3		M+CS+C3→ MT		Average	
			Rank-1	*m*AP	Rank-1	*m*AP	Rank-1	*m*AP	Rank-1	*m*AP
Training Sets	SNR (Jin et al., 2020)	ResNet	62.7	34.6	8.9	8.9	19.9	6.8	30.5	16.8
	DualNorm (Jia et al., 2019)		50.9	25.0	6.8	7.3	23.8	8.5	27.2	13.6
	DDAN (Chen et al., 2021)		70.8	42.9	16.3	16.8	18.4	5.6	35.2	21.7
	CBN (Zhuang et al., 2020)		78.5	53.5	27.4	27.6	33.2	13.7	46.4	31.6
	M^3L (Zhao et al., 2021)		79.9	58.4	31.9	20.9	36.9	15.9	49.6	31.7
	IL (Tan et al., 2023)		80.2	57.2	31.6	31.3	30.8	12.6	47.5	33.7
	MetaBIN (Choi et al., 2021)		80.1	57.9	28.1	28.8	40.2	17.8	49.5	34.8
	QAConv50 (Liao & Shao, 2020)		83.7	63.1	24.8	25.4	45.3	16.4	51.3	35.0
	META (Xu et al., 2022)		86.1	67.5	35.1	36.3	49.9	22.5	57.0	42.1
	ACL (Zhang et al., 2022)		89.3	**74.3**	41.8	41.2	45.9	20.4	59.0	45.3
	Ours		**89.5**	74.1	**42.4**	**42.3**	**49.7**	**23.2**	**60.5**	**46.5**
	TransMatcher (Liao & Shao, 2021)	Res+ViT	83.4	62.3	33.9	31.6	47.1	17.5	54.8	37.1
	PAT (Ni et al., 2023)	ViT	70.0	46.5	23.1	24.8	27.9	12.3	40.3	27.9
	Ours		73.0	51.8	29.5	31.1	28.0	13.4	43.5	32.1
	SOLIDER-ZS (Chen et al., 2023)	Swim	32.5	11.0	3.0	3.5	15.4	3.4	17.0	6.0
	SOLIDER (Chen et al., 2023)		91.6	81.7	61.4	59.7	37.5	18.8	63.5	53.4
	Ours		91.7	82.1	60.6	60.1	43.5	23.0	65.3	55.1
Full Images	DualNorm (Jia et al., 2019)	ResNet	60.5	33.3	13.1	13.5	26.4	9.6	33.4	18.8
	DDAN (Chen et al., 2021)		76.8	51.9	20.4	20.3	21.0	6.8	39.4	26.4
	SNR (Jin et al., 2020)		77.8	52.4	17.1	17.5	22	7.7	39.0	25.9
	CBN (Zhuang et al., 2020)		75.4	48.9	29.3	29.7	35.1	14.7	46.6	31.1
	M^3L (Zhao et al., 2021)		81.2	61.2	33.8	32.3	36.9	16.2	50.6	36.6
	IL (Tan et al., 2023)		84.7	66.9	40.4	39.7	34.5	15.2	53.2	40.6
	$QAConv_{50}$ (Liao & Shao, 2020)		85.0	66.5	33.3	32.9	46.6	17.6	55.0	39.0
	MetaBIN (Choi et al., 2021)		84.5	67.2	43.1	43.0	41.2	18.8	56.3	43.0
	META (Xu et al., 2022)		90.5	76.5	46.2	47.1	52.1	24.4	62.9	49.3
	ACL (Zhang et al., 2022)		90.6	76.8	50.1	49.4	47.3	21.7	62.7	49.3
	Ours		**91.3**	**78.7**	**53.1**	**53.2**	**52.9**	**25.4**	**65.8**	**52.4**
	TransMatcher (Liao & Shao, 2021)	Res+ViT	87.0	67.5	36.8	35.1	44.7	16.9	56.2	39.8
	PAT (Ni et al., 2023)	ViT	73.8	52.6	32.4	32.7	23.2	10.8	43.1	32.0
	Ours		74.2	53.0	33.1	33.5	27.8	13.2	45.0	33.2
	SOLIDER-ZS (Chen et al., 2023)	Swim	32.5	11.0	3.0	3.5	15.4	3.4	17.0	6.0
	SOLIDER (Chen et al., 2023)		91.9	82.2	60.1	60.7	41.0	20.9	64.3	54.6
	Ours		92.3	82.5	60.9	61.5	44.2	22.5	65.8	55.5

The ResNet-based best results are highlighted in bold, and the ViT-based best results are underlined.
'Training Sets' denotes that only the training sets in the source domains are used for training and 'Full Images' denotes that all images are used for training.

