Revisiting Stochastic Learning for Generalizable Person Re-identification

Jiajian Zhao*
Beihang University
zhaojiajian@buaa.edu.cn

Yifan Zhao*
Peking University
zhaoyf@pku.edu.cn

Xiaowu Chen
Beihang University
chen@buaa.edu.cn

Jia Li†
Beihang University
Peng Cheng Laboratory
jiali@buaa.edu.cn

ABSTRACT

Generalizable person re-identification aims to achieve a well generalization capability on target domains without accessing target data. Existing methods focus on suppressing domain-specific information or simulating unseen environments by meta-learning strategies, which could damage the capture ability on fine-grained visual patterns or lead to overfitting issues by the repetitive training of episodes. In this paper, we revisit the stochastic behaviors from two different perspectives: 1) Stochastic splitting-sliding sampler. It splits domain sources into approximately equal sample-size subsets and selects several subsets from various sources by a sliding window, forcing the model to step out of local minimums under stochastic sources. 2) Variance-varying gradient dropout. Gradients in parts of network are also selected by a sliding window and multiplied by binary masks generated from Bernoulli distribution, making gradients in varying variance and preventing the model from local minimums. By applying these two proposed stochastic behaviors, the model achieves a better generalization performance on unseen target domains without any additional computation costs or auxiliary modules. Extensive experiments demonstrate that our proposed model is effective and outperforms state-of-the-art methods on public domain generalizable person Re-ID benchmarks.

CCS CONCEPTS

• **Computing methodologies** → **Object identification**; **Object recognition**.

KEYWORDS

person re-identification, domain generalization, stochastic behaviors

*Both authors contribute equally to this work.
†Correspondence should be addressed to Jia Li.
Website: http://cvteam.buaa.edu.cn/

MM '22, October 10–14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3547812

ACM Reference Format:
Jiajian Zhao, Yifan Zhao, Xiaowu Chen, and Jia Li. 2022. Revisiting Stochastic Learning for Generalizable Person Re-identification. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3503161.3547812

1 INTRODUCTION

Person re-identification (Re-ID) aims to retrieve images of the same person identity captured by non-overlapped cameras in a gallery when given a query image. For decades, advanced deep learning methods [11, 31, 33, 52] and proposals of large person datasets [38, 46–48] greatly boost the development of supervised person Re-ID tasks. However, these supervised methods suffer from significant performance degradation when directly applied in unseen domains, owing to the strong inductive biases in learning systems. To reduce these biases, unsupervised domain adaptation methods [5, 44, 53] are proposed to exploit unlabeled data in target domains to further finetune these well-trained models. However, when applying to the real-world applications, the target domain knowledge is usually unknown, *e.g.*, surveillance videos with different lightness or occlusions, existing domain adaptation methods fail to handle all circumstances with the various change of unknown data. To tackle this challenge, researchers resort to the task of Domain Generalizable (DG) person Re-ID whose goal is to train a model with a strong generalization capability on unseen target domains by only using labeled source domain data.

Different from other domain generalization tasks on common data, Re-ID datasets have strong similarities in their distributions (*e.g.*, head-body-leg from top to bottom), leading to severe overfitting in the learning process. Hence in the domain generalizable Re-ID task, one intuitive and prevailing idea is to prevent the model from over-fitting in source domains. Motivated by this idea, most of existing research concentrates on two lines of techniques: domain-specific information suppression and meta-learning strategies. For example, Jin *et al.* [14] utilized an instance normalization operation to eliminate style discrepancies among different identities, enhancing the identity-irrelevant features with the proposed restitution module. In [3, 29, 45], meta-learning strategies were adopted to simulate the process of model inference during training stages to reduce domain shifts between training and testing. Besides these methods, Dai *et al.* [4] combined these two lines of techniques to extract source-discriminative characteristics and distill irrelevant

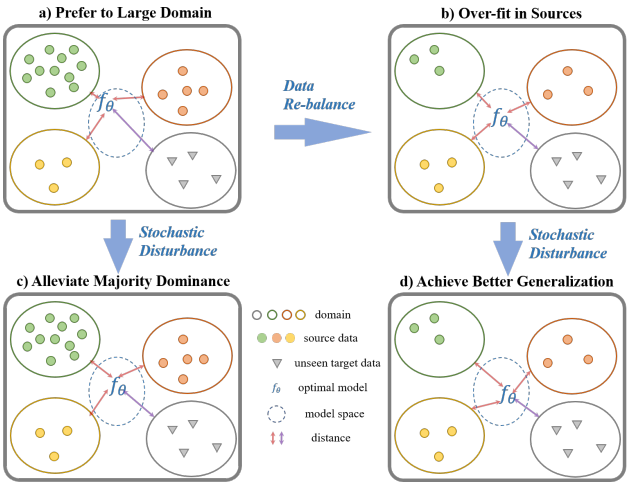


Figure 1: The motivation of our proposed method. a) On multiple sample-imbalanced source domains, one model prefers to learn characteristics from sources with majority samples and neglects the minority, hindering the model from extracting generalizable representations. b) A common way is to re-balance sample size in each source, making the model learn an 'average' representation that is only fitting to sources. c) and d) Taking stochastic disturbances in training stage, the model is forced to jump out of current local minimums and search for a more generalizable solution in feature spaces.

information in a meta-learning manner. Although these two lines of techniques show their effectiveness in generalizable representations, major deficiencies behind these ideas are still under-explored: (1) the fine-grained recognition especially re-identification heavily relies on low-level patterns, including colors and textures, but these patterns are easily damaged by the domain-specific feature suppressing procedure; (2) although meta-learning methods tend to improve the generalization ability of training model by evaluating its performance on pseudo test sets. These pseudo data still come from the known source domain and also lead to over-fitting issues by the repetitive training of hundreds of episodes.

Keeping these concerns in mind, a question naturally arises: how to prevent the model stuck in over-fitted local minimums without additional costs and loss of useful information? In this paper, we propose to revisit the domain generalizable person Re-ID from two distinctive views: 1) the data distribution in source sampling and 2) the gradient back-propagation in training process. Numerous researches [7, 9, 32] have shown that learnable models incline to be bias to majority classes under the data-imbalance situation. Guiding by this finding, if we take one domain as a class, a domain generalization model tends to learn characteristics of domains with majority proportions of data and fails to keep informative representations of minority domains, as shown in Fig. 1 a). The most intuitive way to alleviate this issue is to re-balancing samples of each source domain. However, this training manner only forms an *averaged* representation of known sources but could not help the model out of over-fitting issues. Thus we propose to take the advantage of stochasticity by the nature of dataset itself, promoting a model to step out of local minimums owing to the instability of sampled data. We develop a novel stochastic splitting-sliding sampler to re-balance source domains and take stochasticity into the model with a sliding-window behavior. There are two typical steps in the proposed stochastic splitting-sliding sampler: 1) it first splits each source domain into several subsets with approximately equal sample sizes and then orders these subsets into a queue where each subset and its neighborhoods belong to different sources; 2) a sliding window is then taken to select subsets whose sizes are less than the number of domains in source data. The sliding-window behavior guarantees that domains in each adjacent episode are various, forcing the model to jump out of the local minimum in current sources, which can be found in Fig. 1 d).