Table 12 Contribution of each component to generalization performance (%) from indoor domains to outdoor domains under protocol-1

	Rank-1	Rank-5	Rank-10	mAP
PKU				
Baseline	62.5	80.4	87.5	46.3
w/ DeepMOE	63.4	80.9	86.1	47.3
w/ DSB	67.9	83.9	89.3	50.9
w/ FDB	68.3	84.5	90.0	51.7
DF^2	**71.4**	**85.7**	**90.5**	**52.2**
LPW				
Baseline	40.8	53.6	59.2	25.2
w/ DeepMOE	41.0	54.2	60.7	26.3
w/ DSB	42.8	59.5	62.4	29.0
w/ FDB	47.3	63.1	69.3	33.8
DF^2	**52.4**	**70.5**	**77.0**	**37.7**
WildTrack				
Baseline	20.7	31.3	37.7	5.0
w/ DeepMOE	22.8	35.5	40.9	6.3
w/ DSB	25.3	39.0	42.2	7.2
w/ FDB	27.0	40.3	44.2	8.0
DF^2	**27.5**	**41.4**	**45.5**	**8.3**
Market1501				
Baseline	62.9	78.6	84.1	32.4
w/ DeepMOE	63.9	81.3	86.5	33.7
w/ DSB	66.6	82.5	87.9	38.8
w/ FDB	70.7	84.1	89.0	41.4
DF^2	**74.4**	**87.1**	**90.8**	**44.8**

Best results are highlighted in bold
DeepMoE (Wang et al., 2020) is the similar method to DSB that also conducts the selection and scaling of channels for each input. Bold values indicate the best results among all compared methods

Table 13 Contribution of each whitening technique to generalization performance (%) under protocol-1

	Rank-1	Rank-5	Rank-10	mAP
PKU				
Baseline	62.5	80.4	87.5	46.3
IBN	64.0	81.4	88.4	47.8
DSBN	68.9	82.6	88.7	50.0
Ours (IW)	69.0	82.7	88.9	50.5
Ours (BW)	70.1	83.8	89.1	51.1
Ours (IBW)	**71.4**	**85.7**	**90.5**	**52.2**
LPW				
Baseline	40.8	53.6	59.2	25.2
IBN	45.4	65.1	74.1	32.8
DSBN	47.9	66.0	75.4	34.8
Ours (IW)	49.0	66.1	75.3	34.9
Ours (BW)	50.1	68.1	76.2	36.9
Ours (IBW)	**52.4**	**70.5**	**77.0**	**37.7**
WildTrack				
Baseline	20.7	31.3	37.7	5.0
IBN	25.9	36.8	38.4	6.3
DSBN	26.0	37.0	39.5	6.9
Ours (IW)	26.4	38.0	40.9	7.0
Ours (BW)	28.1	39.5	42.6	7.6
Ours (IBW)	**27.5**	**41.4**	**45.5**	**8.3**
Market1501				
Baseline	62.9	78.6	84.1	32.4
IBN	68.0	83.1	87.1	36.8
DSBN	69.8	84.1	88.7	37.8
Ours (IW)	70.5	84.8	89.3	38.5
Ours (BW)	71.1	86.2	90.0	40.8
Ours (IBW)	**74.4**	**87.1**	**90.8**	**44.8**

Best results are highlighted in bold
Bold values indicate the best results among all compared methods for the current dataset

analysis reveals that both the DSB and FDB substantially enhance the model's generalization abilities. For instance, when considering the PKU dataset, the introduced DSB exhibits a notable improvement of 5.4% on Rank-1 accuracy and 2.6% on mAP compared to the baseline. The contribution of the FDB is even more remarkable, resulting in a substantial performance increase of 6.2% in Rank-1 accuracy and 5.4% in mAP. These enhancements consistently hold true across all datasets. Besides, in order to assess the exclusivity compared to similar modules, we make the comparison to the DeepMoE (Wang et al., 2020), which also conducts the selection and scaling of channels for each input. We can find the designed DSB outperforms DeepMoE to a large extent, which demonstrates the effectiveness and exclusivity of DSB.

Effectiveness of designed whitening. Our designed instance-batch whitening method comprises two key components: instance whitening (IW) and batch whitening (BW). To thoroughly assess the effectiveness of each component, we conduct a performance comparison by equipping the model with different whitening techniques within the channel decorrelation block. Additionally, we further consider the popular feature regularization methods IBN (Pan et al., 2018) and DSBN (Chang et al., 2019), which have been extensively studied in prior works (Jin et al., 2020; Choi et al., 2021; Xu et al., 2022), to provide a comprehensive evaluation. As indicated in Table 13, the incorporation of normalization and whitening techniques consistently enhances model performance. Notably, the improvements attributed to whitening techniques surpass those of normalization, underscoring their significant potential. For instance, the introduction of DSBN yields a modest performance boost, such as a 6.4% increase in Rank-1 accuracy on the PKU dataset. However, when we employ IBW, the gain is more substantial, amounting to a 8.9% improvement in Rank-1 accuracy on PKU. Furthermore, it's worth noting that batch whitening, which operates