Expect for the stochastic behavior in the data view, we revisit another stochastic technique in the optimization process: dropout. Vanilla dropout [30] is effective to suppress the over-fitting issue, but applying a dropout layer between learnable blocks would change network structures and further harm the pre-training knowledge. To solve this problem, several works [2, 35] propose to drop the gradient in all layers with a probability generated by a strategy, *e.g.*, Bernoulli distribution. However, it would lead to severe problems to directly drop out gradients in a p probability Bernoulli distribution: if the model has been already stuck in local minimums in Fig. 1 a) and b), applying dropout to gradients of all layers only changes the variance of p times. Besides theoretical analyses, we propose a novel variance-varying gradient dropout with sliding windows of several layers, keeping the stochasticity in learning process. In this dropout, we first divide the backbone into several groups, *e.g.*, taking blocks in a stage of ResNet as a group. Then gradients of the groups selected with the sliding window are multiplied by binary masks generated by a Bernoulli distribution. By our proposed variance-varying gradient dropout, the gradients of the model can vary constantly, further to achieve generalization capability as in Fig. 1 c) and d). Our proposed method does not require any additional computation costs or auxiliary training modules and achieves the state-of-the-art performance in domain generalizable person re-identification tasks.

Our main contributions in this paper are three-fold:

1) In the view of data distribution, we design a novel stochastic splitting-sliding sampler to split imbalanced sources into several subsets with approximately equal sample sizes and utilize a sliding window to select subsets in different sources in each training episode, preventing the model from preferring to source domains.

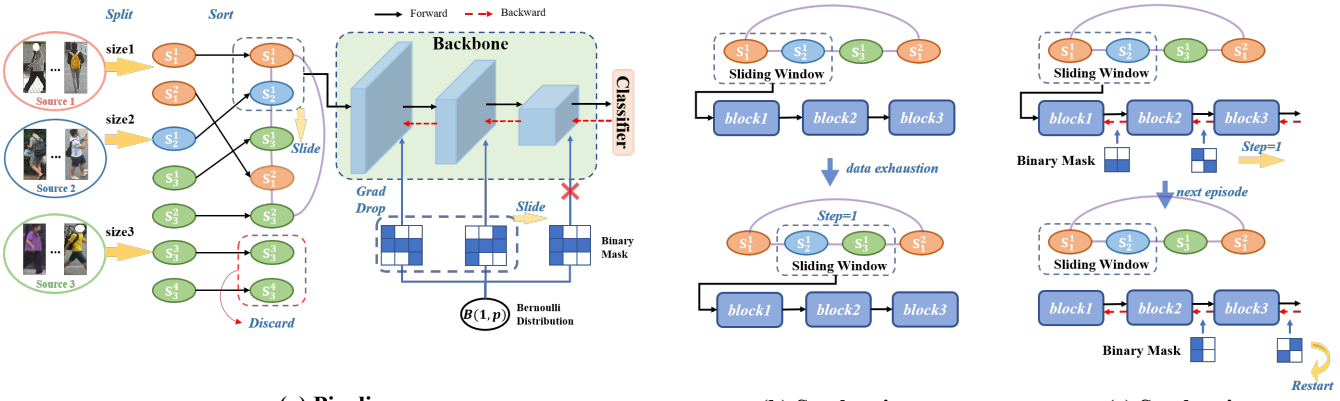
2) In the view of optimization, we propose a variance-varying gradient dropout to set the variance of gradients in a constant change state, promoting the model jump out of the local minimum.

3) We make theoretical and experimental analyses to reveal the importance of stochastic training in generalization and conduct extensive experiments to verify the superiority of the proposed method on public domain generalizable person Re-ID benchmarks.

2 RELATED WORK

2.1 Person Re-identification

Deep supervised person re-identification in a single domain have gained a significant improvement for decades, including three main components [43]: Feature Representation Learning [31, 33, 48, 50],



(a) Pipeline

(b) Stochastic Domain Sources

(c) Stochastic Variance GradDrop

Figure 2: Illustration of our approach. The pipeline consists of the network structure, stochastic splitting-sliding sampler and variance-varying gradient dropout. The proposed sampler splits K source domains into several approximately equal sample-size subsets and sorts these subsets in the condition (3) to form a circular queue. Then a sliding-window behavior is adopted to select $L_S(< K)$ subsets in the queue from head to tail in T_S steps to keep data from stochastic sources in two adjacent episodes, preventing over-fitting in sources. The proposed gradient dropout divides the network into several groups and also takes a sliding window in T_G steps to select L_G groups. Gradients of these groups will multiply by a binary mask generated by Bernoulli distribution with probability p to keep gradient variance in a constant change state, helping to step out of local minimums.

Deep Metric Learning [11, 39, 41] and Ranking Optimization [28, 52]. Further to promote Re-ID methods to be applied in the real world, unsupervised domain adaptation (UDA) methods are proposed to eliminate domain gaps without labeling the target dataset in recent years. The unsupervised domain adaptation can be roughly categorized into two groups which are GAN-based transfer [5, 38] and target domain supervision mining [44, 53]. In [38], PTGAN was proposed to realize the style transfer between different domain images, aiming to bridge the domain gap. Unlike using GAN, Zhong *et al.* [53] designed an exemplar memory to memory target domain features with three invariance properties. For domain generalizable person re-identification [1, 3, 4, 13, 14, 18, 22, 29, 34, 37, 54], it can be trained only in the source domains and directly evaluate in unseen target domains. Song *et al.* [29] proposed Domain-Invariant Mapping Network to learn domain-invariant features by utilizing a hyper-network and memory bank. Inspired by the style variations suppression of IN, Jin *et al.* [14] developed a Style Normalization and Restitution module to learning a robust representation by eliminating the identity-irrelevant features. Choi *et al.* [3] combined batch normalization and instance normalization by a learnable parameter, learned in a meta-learning manner. Different from the above methods, we propose to adopt stochastic behaviors to avoid over-fitting in source domains without any additional computation costs or auxiliary training modules.

2.2 Domain Generalization

Domain Generalization (DG) has attracted more and more researchers in the past ten years. Domain-invariant representation learning [17, 21, 25, 26] and data augmentation [36, 42, 55, 56] are two main methods in domain generalization. The crucial thought in domain-invariant representation is to minimize the domain gap among source domains, *i.e.*, reducing the negative effect caused by domain shifts in the final representation. Data augmentation aims to simulate various domain shift by augmenting the source data, which makes the model avoid over-fitting in source domain characteristics. Besides two mainstreams, researchers also explore other DG methods, *e.g.*, meta learning [16], casual matching [24], disentangled representation [15] and adaptive methods [6]. For domain generalization, source domains and target domains are in the same label space. However, the label spaces of unseen target domains in generalizable person re-identification are completely different from source domains. This means that preventing Re-ID from over-fitting in source domains is more important than learning domain-invariant features.

2.3 Gradient Dropout

Dropout [30] is a simple yet effective technique to prevent neural networks from overfitting by dropping out network units with probability p at the training time. In [2], Chen *et al.*designed a GradDrop to balance gradients derivated by multiple losses in multitask. Tseng *et al.* [35] applied the GradDrop method to meta learning tasks and presented two Bernoulli and Gaussian dropout terms. In this paper, we adopt gradient dropout in a sliding-window manner to set gradients of different network layers in various variances, ensuring that a model can step out of local minimums.