Table 14 Ablation study on different losses

Method	Target: PKU		Target: LPW		Target: WildTrack		Target: Market1501		Average	
	Rank-1	*m*AP	Rank-1	*m*AP	Rank-1	*m*AP	Rank-1	*m*AP	Rank-1	*m*AP
w All	71.4	52.2	52.4	37.7	27.6	8.3	74.4	44.8	56.5	35.8
w/o $\mathcal{L}_{div}$	− 4.0	− 3.6	− 2.0	− 0.8	− 1.2	− 0.9	− 2.2	− 1.6	− 2.4	− 1.7
w/o $\mathcal{L}_{orth}$	− 1.2	− 1.0	− 1.1	− 0.9	− 2.7	− 1.5	− 1.6	− 1.0	− 1.7	− 1.1
w/o $\mathcal{L}_{sep}$	− 2.8	− 2.6	− 0.9	− 0.7	− 0.8	− 0.4	− 1.0	− 0.7	− 1.4	− 1.1
w/o $\mathcal{L}_{dom}$	− 1.8	− 1.3	− 3.9	− 2.2	− 2.3	− 1.0	− 4.9	− 4.1	− 3.2	− 2.2

on the mini-batch level to decorrelate features, outperforms instance whitening on a per-example basis. Nevertheless, the most favorable results are achieved when combining both techniques.

Effectiveness of each loss term. We conduct ablation studies to investigate the effectiveness of the each loss in the proposed DF^2. As shown in Table 14, '*w* ALL' denotes the results of the proposed DF^2 with all losses, and '*w/o* $\mathcal{L}_i$' denotes the results of the proposed DF^2 under the supervision without $\mathcal{L}_i$. We can observe that removing every loss term consistently degrades the performance on Rank-1 accuracy and mAP. Specifically, when removing the $\mathcal{L}_{div}$, which is designed to maximize the activation of all spatial locations, the performance degrades 2.4% on Rank-1 accuracy and 1.7% on mAP. Besides, there are similar impacts of $\mathcal{L}_{orth}$ and $\mathcal{L}_{sep}$, which are proposed to promote the orthogonality and separation between the subspace bases. Compared to them, the supervision of $\mathcal{L}_{dom}$ can bring more improvements. This phenomenon demonstrates the significant of modeling the domain characteristics.

5.5 Hyper-parameter Analysis

Influence of the group size in FDB. To understand the parameter selection of the proposed DF^2, we make the analysis from both the model performance and computation efficiency aspects. Regarding the influence on model performance, we train it under protocol-2 with different parameter configurations to the defaults as shown in Fig. 6a. It presents the comparison of the average performance with different group sizes in FDB. First of all, we observe a gradual enhancement in performance as we increased the group size (g).

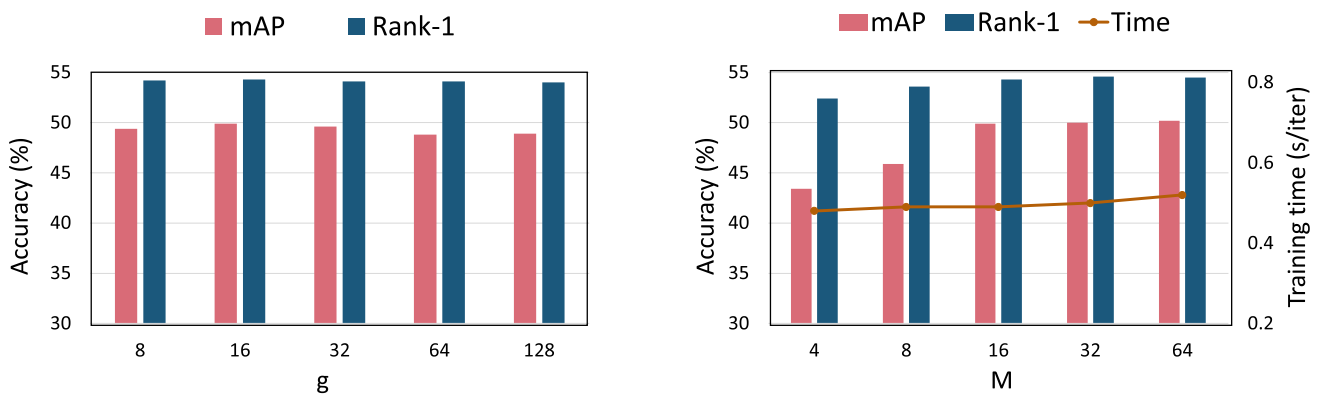
Regarding the computation cost, from the theoretical aspects, a convolutional layer with a batch input $N \times C \times H \times W$ input, and C filters of size $F_h \times F_w$ costs $C^2NHWF_hF_w$. Adding FDB with a group size g incurs an overhead of $4gCNHW + 6g^2C$. The relative overhead is $\frac{6g^2}{CNHWF_gF_w} + \frac{4g}{CF_hF_w}$, which is negligible when g is small (e.g. 16). This demonstrates the computation cost of the proposed FDB is comparable to the convolution operation. From the empirical aspect, we compare the FLOPs and the training time for each iteration of the proposed model with different group sizes to the baseline model that is not equipped with any designed modules. The results are presented in Table 15 below. When the g is small ($<= 32$), the additional cost is blow 10% compared to the baseline. With larger group sizes, the training time for each iteration also increases significantly. Consequently, to strike a balance between accuracy and computational efficiency, it is set as 16.