3 METHODOLOGY

3.1 Overview

In this section, we elaborate the pipeline of the proposed method which is exhibited in Fig. 2 (a). In training stages of this generalization task, there exist K source domains $\mathcal{D} = \{\mathcal{D}_k\}_{k=1}^{K}$, where

$\mathcal{D}_k = \{(x_k^n, y_k^n, d_k)\}_{n=1}^{N_k}$ includes N_k image-identity pairs with domain label d_k in the k-th domain. Firstly, we utilize the proposed splitting-sliding sampler $\mathcal{S}$ to split each source domain into approximately equal sample-size subsets, where $\{\mathcal{D}_k^i\}_{i=1}^{N_s^k}$ represents k-th source domain is split into N_s^k subsets. Then these subsets are rearranged as a queue where each subset and its neighbors should belong to different domains, and $L_S(< K)$ subsets are taken by a sliding window in one episode, as illustrated in Fig. 2 (b). After a minibatch of input is sampled, we feed the input into a backbone $\mathcal{F}_\phi$ followed by a classifier $\mathcal{F}_\psi$ and compute gradients with back-propagation. Keeping gradient update in various variance space, we adopt a variance-varying gradient dropout $\mathcal{G}$ to divide gradients of $\mathcal{F}_\phi$ and $\mathcal{F}_\psi$ into several groups and also adopt a sliding-window behavior to mask gradient groups $\mathcal{W}$ in the range of window with a p probability Bernoulli distribution, as shown in Fig. 2 (c).

3.2 Stochastic Splitting-sliding Sampler

Existing pioneer works have demonstrated that deep models incline to keep memories of visual patterns from majority classes which own most of training samples. This leads the trained model to perform a high accuracy on majority classes but an inferior performance on minority classes, damaging the generalization capability. Taking a domain analogy to a class, total samples in different source domains are seriously imbalanced, *e.g.*, about 7,000 samples in CUHK02 [19] and 36,000 samples in DukeMTMC-reID [51]. If we directly combine data from multiple domains in a training episode, the model would only remember characteristics of majority source domains and forget minority ones. To alleviate over-fitting issues from the perspective of data distributions, we resort to the stochastic splitting-sliding sampler, typically consisting of two steps.

In the first step, source domains are divided into several subsets by a split operation. We determine a subset size S_l empirically, which should be not greater than the minimal size of each domain. In this manner, training samples of each source domain are split into several subsets with S_l samples in each. However, it is usually infeasible to divide these datasets to the size of S_l with no remainder. If we promise a strictly equal sample-size subset, two common ways are discarding the remainder or repeating samples to make up for the lack. However, no matter which solution we take, it is sticky to choose which samples to be discarded or repeated. The third solution is to guarantee the size of first few subsets is equal to S_l and the last subset includes all the remainder. Further to avoid only a few in the last subset and keep all subset sizes as equal as possible, we propose to utilize a rounding to compute an approximately equal sample size S_l^k for the k-th source domain, computed by:

$$S_l^k = \lfloor \frac{S_k}{\mathcal{R}(S_k/S_l)} \rfloor, \tag{1}$$

where S_k is the size of the k-th source domain and $\mathcal{R}$ is a rounding operation. After simple rounding, we can prevent the size of last subset from becoming a outlier under the approximate equality of S_k and S_l. In the first step, balanced subsets $\{\mathcal{D}_k^i\}_{i=1}^{N_s^k}$ in k-th source domain are obtained, which can be defined by

$$\mathcal{D}_k^i = \begin{cases} \{(x_k^{(i-1)S_l^k+j}, y_k^{(i-1)S_l^k+j}, d_k)\}_{j=1}^{S_l^k} & i < N_s^k \\ \{(x_k^j, y_k^j, d_k)\}_{j=(i-1)S_l^k}^{N_k} & i = N_s^k \end{cases}, \tag{2}$$

where N_k is the number of total samples in k-th source domain, and N_s^k is the number of subsets in k-th source domain and equal to $\mathcal{R}(S_k/S_l)$.

In the second step, we develop a sliding window strategy SW_S to prevent the model from over-fitting in source domains. After each source domain has been split into several subsets with nearly the same size in the first step, the next crucial problem is how to help the model to jump out of the local minimum. As shown in Fig. 1 b), feeding re-balanced samples of different source subsets into the model is effective to alleviate the preference for large-scale domains and the optimal model is robust to source domains. However, our key idea to solve domain generalizable Re-ID tasks is that one model should achieve generalizable capabilities on unseen target domains rather than robustness on source domains. The optimal

Algorithm 1: Training Scheme of our proposed method

Input: Source domains $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_k\}$,pre-trained parameters $\theta_{\mathcal{F}_\phi}$,hyperparameters α, β
Output: Trained parameters $\theta_{\mathcal{F}_\phi}, \theta_{\mathcal{F}_\psi}$

Initialize classifier parameters $\theta_{\mathcal{F}_\psi}$
for $k = 1, 2, \ldots, K$ **do**
  Split k-th source domain as N_s^k subsets $\{\mathcal{D}_k^i\}_{i=1}^{N_s^k}$ // Eq.(1)-Eq.(2)
end
Sort all subsets $\bigcup_{k=1}^{K}\{\mathcal{D}_k^i\}_{i=1}^{N_s^k}$ followed by the condition (3) to form a sorted set $\mathcal{D}^{'}$
Keep at most $L_S + \lfloor \frac{T_S}{2} \rfloor$ subsets from the same domain in the tail of $\mathcal{D}^{'}$
Divide $\theta_{\mathcal{F}_\phi}$ and $\theta_{\mathcal{F}_\psi}$ to several groups $\mathcal{W}$
$\mathcal{D}_S^{'} = SW_S(\mathcal{D}^{'}, L_S)$ // sliding operation in Sec 3.2
$\mathcal{M}^{'} = SW_G(\mathcal{W}, p, L_G)$ // sliding operation in Sec 3.3
for *epoch* $= 1, 2, \ldots$ **do**
  if *epoch* % $N_G == 0$ **then**
    $\mathcal{M}^{'} = SW_G(\mathcal{W}, p, L_G, T_G)$
  end
  Sample a mini-batch $\mathcal{B}$ from $\mathcal{D}_S^{'}$
  if *data exhaustion* **then**
    $\mathcal{D}_S^{'} = SW_S(\mathcal{D}^{'}, L_S, T_S)$
  end
  Compute $\mathcal{L}_E(\theta_{\mathcal{F}_\phi}, \theta_{\mathcal{F}_\psi}, \mathcal{B})$
  $g = \nabla\mathcal{L}_E(\theta_{\mathcal{F}_\phi}, \theta_{\mathcal{F}_\psi}, \mathcal{B})$
  $g^{'} = g \cdot \mathcal{M}^{'}$
  $\theta_{\mathcal{F}_\phi} \leftarrow \theta_{\mathcal{F}_\phi} - \alpha \cdot g_{\mathcal{F}_\phi}^{'}$
  $\theta_{\mathcal{F}_\psi} \leftarrow \theta_{\mathcal{F}_\psi} - \beta \cdot g_{\mathcal{F}_\psi}^{'}$
end
return trained network parameters $\theta_{\mathcal{F}_\phi}, \theta_{\mathcal{F}_\psi}$

solution in source domains is not suitable for unseen domains due to large distribution discrepancies between source and unseen domain targets. Thus it is necessary to apply stochastic disturbances for break the model training from local optimal traps. For this intuition, one feasible way is to select part of source domains in a training episode. Various domains force the model to continuously search for a more generalizable solution. Meanwhile, it is important to avoid several domains rarely selected due to randomness. Hence we propose to apply a sliding-window behavior in the ordered subset circular queue. All split subsets $\cup_{k=1}^{K}\{\mathcal{D}_k^i\}_{i=1}^{N_s^k}$ are sorted according to the i, and the k for subsets if i is identical, defined as

$$\begin{cases} Pos_{\mathcal{D}'}(\mathcal{D}_a^i) < Pos_{\mathcal{D}'}(\mathcal{D}_b^j) & i < j \ \ \& \ \ \forall a, b \leq K \\ Pos_{\mathcal{D}'}(\mathcal{D}_a^i) < Pos_{\mathcal{D}'}(\mathcal{D}_b^i) & a < b \end{cases}, \tag{3}$$