Influence of the number of bases. Figure 6b presents the comparison of the average performance and the training time per iteration with different number of bases M in each subspace. First of all, we can find that the number of bases had a less pronounced impact on the training time. Besides, the performance is slightly influenced by the number of bases, where relatively small numbers may lead to inadequate modeling of domain information. Consequently, we set it to $M = 16$ for optimal results.

Influence of the balanced weights of each loss. In this section, we evaluate the sensitivity of the four balanced weights of losses. The default balanced weights λ_{div}, λ_{orth}, λ_{sep} and λ_{dom} are 1.0, 0.01, 0.1 and 1.0, respectively. We conduct the thorough experiments with different values of each balanced weight. As shown in Fig. 7, the proposed DF^2 achieves competitive performances robustly under a wide range of their values. Taking the Market1501 as an example, when λ_{div} increases from 0.1 to 2.0, the range of performance changes is within 1.6%. Furthermore, different weight values do not cause drastic fluctuations in the performance, indicating that most of these weights are insensitive to numerical changes.

5.6 More Analysis

Computational cost analysis. In this experiment, we demonstrate that DF^2 not only achieves superior performance in terms of DG-ReID accuracy, but is also advantageous in terms of time and space complexities. To facilitate a fair comparison, we utilize the same batch size and the same NVIDIA 1080Ti GPU for all methods in Table 16 and the inference time is averaged over 500 trials. First of all, the computation cost of ViT-based methods is notably higher than ResNet-based methods, especially on the network param-



(a) Comparison with different group sizes

(b) Comparison with different number of bases

Fig. 6 Results of hyper-parameter analysis under protocol-2. **a** Present the comparison of the average performance with different group size in FDB. **b** Presents the comparison of the average performance with different number of basis per subspace in DSB

Table 15 Analysis of computational complexity with different group sizes g in FDB block

Group size	8	16	32	64	128	Baseline
FLOPs (G)	8.40	8.50	8.70	9.12	10.10	8.26
△ Cost	1.7%	2.9%	5.3%	10.4%	22.2%	–
Time (s/iter)	0.43	0.49	0.56	0.68	0.75	0.30

△ Cost denotes the additional flops compared to the baseline model, ie., ResNet50

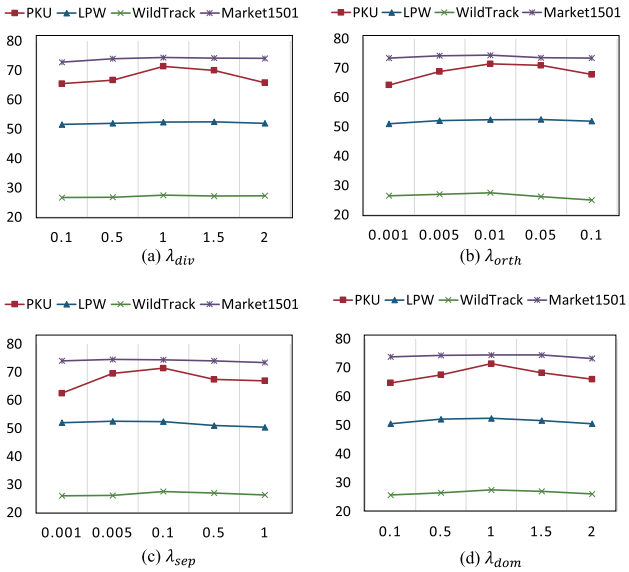


Fig. 7 Influence of the balanced weights of each loss

eters. Furthermore, it's evident that the computational cost of our proposed method is notably lower during the training stage when compared to recent methods. This reduction in computational cost can be attributed to DF^2's independence from the meta-learning training strategy compared to the M^{3}L, which typically involves a high computational overhead due to the necessity of two backward propagations. In comparison to IL (Tan et al., 2023), our approach introduces only a marginal increase in computational requirements while delivering a substantial performance boost. Additionally, DF^2 maintains a rapid inference speed, on par with that of the IL. To sum up, our findings reveal that the additional computation cost introduced by our designed blocks are minimal.

Comparison of covariance matrices. To illustrate the effectiveness of the feature decorrelation, we employ a visualization technique to examine the correlation coefficients of intermediate feature maps generated by the model both with and without feature diversification block (FDB). This analysis is depicted in Fig. 8, wherein the correlation coefficients are computed based on the feature output from the second convolution layer. The presented pairs of visualizations offer insights into two aspects of correlation. The above pairs show the overall correlation, which demonstrates the global correlation among all channels. The pairs below show detailed correlation, focusing on the specific correlations within the first 16 channels, enhancing clarity in our presentation. It's noteworthy that domain-specific style information is inherently embedded within the features, as established in prior research (Tishby & Zaslavsky, 2015; Pan et al., 2018). Consequently, feature maps with high correlations tend to encapsulate limited style information, potentially resulting in overfitting to the source domains. By comparing these correlation coefficients, we can observe a significant distinction: feature channels generated by the model with FDB exhibit considerably lower correlations than those without FDB. This observation underscores the effectiveness of the proposed FDB in the process of decorrelating learned features, thereby encouraging the emergence of diverse representations.

Subspaces analysis. To demonstrate the DSB's ability to construct these subspaces, we visualize the orthogonal-

Table 16 Comparison of computational cost to the state-of-the-art methods

Model		# of Params (M)	Train (s/iter)	Inference (ms/img)
ResNet	M^3L (Zhao et al., 2021)	23.5	1.974	0.21
	MetaBIN (Choi et al., 2021)	23.6	0.585	0.32
	ACL (Zhang et al., 2022)	29.6	0.588	0.36
	IL (Tan et al., 2023)	23.5	0.352	0.21
	Ours	23.6	0.488	0.21
ViT	PAT (Ni et al., 2023)	93.1	1.016	0.51
	Ours	88.4	0.908	0.44

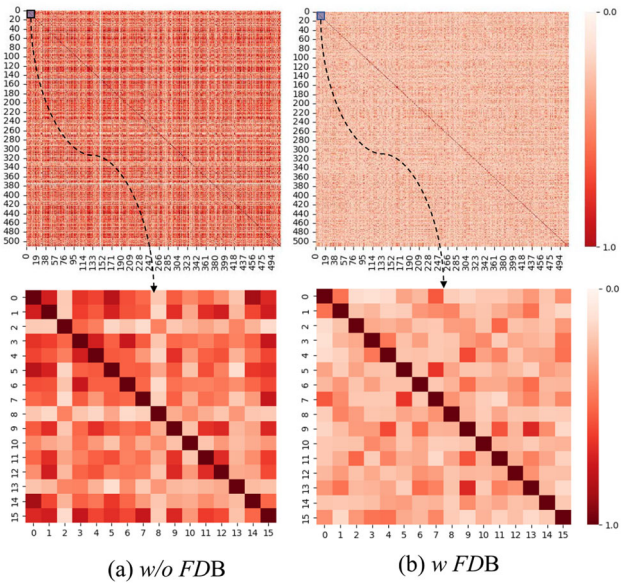


Fig. 8 Visualization of correlation coefficients extracted from from two distinct models: one without the proposed channel decorrelation block (referred to as *w/o* FDB) and the other euipped with the channel decorrelation block (referred to as *w* FDB)

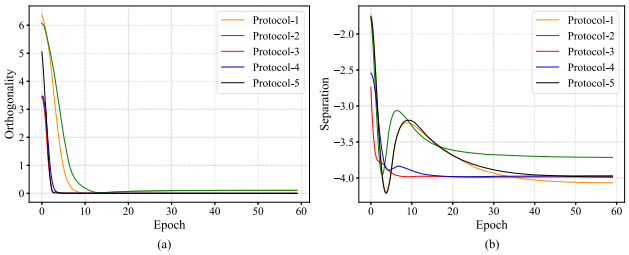


Fig. 9 Visualization of the orthogonality loss and the separation loss during the training process

Fig. 10 Failure cases analysis

ity loss and the separation loss, as derived from Eqs. (19) and (21), during the training process under all protocols. The values of these losses are averaged across each epoch. As shown in Fig. 9a, there is a clear trend where the orthogonality of subspaces progressively improves, reaching an optimal state of absolute orthogonality, where the orthogonality loss converges to zero. Besides, to make sure different domains distinctive, different subspaces are supposed to be well separated in the common feature space. To this end, we visualize the separation loss, where a lower value signifies greater subspace separation. As depicted in Fig. 9b, despite initial fluctuations, the separation loss demonstrates a declining trend, eventually stabilizing at a minimized value. Such outcomes affirm the subspace construction's efficacy facilitated by the DSB block.

Failure cases analysis. To better understand the proposed method, we present some failure cases in Fig. 10. Taking the results under protocol-1 as an example, we selected the cases that most retrieved images of the in the top-10 ranking list are wrong. We plot the query image and corresponding retrieved gallery images in the first and subsequent columns of these figures respectively, where the blue box indicates that the retrieved result is incorrect, and the red box indicates that the retrieved result is correct. Our observation reveals that the proposed DF^2 faces challenges in extreme scenarios, where the cross-class similarity significantly outweighs the in-class similarity. For example, as demonstrated in the first row, the incorrectly retrieved images appear more similar than the actual ones due to light variation. Addressing these challenging samples with greater precision is one of our future research directions.