where $\mathcal{D}^{'}$ is the sorted set and $Pos_{\mathcal{D}'}$ represents the position number of $\mathcal{D}^{'}$. Then this operation utilizes a sliding window with a size of L_S to select L_S subsets from $\mathcal{D}^{'}$. After samples of these subsets are exhaustively enumerated, we move the sliding window forward T_S steps, as shown in Fig. 2 (b). In the front of the queue, there are approximately equal training samples in each selected source but domains are various in two adjacent episodes. The source domain without appearing in the last episode can force the model to jump out of the over-fitting in sources, as shown in Fig. 1 d). Due to the imbalance among source domains, subsets from the minority are depleted and disappear in the tail of the queue. When the number of rest domains is less than window size L_S, there does not exist an unseen domain in the next episode. However, sliding-window behavior still keeps the model in a stochastic state by setting $T_S < L_S$. When $T_S < L_S$, samples in $L_S - T_S$ subsets are reused and domains which $L_S - T_S$ subsets belong to hold the dominance. By careful design, reused domains are different in most of episodes. Moreover, if there exists a source domain that is much larger than the others, it would lead to severe over-fitting phenomenon because only subsets from this domain appear in the late training episode. To solve the problem, we just keep at most $L_S + \lfloor \frac{T_S}{2} \rfloor$ subsets of this domain in the tail of the queue and discard the rest, ensuring that newly selected subsets in two continuous episodes are mainly from different domains.

3.3 Variance-varying Gradient Dropout

Dropout [30] is a powerful stochastic behavior to alleviate neural networks from over-fitting, which effectiveness has been widely verified. However, when the Dropout operation is inserted into a model, it usually changes the model structure and damages the pre-trained knowledge. For downstream tasks, it brings a huge impact because the pre-trained knowledge is crucial for the downstream tasks to converge faster while achieving a generalization simultaneously. To avoid losing this key knowledge, recent researches propose the gradient dropout operations [2, 35], which apply dropout in the gradient updating process:

$$g^{'} = \mathcal{M} \cdot g, \tag{4}$$

where g is original gradients yield from back-propagation, $g^{'}$ is gradients after dropout and $\mathcal{M}$ is a binary mask generated by a dropout strategy. For simplicity, we choose a Bernoulli distribution with a probability p as done in [30] to generate a binary mask.

In domain generalizable Re-ID task, our motif is to prevent the model from over-fitting in source domains. If we directly mask all gradients of the model with p probability, it would not promote the model to jump out of the local minimum in certain cases because this gradient dropout behavior only makes the original gradient variance multiplied by p, proved as follows. In the optimization, the purpose is to minimize the expected risk $\mathcal{L}_E$, defined as

$$\mathcal{L}_E = \mathbb{E}_{(x,y)\in\mathcal{D}}[l(\mathcal{F}_\psi(\mathcal{F}_\phi(x)), y)], \tag{5}$$

Lemma 1. *When a model converges into a local minimum value, the gradient g derived by $\mathcal{L}_E$ is approximately equal to zero,* i.e., $g \approx 0$.

Proof. From the definition of the local minimum, it is easy to conclude that weights in the model tend to be stable and gradients approach to be zero values.

Proposition 1. *Let the binary mask $\mathcal{M}$ from Bernoulli distribution with a probability p and the new gradient $g^{'} = \mathcal{M} \cdot g$, the expectation $\mathbb{E}$ of $g^{'}$ approaches to be zero and the variance approaches to p times the variance of g, when the model traps into a local minimum value.*

Proof. When the model reaches a local minimum during training, *i.e.*, $g \approx 0$, the expectation $\mathbb{E}[g] \approx 0$. For the binary mask $\mathcal{M}$ independent of gradients, the expectation of $g^{'}$ is

$$\begin{aligned} \lim_{\mathcal{L}_E \to \mathcal{L}_E^*} \mathbb{E}[g^{'}] &= \lim_{\mathcal{L}_E \to \mathcal{L}_E^*} \mathbb{E}[\mathcal{M}]\mathbb{E}[g] \\ &= \lim_{\mathcal{L}_E \to \mathcal{L}_E^*} p\mathbb{E}[g] \\ &= 0 \end{aligned}, \tag{6}$$

where $\mathcal{L}_E^*$ represents the local minimum loss.

Then the variance of $g^{'}$ can be defined as:

$$\begin{aligned} \lim_{\mathcal{L}_E \to \mathcal{L}_E^*} Var[g^{'}] &= \lim_{\mathcal{L}_E \to \mathcal{L}_E^*} \mathbb{E}[\mathcal{M}^2 g^2] - \mathbb{E}^2[\mathcal{M}g] \\ &= \lim_{\mathcal{L}_E \to \mathcal{L}_E^*} \mathbb{E}[\mathcal{M}^2]\mathbb{E}[g^2] \\ &= \lim_{\mathcal{L}_E \to \mathcal{L}_E^*} p\mathbb{E}[g^2] - 0 \\ &= \lim_{\mathcal{L}_E \to \mathcal{L}_E^*} p(\mathbb{E}[g^2] - \mathbb{E}^2[g]) \\ &= pVar[g] \end{aligned}. \tag{7}$$

Proposition 2. *Let the binary mask $\mathcal{M}$ sampled from Bernoulli distribution with a probability p and the new gradient $g^{'} = \frac{\mathcal{M}\cdot g}{p}$, the expectation of $g^{'}$ approaches to be zero and the variance approaches $\frac{1}{p}$ times the variance of g when the model traps into a local minimum.*

Proof. Replace $\mathbb{E}[\mathcal{M}]$ and $\mathbb{E}[\mathcal{M}^2]$ with $\frac{\mathbb{E}[\mathcal{M}]}{p}$ and $\frac{\mathbb{E}[\mathcal{M}^2]}{p^2}$ in the proof of Proposition 1. Detailed proofs are elaborated in Appendix.

In a fixed variance space, the magnitude of the gradient is limited, and the model cannot escape from local minimums, leading to a generalization degradation. Therefore, we propose to only apply gradient dropout in several model layers, keeping the variance of the gradient with constant change. As the stochastic splitting-sliding sampler, we adopt a sliding-window behavior SW_G to achieve various gradient variances. Firstly, backbone $\mathcal{F}_\phi$ is divided into z groups according to several modules, *e.g.*, blocks in a ResNet stage. Then these groups and classifier $\mathcal{F}_\psi$ are unified into a candidate sequence. Before gradient update in each iteration, gradients of layers are selected by a sliding window with the size L_G, and multiplied with a binary mask generated by Bernoulli distribution with a probability p. After a training episode (N_G epochs), the sliding window moves forward T_G step(s) in the group sequence to change gradient

Table 1: Performance (%) comparisons with state-of-the-arts on DG Re-ID benchmarks.

Source	Method	VIPeR				PRID				GRID				i-LIDS			
		R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP
M+C2+ D+C3+ CS	DIMN [29]	51.2	70.2	76.0	60.1	39.2	67.0	76.7	52.0	29.3	53.3	65.8	41.1	70.2	89.7	94.5	78.4
	MetaBIN [3]	59.9	**78.4**	82.8	68.6	74.2	89.7	92.2	81.0	48.4	**70.3**	77.2	57.9	81.3	95.0	97.0	87.0
	DualNorm [13]	59.4	-	-	-	69.6	-	-	-	43.7	-	-	-	78.2	-	-	-
	SNR [14]	52.9	-	-	61.3	52.1	-	-	66.5	40.2	-	-	47.7	84.1	-	-	89.9
	RaMoE [4]	56.6	-	-	64.6	57.7	-	-	67.3	46.8	-	-	54.2	**85.0**	-	-	**90.2**
	Baseline	56.9	73.4	80.0	64.6	59.2	81.7	89.3	69.6	43.0	64.0	73.1	53.2	77.0	92.2	96.7	83.7
	Ours	**61.5**	77.3	**82.9**	**69.2**	**74.4**	**91.7**	**96.4**	**81.6**	**49.9**	70.0	**78.6**	**59.6**	80.7	**95.7**	**97.7**	87.2
M+D+ C3+MT	SNR [14]	55.1	-	-	65.0	49.0	-	-	60.0	30.4	-	-	41.3	87.0	-	-	91.9
	RaMoE [4]	63.4	-	-	72.2	56.9	-	-	66.8	43.4	-	-	**53.9**	**88.4**	-	-	**92.3**
	Baseline	59.8	79.7	84.5	68.8	55.8	76.9	82.4	65.5	34.5	57.9	64.1	45.4	76.0	90.3	95.5	82.6
	Ours	**65.1**	**81.2**	**87.1**	**72.6**	**71.6**	**86.6**	**90.9**	**78.9**	**43.8**	**63.1**	**71.0**	52.5	80.7	**94.7**	**97.3**	86.9

Table 2: Ablation studies of our proposed method.

Source	Sampler	GradDrop	PRID		GRID	
			R-1	mAP	R-1	mAP
M+C2+ D+C3+ CS			59.2	69.6	43.0	53.2
	✓		73.0	79.6	49.5	58.9
		✓	68.0	76.7	**51.8**	59.3
	✓	✓	**74.4**	**81.6**	49.9	**59.6**

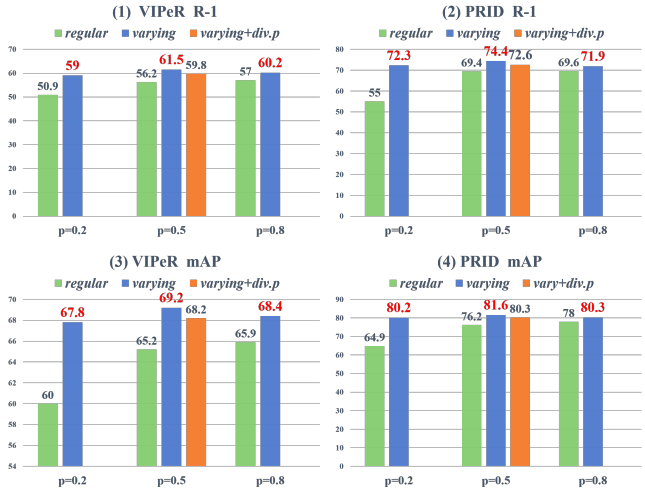


Figure 3: Comparisons between variance-varying gradient dropout and regular gradient dropout. Regular gradient dropout is sensitive to the probability value while proposed gradient dropout is stable in high performances. It demonstrates that our proposed method whether to divide by p is effective to prevent over-fitting compared with the regular.

variance. After moving to the last group in the sequence, the sliding window will restart from the first group.

4 EXPERIMENTS

4.1 Datasets and Evaluation Settings

To evaluate the effectiveness of our proposed method to improve the generalization capability, we conduct experiments on the public person Re-ID datasets, In the DG Re-ID task, source domains include 6 datasets CUHK02 [19], CUHK03 [20], Market1501 [47], DukeMTMC-reID [51], MSMT17 [38] and CUHK-SYSU Person-Search [40]. Unseen target domains are 4 small Re-ID datasets which are VIPeR [8], PRID [12], GRID [23], and QMUL i-LIDS [49]. We denote CUHK02 as C2, CUHK03 as C3, Market1501 as M, DukeMTMC-reID as D, CUHK-SYSU as CS, and MSMT17 as MT. In training stage, both train and test split subsets on each source domain are used. In evaluation stage, the evaluation protocol follows by [29] on four unseen target domains. Evaluation metrics are mean Average Precision (mAP) and Cumulative Matching Characteristic (CMC) at Rank-k, commonly adopted in Re-ID task.

4.2 Implementation Details

We adopt ResNet-50 with ibn-a blocks [27] pre-trained on ImageNet as our backbone and a classifier with total identities on all training source domains. In training stage, the mini-batch size is 64 (32 IDs, 2 instances). We take the random flipping with a probability of 0.5 and pad 10 pixels on the image border, then randomly cropped to 256 × 128. The model is trained for 100 epochs by SGD optimizer with a momentum of 0.9 and a weight decay of 5e-4. The learning rate warms up from 7.7e-5 to 1e-2 in the first 10 epochs and the backbone is frozen in the warm-up time. Then the learning rate is divided by 10 in the 50th and 90th epochs, respectively. The learning rate of classifier is multiplied by 10, followed by [13]. For the proposed gradient dropout, we divide the network into 6 groups which the number of groups is 5 in backbone and 1 in classifier. In ResNet, the first group contains the first convolution layer and BN layer. The other groups correspond to each ResNet stage, respectively.

Table 3: Analyses of our proposed method on extremely imbalanced sources. SW: sliding-window behavior. sub: subsets

Source	Method	VIPeR		PRID	
		R-1	mAP	R-1	mAP
M+D+C3+MT	Baseline	59.8	68.8	55.8	65.5
	GradDrop	63.1	71.6	61.6	72.7
	SW W/o split	62.2	71.3	61.7	72.4
	SW+all sub	58.6	67.3	52.3	62.4
	SW+repeat sub	59.3	68.9	60.8	69.1
	SW+discard sub	62.2	71.0	67.4	75.5
	+GradDrop	**65.1**	**72.6**	**71.6**	**78.9**

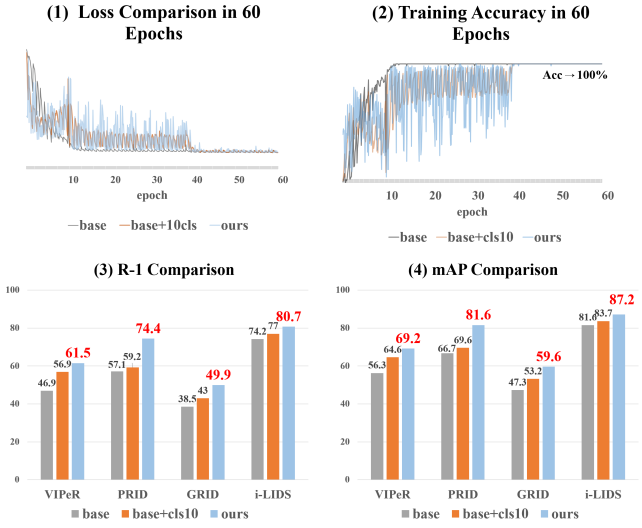


Figure 4: Quantitative analyses on the effectiveness of stochasticity. By comparison between the baseline with 1x learning rate and it with 10x learning rate in classifier, the stochasticity can effectively improve the generalization capability. By observing the loss curve and the training accuracy curve, it verifies our approach brings more positive stochasticity and prevent a model from over-fitting.

4.3 Comparisons with State-of-the-Arts

In this section, we evaluate our proposed method with state-of-the-arts on two DG Re-ID benchmarks, as shown in Tab. 1.