6 Conclusion

In this paper, we conducted a comprehensive study on the domain generalizable person re-identification (DG-ReID)

problem. Firstly, we gave a focused discussion on the construction of benchmarks in DG-ReID, especially pointing out the highlighted challenges worth being reflected in benchmarks. In the hope of promoting more advanced DG-ReID research, we proposed a large-scale benchmark with Enhanced Distributional VARiety and Shifts (EDVARS) comprised of diverse collected datasets and rational evaluation protocols. Taking a step further, we designed a novel DG-ReID framework based on Diver spacE Learning with domain FActorization (DF^2) in response to the highlighted challenges. In pursuit of generalization ability with scalable additional memory and computation costs, DF^2 consists of two types of proposed blocks, ie., the Feature Diversification Block (FDB) and the Domain-adaptive Shielding Block (DSB). FDB promotes a diverse feature space capable of learning domain-specific characteristics under rich distributional variety, whose core design is composed of the instance-batch whitening together with a diversity loss. DSB, consisting of a project layer and a shield layer, applies channel-wise shielding operations based on subspace-based domain factorization in order to prevent the model from prediction bias caused by distributional shifts. Finally, we conducted extensive experiments to demonstrate the effectiveness of the proposed DF^2 framework.

Funding This work was supported by National Natural Science Foundation of China (NSFC) under Grants 62225207 and 62106245.

References

Barbosa, I. B., Cristani, M., Caputo, B., Rognhaugen, A., & Theoharis, T. (2018). Looking beyond appearances: Synthetic training data for deep CNNs in re-identification. *Computer Vision and Image Understanding, 167*, 50–62.

Bialkowski, A., Denman, S., Sridharan, S., Fookes, C., & Lucey, P. (2012). A database for person re-identification in multi-camera surveillance networks. In *International conference on digital image computing techniques and applications (DICTA)* (pp. 1–8). IEEE.

Bini, D. A., Higham, N. J., & Meini, B. (2005). Algorithms for the matrix p-th root. *Numerical Algorithms, 39*(4), 349–378.

Cai, Y., Takala, V., & Pietikainen, M. (2010). Matching groups of people by covariance descriptor. In *20th International conference on pattern recognition* (pp. 2744–2747). IEEE.

Chang, W. G., You, T., Seo, S., Kwak, S., & Han, B. (2019). Domain-specific batch normalization for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7354–7362).

Chavdarova, T., Baqué, P., Bouquet, S., Maksai, A., Jose, C., Bagautdinov, T., Lettry, L., Fua, P., Van Gool, L., & Fleuret, F. (2018). Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5030–5039).

Chen, P., Dai, P., Liu, J., Zheng, F., Xu, M., Tian, Q., & Ji, R. (2021). Dual distribution alignment network for generalizable person re-identification. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 1054–1062).

Chen, W., Xu, X., Jia, J., Luo, H., Wang, Y., Wang, F., Jin, R., & Sun, X. (2023). Beyond appearance: A semantic controllable self-supervised learning framework for human-centric visual tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 15050–15061).

Cho, W., Choi, S., Park, D. K., Shin, I., & Choo, J. (2019). Image-to-image translation via group-wise deep whitening-and-coloring transformation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10639–10647).

Cho, Y., Cho, H., Kim, Y., & Kim, J. (2021). Improving generalization of batch whitening by convolutional unit optimization. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 5321–5329).

Choi, S., Kim, T., Jeong, M., Park, H., & Kim, C. (2021). Meta batch-instance normalization for generalizable person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3425–3435).

Dai, Y., Li, X., Liu, J., Tong, Z., & Duan, L. Y. (2021). Generalizable person re-identification with relevance-aware mixture of experts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16145–16154).

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 248–255).

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S, Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International conference on learning representations*.

Edelman, A., Arias, T. A., & Smith, S. T. (1998). The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications, 20*(2), 303–353.

Fu, D., Chen, D., Bao, J., Yang, H., Yuan, L., Zhang, L., Li, H., & Chen, D. (2021). Unsupervised pre-training for person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 14750–14759).

Gou, M., Wu, Z., Rates-Borras, A., Camps, O., & Radke, R. J. (2018). A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(3), 523–536.

Gray, D., & Tao, H. (2008). Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European conference on computer vision* (pp. 262–275). Springer.

Harandi, M., Sanderson, C., Shen, C., & Lovell, B. C. (2013). Dictionary learning and sparse coding on Grassmann manifolds: An extrinsic solution. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3120–3127).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 770–778).