Sources: M+C2+D+C3+CS. Among all methods, RaMoE [4] achieved impressive results on i-LIDS dataset, the smallest target domain. Our proposed method outperforms significantly RaMoE on three other larger datasets. MetaBIN [3] has a superior performance on VIPeR dataset in evaluation metric R-5, but our method performs better on other target domains and the mAP on four datasets is higher than MetaBIN. Meanwhile, we trained MetaBIN and our model for 184,000 iterations with one RTX3090 NVIDIA GPU and the same coding framework [10]. The training time of MetaBIN is about 40 hours by meta learning while ours is about 9 hours.

Table 4: Analyses of subset size and sliding window in proposed sampler.

Source	Sub.size	Win.size	Win.step	mAP	
				PRID	GRID
M+C2+D+C3+CS	2000	4	3	78.9	57.6
	3500			78.7	58.3
	7264 (min.source)			**81.6**	**59.6**
		5	5	79.9	56.0
		4	4	79.8	57.7
		4	2	77.6	59.2
		3	3	76.9	55.9
		3	2	77.1	58.5
		2	1	70.2	59.5

Sources: M+D+C3+CS. Our method outperforms RaMoE in three larger datasets. It verifies that ours is effective to help the model step out of local minimums and search for a better solution, *i.e.*, improving holistic generalization capability of a model. Moreover, similar performances on two benchmarks further demonstrate that our model is stable and effective.

4.4 Perfomance Analyses

Ablation studies of proposed method. In Tab. 2, it is obvious that two proposed methods can greatly improve the generalization capability, respectively. Compared with the respective roles of the two proposed methods, the combination of them achieves a more generalizable performance, especially in mAP.

Analyses of the proposed method on extremely imbalanced sources. For the C2-C3-M-D-CS setting, the number of samples on these domain is relatively balanced. Further to demonstrate our proposed sampler is effective to tackle the problem of extremely imbalanced source domains, we conduct experiments on C3-M-D-MT setting. The number of samples on MSMT17 dataset is about 126,000 while three other dataset is about 80,000 in total. Therefore, there exists extreme imbalance among these source domains. As shown in Tab. 3, we verify different operations of our proposed sampler. If directly combine all source domains into training, the performance on VIPeR and PRID is unsatisfactory, but it has an improvement by taking our sliding-window behavior for sampling or applying the proposed gradient dropout due to preventing the model from preferring to dominant source domain. After splitting source domains and sorting subsets on the condition (3), taking all subsets in the sorted queue significantly degrades the generalization capability because subsets in the tail of the queue all belong to MSMT17, misleading the model to trap into the local minimum on MSMT17. When re-balancing source domains by repeating subsets to make sure each source domain with the same number of subsets, its generalization improvement is limited. The proposed operation that only keeps $L_S + \lfloor \frac{T_S}{2} \rfloor$ subsets of MSMT17 in the tail of the queue and discards others achieves impressive results. This operation not only bring a better generalization capability but also needs less samples on the absolutely dominant source domains.

Analyses of the proposed gradient dropout and regular gradient dropout. We demonstrate the advantage of the proposed

Figure 5: The t-SNE visualization on four unseen target datasets. From the visualization, the baseline fails to extract discriminative identity features on unseen targets due to over-fitting in sources, but our method is effective to pull the distance between two features with the same identity, illustrating that ours promotes the baseline to achieve a more generalizable capability.

gradient dropout compared with the regular gradient dropout. In the regular gradient dropout, gradients of all layers in the network will be dropped with a probability, which just changes the original variance proportional to the probability p in local minimums. In the fixed variance, it is very possible for the model to jump out of the local minimum unsuccessfully. In Fig. 3, we conduct experiments among regular gradient dropout, variance-varying gradient dropout and variance-varying gradient dropout divided by probability p, which are abbreviated as regular, varying and varying div.p. These three gradient dropouts all take Bernoulli distribution to generate binary masks. It is observed that the probability of the regular can greatly influence the generalization performance on unseen target domains. However, by adopting a sliding window to force the gradient in constantly varying variance, it is effective for the model to prevent over-fitting in sources. From Fig. 3, the generalization capability of the model is stable and completely outperforms regular gradient dropouts whatever the dropout probability is. Moreover, the method with masked gradients divided by p also achieves excellent generalization performance, exceeding the regular method. Nonetheless, it is inferior to the original method, because its larger gradient variance makes the model update in a large step.

Quantitative analyses on the effectiveness of stochasticity. Further to demonstrate that stochasticity can improve the generalization capability and verify the ability of our method to bright the positive stochasticity for a model, we conduct experiments on three settings: 1) origin baseline; 2) baseline with ten times learning rate of classifier; 3) baseline with ten times learning rate of classifier and our two proposals. Experimental results are exhibited in three aspects which are loss, training accuracy and evaluation metrics on four unseen target domains. In Fig. 4, when the learning rate of classifier increases tenfold, the loss and training accuracy obviously fluctuate back and forth, meaning the model is in stochasticity. This stochasticity leads to a higher performance of R-1 and mAP on targets. Compared with setting classifier at ten times learning rate, our proposed methods promote the model to experience more violent randomness and experiments demonstrate that randomness brought by the proposal is positive.

Analyses of subset size and sliding window in proposed sampler. As shown in Tab. 4, we conduct experiments about three attributes in our proposed sampler: 1) the subset size (S_l in Sec 3.2); 2) the sliding window size; 3) the sliding window step. For the subset size, it is the best solution to directly select the minimum size among source domains. If the size of subset is small, data exhaustion reaches in a short time. It will cause frequent domain changes and hinder the model learn a stable representation. For window size, the small size means that the model cannot learn more complex cross-domain information. For window step, it is beneficial to retain one of the subsets in the last episode.

The t-SNE visualization. Intuitive to demonstrate our method can effectively improve generalization capability compared with the baseline, we visualize query features and gallery features on 4 target domains by t-SNE, shown in Fig. 5. For baseline, it is obvious that baseline is over-fitting to source domains and lacks the ability to extract identical features from the same identity on the query and the gallery. However, our approach prevents the model from over-fitting in sources and is generalizable to unseen target domains.

5 CONCLUSIONS

In this paper, we revisit the stochastic learning in DG person Re-ID and propose two stochastic behaviors, *i.e.*, stochastic splitting-sliding sampler from view of data distribution and variance-varying gradient dropout from view of optimization process. The proposed sampler keeps various source dominance in two adjacent training episodes and the proposed gradient dropout assists the model to step out of local minimums and search for an optimal solution by constantly varying its gradient variance. Experiments demonstrate that our method can improve the generalization capability on relatively balanced or extremely imbalanced source domains, and outperforms state-of-the-art methods on public DG person Re-ID benchmarks, verifying the effectiveness of our proposal.

ACKNOWLEDGMENTS

This work was supported by grants from National Natural Science Foundation of China (No.62132002 and No.61922006).

REFERENCES

[1] Peixian Chen, Pingyang Dai, Jianzhuang Liu, Feng Zheng, Qi Tian, and Rongrong Ji. 2021. Dual distribution alignment network for generalizable person re-identification. In *Proceedings of AAAI Conference on Artificial Intelligence*, Vol. 6.

[2] Zhao Chen, Jiquan Ngiam, Yanping Huang, Thang Luong, Henrik Kretzschmar, Yuning Chai, and Dragomir Anguelov. 2020. Just pick a sign: Optimizing deep multitask models with gradient sign dropout. *Advances in Neural Information Processing Systems* 33 (2020), 2039–2050.

[3] Seokeon Choi, Taekyung Kim, Minki Jeong, Hyoungseob Park, and Changick Kim. 2021. Meta batch-instance normalization for generalizable person re-identification. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*. 3425–3435.

[4] Yongxing Dai, Xiaotong Li, Jun Liu, Zekun Tong, and Ling-Yu Duan. 2021. Generalizable person re-identification with relevance-aware mixture of experts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16145–16154.

[5] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. 2018. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 994–1003.

[6] Abhimanyu Dubey, Vignesh Ramanathan, Alex Pentland, and Dhruv Mahajan. 2021. Adaptive methods for real-world domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14340–14349.

[7] Alberto Fernández, Salvador García, Mikel Galar, Ronaldo C Prati, Bartosz Krawczyk, and Francisco Herrera. 2018. *Learning from imbalanced data sets*. Vol. 10. Springer.

[8] Douglas Gray and Hai Tao. 2008. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European conference on computer vision*. Springer, 262–275.

[9] Haibo He and Edwardo A Garcia. 2009. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering* 21, 9 (2009), 1263–1284.

[10] Lingxiao He, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng, and Tao Mei. 2020. Fastreid: A pytorch toolbox for general instance re-identification. *arXiv preprint arXiv:2006.02631* (2020).

[11] Alexander Hermans, Lucas Beyer, and Bastian Leibe. 2017. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737* (2017).

[12] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof. 2011. Person re-identification by descriptive and discriminative classification. In *Scandinavian conference on Image analysis*. Springer, 91–102.

[13] Jieru Jia, Qiuqi Ruan, and Timothy M Hospedales. 2019. Frustratingly easy person re-identification: Generalizing person re-id in practice. *arXiv preprint arXiv:1905.03422* (2019).

[14] Xin Jin, Cuiling Lan, Wenjun Zeng, Zhibo Chen, and Li Zhang. 2020. Style normalization and restitution for generalizable person re-identification. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3143–3152.

[15] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. 2017. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*. 5542–5550.

[16] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. 2018. Learning to generalize: Meta-learning for domain generalization. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[17] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. 2018. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5400–5409.

[18] He Li, Mang Ye, and Bo Du. 2021. Weperson: Learning a generalized re-identification model from all-weather virtual data. In *Proceedings of the 29th ACM International Conference on Multimedia*. 3115–3123.

[19] Wei Li and Xiaogang Wang. 2013. Locally aligned feature transforms across views. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3594–3601.

[20] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. 2014. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 152–159.

[21] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. 2018. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 624–639.

[22] Shengcai Liao and Ling Shao. 2020. Interpretable and generalizable person re-identification with query-adaptive convolution and temporal lifting. In *European Conference on Computer Vision*. Springer, 456–474.

[23] Chen Change Loy, Tao Xiang, and Shaogang Gong. 2010. Time-delayed correlation analysis for multi-camera activity understanding. *International Journal of Computer Vision* 90, 1 (2010), 106–129.

[24] Divyat Mahajan, Shruti Tople, and Amit Sharma. 2021. Domain generalization using causal matching. In *International Conference on Machine Learning*. PMLR, 7313–7324.

[25] Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. 2017. Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE international conference on computer vision*. 5715–5725.

[26] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. 2013. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*. PMLR, 10–18.

[27] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. 2018. Two at once: Enhancing learning and generalization capacities via ibn-net. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 464–479.

[28] M Saquib Sarfraz, Arne Schumann, Andreas Eberle, and Rainer Stiefelhagen. 2018. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 420–429.

[29] Jifei Song, Yongxin Yang, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. 2019. Generalizable person re-identification by domain-invariant mapping network. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*. 719–728.

[30] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* 15, 1 (2014), 1929–1958.

[31] Chi Su, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. 2016. Deep attributes driven multi-camera person re-identification. In *European conference on computer vision*. Springer, 475–491.

[32] Yanmin Sun, Andrew KC Wong, and Mohamed S Kamel. 2009. Classification of imbalanced data: A review. *International journal of pattern recognition and artificial intelligence* 23, 04 (2009), 687–719.

[33] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. 2018. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European conference on computer vision (ECCV)*. 480–496.

[34] Masato Tamura and Tomokazu Murakami. 2019. Augmented hard example mining for generalizable person re-identification. *arXiv preprint arXiv:1910.05280* (2019).

[35] Hung-Yu Tseng, Yi-Wen Chen, Yi-Hsuan Tsai, Sifei Liu, Yen-Yu Lin, and Ming-Hsuan Yang. 2020. Regularizing meta-learning via gradient dropout. In *Proceedings of the Asian Conference on Computer Vision*.

[36] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. 2018. Generalizing to unseen domains via adversarial data augmentation. *Advances in neural information processing systems* 31 (2018).

[37] Yanan Wang, Shengcai Liao, and Ling Shao. 2020. Surpassing real-world source training data: Random 3d characters for generalizable person re-identification. In *Proceedings of the 28th ACM international conference on multimedia*. 3422–3430.

[38] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. 2018. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 79–88.

[39] Nicolai Wojke and Alex Bewley. 2018. Deep cosine metric learning for person re-identification. In *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 748–756.

[40] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. 2016. End-to-end deep learning for person search. *arXiv preprint arXiv:1604.01850* 2, 2 (2016), 4.

[41] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. 2017. Joint detection and identification feature learning for person search. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3415–3424.

[42] Zhenlin Xu, Deyi Liu, Junlin Yang, Colin Raffel, and Marc Niethammer. 2020. Robust and generalizable visual representation learning via random convolutions. *arXiv preprint arXiv:2007.13003* (2020).

[43] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. 2021. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

[44] Xinyu Zhang, Jiewei Cao, Chunhua Shen, and Mingyu You. 2019. Self-training with progressive augmentation for unsupervised cross-domain person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8222–8231.

[45] Yuyang Zhao, Zhun Zhong, Fengxiang Yang, Zhiming Luo, Yaojin Lin, Shaozi Li, and Nicu Sebe. 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*. 6277–6286.

[46] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. 2016. Mars: A video benchmark for large-scale person re-identification. In *European conference on computer vision*. Springer, 868–884.

[47] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*. 1116–1124.

[48] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. 2017. Person re-identification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1367–1376.

[49] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. 2009. Associating Groups of People.. In *BMVC*, Vol. 2. 1–11.
[50] Zhedong Zheng, Liang Zheng, and Yi Yang. 2017. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE international conference on computer vision*. 3754–3762.
[51] Zhedong Zheng, Liang Zheng, and Yi Yang. 2017. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE international conference on computer vision*. 3754–3762.
[52] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. 2017. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1318–1327.
[53] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. 2019. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 598–607.
[54] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. 2021. Learning generalisable omni-scale representations for person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).
[55] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. 2020. Learning to generate novel domains for domain generalization. In *European conference on computer vision*. Springer, 561–578.
[56] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. 2021. Domain Generalization with MixStyle. In *ICLR*.

A PROOF OF PROPOSITION 2

Proposition 2. *Let the binary mask $\mathcal{M}$ sampled from Bernoulli distribution with a probability p and the new gradient $g^{'} = \frac{\mathcal{M} \cdot g}{p}$, the expectation $\mathbb{E}$ of $g^{'}$ approaches to be zero and the variance approaches $\frac{1}{p}$ times the variance of g, when the model traps into a local minimum.*

Proof. When the model reaches a local minimum during training, *i.e.*, $g \approx 0$, so the expectation $\mathbb{E}[g] \approx 0$. For the binary mask $\mathcal{M}$ independent of gradients, the expectation of $g^{'}$ is

$$\begin{aligned} \lim_{\mathcal{L}_E \to \mathcal{L}_E^*} \mathbb{E}[g^{'}] &= \lim_{\mathcal{L}_E \to \mathcal{L}_E^*} \frac{\mathbb{E}[\mathcal{M}]\mathbb{E}[g]}{p} \\ &= \lim_{\mathcal{L}_E \to \mathcal{L}_E^*} \frac{p\mathbb{E}[g]}{p} \\ &= 0 \end{aligned} , \tag{8}$$

where $\mathcal{L}_E^*$ represents the local minimum loss.