He, L., Liu, W., Liang, J., Zheng, K., Liao, X., Cheng, P., & Mei, T. (2021) Semi-supervised domain generalizable person re-identification. arXiv preprint arXiv:2108.05045

Hirzer, M., Beleznai, C., Roth, P. M., & Bischof, H. (2011). Person re-identification by descriptive and discriminative classification. In *Image analysis: 17th Scandinavian conference, SCIA 2011, Ystad, Sweden, May 2011* (pp. 91–102). Proceedings 17, Springer.

Huang, L., Yang, D., Lang, B., & Deng, J. (2018). Decorrelated batch normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 791–800).

Huang, L., Zhou, Y., Zhu, F., Liu, L., & Shao, L. (2019). Iterative normalization: Beyond standardization towards efficient whitening. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4874–4883).

Huang, Y., Wu, Q., Xu, J., Zhong, Y., & Zhang, Z. (2021). Unsupervised domain adaptation with background shift mitigating for person re-identification. *International Journal of Computer Vision, 129*(7), 2244–2263.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning* (pp. 448–456). PMLR.

Jia, J., Ruan, Q., & Hospedales, T. M. (2019). Frustratingly easy person re-identification: Generalizing person re-id in practice. arXiv preprint arXiv:1905.03422

Jiao, B., Liu, L., Gao, L., Lin, G., Yang, L., Zhang, S., Wang, P., & Zhang, Y. (2022). Dynamically transformed instance normalization network for generalizable person re-identification. In *European conference on computer vision* (pp. 285–301). Springer.

Jin, X., Lan, C., Zeng, W., Chen, Z., & Zhang, L. (2020). Style normalization and restitution for generalizable person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3143–3152).

Li, S., Ren, W., Wang, F., Araujo, I. B., Tokuda, E. K., Junior, R. H., Cesar-Jr, R. M., Wang, Z., & Cao, X. (2021). A comprehensive benchmark analysis of single image deraining: Current challenges and future perspectives. *International Journal of Computer Vision, 129*, 1301–1322.

Li, W., & Wang, X. (2013). Locally aligned feature transforms across views. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3594–3601).

Li, W., Zhao, R., Xiao, T., & Wang, X. (2014). Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 152–159).

Liao, S., & Shao, L. (2020). Interpretable and generalizable person re-identification with query-adaptive convolution and temporal lifting. In *European conference on computer vision* (pp. 456–474). Springer.

Liao, S., & Shao, L. (2021). Transmatcher: Deep image matching through transformers for generalizable person re-identification. *Advances in Neural Information Processing Systems, 34*, 1992–2003.

Lin, Y., Zheng, L., Zheng, Z., Wu, Y., Hu, Z., Yan, C., & Yang, Y. (2019). Improving person re-identification by attribute and identity learning. *Pattern Recognition, 95*, 151–161.

Liu, J., Zha, Z. J., Tian, Q. I., Liu, D., Yao, T., Ling, Q., & Mei, T. (2016). Multi-scale triplet CNN for person re-identification. In *Proceedings of the 24th ACM international conference on multimedia* (pp. 192–196).

Liu, J., Zha, Z. J., Chen, D., Hong, R., & Wang, M. (2019a) Adaptive transfer network for cross-domain person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. https://doi.org/10.1109/CVPR.2019.00737

Liu, J., Zha, Z. J., Hong, R., Wang, M., & Zhang, Y. (2019b). Deep adversarial graph attention convolution network for text-based person search. In *Proceedings of the 27th ACM international conference on multimedia*.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*.

Loy, C. C., Liu, C., & Gong, S. (2013). Person re-identification by manifold ranking. In *IEEE international conference on image processing* (pp. 3567–3571). IEEE.

Luo, H., Jiang, W., Gu, Y., Liu, F., Liao, X., Lai, S., & Gu, J. (2019). A strong baseline and batch normalization neck for deep person re-identification. *IEEE Transactions on Multimedia, 22*(10), 2597–2609.

Ma, L., Liu, H., Hu, L., Wang, C., & Sun, Q. (2016). Orientation driven bag of appearances for person re-identification. arXiv preprint arXiv:1605.02464

Miao, J., Wu, Y., Liu, P., Ding, Y., & Yang, Y. (2019). Pose-guided feature alignment for occluded person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 542–551).

Ni, H., Li, Y., Gao, L., Shen, H. T., & Song, J. (2023). Part-aware transformer for generalizable person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 11280–11289).

Pan, X., Luo, P., Shi, J., & Tang, X. (2018). Two at once: Enhancing learning and generalization capacities via ibn-net. In *European conference on computer vision* (pp. 464–479).

Pan, X., Zhan, X., Shi, J., Tang, X., & Luo, P. (2019). Switchable whitening for deep representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 1863–1871).

Siarohin, A., Sangineto, E., & Sebe, N. (2018). Whitening and coloring batch transform for GANs. arXiv preprint arXiv:1806.00420

Song, G., Leng, B., Liu, Y., Hetang, C., & Cai, S. (2018). Region-based quality estimation network for large-scale person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*

Song, J., Yang, Y., Song, Y. Z., Xiang, T., & Hospedales, T. M. (2019). Generalizable person re-identification by domain-invariant mapping network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 719–728).