The variance of $g^{'}$ can be defined as

$$\begin{aligned} \lim_{\mathcal{L}_E \to \mathcal{L}_E^*} Var[g^{'}] &= \lim_{\mathcal{L}_E \to \mathcal{L}_E^*} \mathbb{E}[\frac{\mathcal{M}^2 g^2}{p^2}] - \mathbb{E}^2[\frac{\mathcal{M}g}{p}] \\ &= \lim_{\mathcal{L}_E \to \mathcal{L}_E^*} \frac{\mathbb{E}[\mathcal{M}^2]\mathbb{E}[g^2]}{p^2} \\ &= \lim_{\mathcal{L}_E \to \mathcal{L}_E^*} \frac{p\mathbb{E}[g^2]}{p^2} - 0 \\ &= \lim_{\mathcal{L}_E \to \mathcal{L}_E^*} \frac{\mathbb{E}[g^2] - \mathbb{E}^2[g]}{p} \\ &= \frac{Var[g]}{p} \end{aligned} , \tag{9}$$

B MORE IMPLEMENTATION DETAILS

We use the cross-entropy loss with 0.1 label smoothing. The output feature dimension is 2048 and a batch normalization is inserted between the backbone and the classification. For SGD optimizer, Nesterov is set to true. Moreover, results of the baseline shown in all tables are based on baseline with 10x learning rate in classifier.

C SUPPLEMENTARY EXPLANATIONS AND EXPERIMENTAL ANALYSES

C.1 Supplementary explanations about proposed sampler

The retained number of subsets $L_S + \lfloor \frac{T_S}{2} \rfloor$. Firstly, to prevent all selected subsets only from the largest source domain in two continuous episodes, the number of retained subsets of the largest domain in the tail must be less than $L_S + T_S$. Secondly, considering the number of different subsets between two continuous episodes is T_S, we set the number of retrained subsets to be $L_S + \lfloor \frac{T_S}{2} \rfloor$ to ensure that newly selected subsets in the second episode are mainly from different domains.

The source order sensitivity in the proposed sampler. We conduct experiments on the source order "C2-C3-M-D-CS" and the randomly selected source order "D-CS-C2-M-C3". The performances are only slightly different. On VIPeR dataset, the rank-1 value changes from 61.5 to 61.3, which is 0.2% lower, while on the mAP metric, the random order performance changes from 69.2 to 69.4, which is 0.2% higher. On GRID dataset, the rank-1 value changes from 49.9 to 50.2, which is 0.3% higher, and the mAP values are the same. It reveals that the proposed sampler is not sensitive to the order of source domains.

C.2 Supplementary explanations about proposed gradient dropout

Comparisons between the sliding-window behavior and randomly selected behavior. We randomly select L_G layers in each iteration and then conduct experiments on the C2-C3-M-D-CS setting. From results in Tab. 5, random gradient dropout is effective to improve the performance compared with the baseline, but inferior to the sliding-window behavior. It verifies that applying gradient dropout in continuous layers is better than that in discrete layers.

Performance comparisons in Fig. 3 of the main paper. In the condition of *varying + div.p*, results of p=0.2 and p=0.8 are shown in Tab. 6. From Tab. 6, it reveals that performance comparisons among three conditions usually follow such rules: "varying">"varying+div.p">"regular". Therefore, we only show results of p=0.5 in the paper as the best performance. Moreover, compared with regular dropout (applied in all layers), it demonstrates that the key to the model to escape from local minimums relies on the change of gradient variance rather than the value of gradient variance. Meanwhile, large variance of gradients will hinder the model converging in the late training episodes. In the early training episodes, preventing the model from trapping into local minimums

Table 5: Comparisons between the sliding-window behavior and random selected behavior.

Source	Behavior	PRID		GRID	
		R-1	mAP	R-1	mAP
M+C2+D	Random	70.3	78.2	48.3	57.8
+C3+CS	sliding-window	**74.4**	**81.6**	**49.9**	**59.6**

Table 6: Extra results of the condition *varying + div.p*.

Source	Probability	VIPeR		PRID	
		R-1	mAP	R-1	mAP
M+C2+D	0.2	57.1	65.5	70.3	78.3
+C3+CS	0.8	60.0	68.0	70.9	80.1

Table 7: Comparison with State-of-the-Arts on Protocol-1 and Protocol-2. Protocol-1: use only train split subset on each source domain. Protocol-2: use train and test split subsets on each source domain.

Evaluation Setting	Method	Source	Market-1501		Source	DukeMTMC-reID	
			R-1	mAP		R-1	mAP
Protocol-1	M^3L (ResNet-50) [45]	D+C3+MT	74.5	48.1	M+C3+MT	69.4	50.5
	M^3L (IBN-Net50) [45]		75.9	50.2		69.2	51.1
	Baseline		73.2	46.5		66.7	47.6
	Ours		**78.0**	**51.0**		**71.0**	**52.3**
Protocol-2	RoMoE [4]	D+C3+MT	82.0	56.5	M+C3+MT	73.6	56.9
	Baseline		78.6	53.9		69.1	52.3
	Ours		**82.5**	**58.3**		**73.7**	**57.6**

Table 8: Analyses of sliding window in proposed GradDrop.

Source	Slide.epoch	Win.size	Win.step	mAP	
				PRID	GRID
M+C2+ D+C3+ CS	5	2	1	80.8	57.4
	20			80.5	58.4
	10			**81.6**	**59.6**
		2	2	77.3	58.1
		1	1	81.3	58.9
		3	1	77.4	56.9
		3	3	79.3	57.2

is effective to aid the model to search for a better solution, but in the late training episodes, the model is expected to converge rather than vary.

Analyses of sliding window in proposed gradient dropout We analyse influences on the model in three aspects: 1) sliding epoch; 2) sliding window size; 3) sliding window step. In Tab. 8, it is harmful to the generalization capability of the model if the sliding window moves too fast or too slowly *i.e.*, necessary to wait for the model to approach a local minimum or assist the model in stepping out of over-fitting traps in time. From experimental results, the suitable number of sliding epoch is set to 10. For window size and window step, it can be observed that small window sizes and steps are beneficial for the model to achieve better generalization capability. This means the value of gradient variance should vary in the relatively small interval based on the original variance. From Tab. 8, the best window size and window step are 2 and 1, respectively.

C.3 Evaluations on large-scale datasets

Further to verify the effectiveness of our proposals when evaluating large-scale datasets, we conduct experiments on two settings which are D+C3+MT->M and M+C3+MT->D. For C3+D+MT->M, sources domains are DukeMTMC-reID, CUHK03 and MSMT17 while evaluating on Market-1501. For M+C3+MT->D, sources domains are Market-1501, CUHK03 and MSMT17 while evaluating on DukeMTMC-reID. Fair to compare with state-of-the-arts, we adopt two evaluation settings, *i.e.*, protocol-1 following [45] and protocol-2 following [4]. Two protocols are different in two aspects: (1) For CUHK03 (C3), the detected subset of the old protocol (26,263 images of 1,367 IDs for training) is used in Protocol-1, but in Protocol-2, they adopt the old protocol (14,097 images of 1,467 IDs for training). (2) For training, in Protocol-1, only train split subset is adopted, while they use train and test split subsets on each source domain in Protocol-2. As shown in Tab. 7, our proposal outperforms the state-of-the-arts on both two evaluation settings, verifying the robustness of our approach.