Su, C., Li, J., Zhang, S., Xing, J., Gao, W., & Tian, Q. (2017). Pose-driven deep convolutional model for person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3960–3969).

Sun, X., & Zheng, L. (2019). Dissecting person re-identification from the viewpoint of viewpoint. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 608–617).

Tan, W., Ding, C., Wang, P., Gong, M., & Jia, K. (2023). Style interleaved learning for generalizable person re-identification. *IEEE Transactions on Multimedia*. https://doi.org/10.1109/TMM.2023.3283878

Tishby, N., & Zaslavsky, N. (2015). Deep learning and the information bottleneck principle. In *2015 IEEE information theory workshop (ITW)* (pp. 1–5). IEEE.

Ulyanov, D., Vedaldi, A., & Lempitsky, V. (2016). Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022

Vinyals, O., Blundell, C., Lillicrap, T., & Wierstra, D. (2016). Matching networks for one shot learning. *Advances in Neural Information Processing Systems, 29*, 3630–3638.

Wang, H., Zhu, X., Gong, S., & Xiang, T. (2018). Person re-identification in identity regression space. *International Journal of Computer Vision, 126*, 1288–1310.

Wang, X., Yu, F., Dunlap, L., Ma, Y. A., Wang, R., Mirhoseini, A., Darrell, T., & Gonzalez, J. E. (2020). Deep mixture of experts via shallow embedding. In R. P. Adams & V. Gogate (Eds.), *Proceedings of the 35th uncertainty in artificial intelligence conference, proceedings of machine learning research* (Vol. 115, pp. 552–562). PMLR. https://proceedings.mlr.press/v115/wang20d.html

Wei, L., Zhang, S., Gao, W., & Tian, Q. (2018). Person transfer GAN to bridge domain gap for person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 79–88).

Xiao, T., Li, S., Wang, B., Lin, L., & Wang, X. (2016). End-to-end deep learning for person search (vol. 2, no. 2, p. 4). arXiv preprint arXiv:1604.01850

Xu, B., Liang, J., He, L., & Sun, Z. (2022). Meta: Mimicking embedding via others' aggregation for generalizable person re-identification. In *European conference on computer vision* (pp. 372–388).

Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., & Hoi, S. C. (2021). Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(6), 2872–2893.

Yin, J., Wu, A., & Zheng, W. S. (2020). Fine-grained person re-identification. *International Journal of Computer Vision, 128*, 1654–1672.

Zhang, J., Niu, L., & Zhang, L. (2020). Person re-identification with reinforced attribute attention selection. *IEEE Transactions on Image Processing, 30*, 603–616.

Zhang, P., Dou, H., Yu, Y., & Li, X. (2022). Adaptive cross-domain learning for generalizable person re-identification. In *European conference on computer vision* (pp. 215–232). Springer.

Zhang, S., Zhang, Q., Yang, Y., Wei, X., Wang, P., Jiao, B., & Zhang, Y. (2020). Person re-identification in aerial imagery. *IEEE Transactions on Multimedia, 23*, 281–291.

Zhang, T., Xie, L., Wei, L., Zhuang, Z., Zhang, Y., Li, B., & Tian, Q. (2021). Unrealperson: An adaptive pipeline towards costless person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11506–11515).

Zhang, X., He, Y., Xu, R., Yu, H., Shen, Z., & Cui, P. (2023). Nico++: Towards better benchmarking for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16036–16047).

Zhang, Z., Lan, C., Zeng, W., Jin, X., & Chen, Z. (2020c). Relation-aware global attention for person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3186–3195).

Zhao, Y., Zhong, Z., Yang, F., Luo, Z., Lin, Y., Li, S., & Sebe, N. (2021). 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* (pp. 6277–6286).

Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., & Tian, Q. (2015a). Scalable person re-identification: A benchmark. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 1116–1124).

Zheng, W. S., Li, X., Xiang, T., Liao, S., Lai, J., & Gong, S. (2015b). Partial person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4678–4686).

Zheng, Z., Zheng, L., & Yang, Y. (2017). Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3754–3762).

Zhong, Z., Gao, Y., Zheng, Y., Zheng, B., & Sato, I. (2023). Real-world video deblurring: A benchmark dataset and an efficient recurrent neural network. *International Journal of Computer Vision, 131*(1), 284–301.

Zhu, X., Zhu, X., Li, M., Morerio, P., Murino, V., & Gong, S. (2021). Intra-camera supervised person re-identification. *International Journal of Computer Vision, 129*, 1580–1595.

Zhuang, Z., Wei, L., Xie, L., Zhang, T., Zhang, H., Wu, H., Ai, H., & Tian, Q. (2020). Rethinking the distribution gap of person re-identification with camera-based batch normalization. In *European conference on computer vision* (pp. 140–157). Springer.