Domain Generalizing DINO for Visual Regression via Latent Distractor Subspace Consistency

Nikhil Reddy[1] Chetan Arora[2] Mahsa Baktashmotlagh[3]
[1]UQ-IIT Delhi Research Academy (UQIDRA) [2]IIT Delhi, India [3]UQ, Australia

Abstract

Vision Foundation Models, such as `DINO` *[20], have demonstrated remarkable generalization in classification; however, their application to out-of-domain visual regression tasks remains a significant and underexplored challenge. Unlike classification, domain generalization in regression poses distinct challenges: regression produces continuous outputs and is particularly sensitive to high-variance, label-irrelevant factors (e.g., illumination, blur, or contrast). These factors can entangle with task-relevant features and induce spurious correlations. While recent regression methods [11, 15, 24, 38, 39] have shown promise, they often rely on CNN backbones and require the pre-specification of known distractors. This demands significant domain expertise and fails to address spurious correlations that emerge during training. To address these challenges, we propose* `LDSC`*, a **L**atent **D**istractor **S**ubspace **C**onsistency framework that disentangles intermediate feature representation into task-relevant and latent distractor subspaces, and regularizes the latter under photometric perturbations to suppress spurious correlations while preserving discriminative features during training. Our proposed method,* `LDSC`*, is the first to effectively adapt the powerful* `DINO` *backbone for domain generalized visual regression.* `LDSC` *achieves state-of-the-art results on seven benchmark regression datasets, demonstrating its strong performance in domain generalization for visual regression with percentage improvements of (41.75%, 20.12%, 52.05%, 8.27%, 22.21%, 3.55%) over state-of-the-art* `DG` *regression methods, respectively. Source code is provided in the supplementary.*

1. Introduction

Vision Foundation Models (`VFMs`), such as `DINO` [20], have shown strong generalization performance in classification settings; their application to out-of-domain visual regression tasks remains a significant and underexplored challenge. Regression tasks introduce unique challenges not present in classification: their continuous label spaces make them sensitive to high-variance, label-irrelevant factors (e.g., illumination, blur). These factors entangle with task-relevant features and induce spurious correlations. Given the prevalence of regression tasks in real-world applications, such as house price prediction [26], age prediction [29], and gaze estimation [1], advancing `DG` in regression is critical to ensure robust generalization of Deep Neural Networks (`DNNs`) in real-world applications.

Existing `DG` methods for regression often adapt techniques developed for classification with discrete labels [2, 3, 12, 16, 27, 41], resulting in suboptimal performance on continuous label spaces. While regression-specific approaches [11, 15, 25, 38, 39] have shown promise, they suffer from three main limitations: (i) they require pre-specification of distractors, demanding substantial domain expertise; (ii) they are not well suited to `VFM` backbones and frequently rely on CNN architectures; and (iii) they fail to address spurious correlations that emerge during training, where irrelevant features such as background clutter become entangled with task-relevant representations. If left unmitigated, these correlations can bias predictions and severely hinder generalization to unseen domains.

Conditional independence methods [24, 28] also focus on prior distractor specification (e.g., protected attributes) and kernelized expectations, but this requirement limits their applicability. Such methods fail to capture unforeseen spurious factors, such as style variations in text or background clutter in images. As a result, both regression-specific techniques [11, 15, 25, 38, 39] and distractor-driven approaches [24, 28] leave spurious correlations largely unmitigated.

To address these limitations, we propose `LDSC`, a **L**atent **D**istractor **S**ubspace **C**onsistency framework that explicitly disentangles task-relevant signals from latent distractors in the representation space. Our key insight is that these distractors manifest as high-variance, label-irrelevant directions in the feature space that capture style-related variations (e.g., illumination, background texture). If left unregularized, these directions can induce spurious correlations.

The key contributions of our work include:

- **Implicit Feature Label Decoupling:** We design a frame-

Problem Setting	Source Data	Target Data	Train Loss	Continuous Target	Open Targets	Data Imbalance
UDA Classification[23]	x^s, y^s	x^t	$\mathcal{L}(x^s, y^s) + \mathcal{L}(x^t, x^s)$	✗	✗	✗
DG Classification[3, 34, 42]	x^s, y^s	✗	$\mathcal{L}(x^s, y^s)$	✗	✓	✗
UDA Regression[19, 35]	x^s, y^s	x^t	$\mathcal{L}(x^s, y^s) + \mathcal{L}(x^t, x^s)$	✓	✗	✗
Deep Imbalanced Regression[38]	x^s, y^s	✗	$\mathcal{L}(x^s, y^s)$	✓	✗	✓
DG Regression[39]	x^s, y^s	✗	$\mathcal{L}(x^s, y^s)$	✓	✓	✗

Table 1. Characteristics of existing paradigms, LDSC fits in *Domain Generalization in Regression* setting, which aims to generalize model trained on source data, to perform well on the unseen target data for regression tasks. UDA, DG refers to Unsupervised Domain Adaptation, Domain Generalization.

work to identify latent distractors prone to inducing spurious correlations. By dynamically identifying and suppressing these spurious factors, our method prevents the model from forming false correlations that degrade performance on unseen domains.

- **Orthogonal Task-Relevant and Latent Distractor Subspaces:** Our method learns two orthogonal subspaces of the feature representations, a task-relevant subspace aligned with the continuous target label, and a latent distractor subspace that captures high-variance, label-irrelevant variability. We employ the Hilbert–Schmidt Independence Criterion (HSIC) to enforce statistical independence between the distractor subspace and the labels, ensuring that latent features do not inadvertently influence predictions.
- **Latent Distractor Subspace Consistency:** We propose LDSC, a novel regularization method that enforces the latent distractor subspace to be consistent under photometric perturbations (e.g., color jitter). By penalizing only the covariance components of the latent features that vary with style changes, LDSC ensures that the model focuses on invariant content rather than style-dependent noise. This effectively suppresses spurious style-based correlations, improving robustness across domains.
- **Extensive Evaluation and Improved Fairness:** We are the first to successfully adapt the powerful DINO Vision Foundation Model for the task of domain generalized visual regression. We provide extensive empirical experiments of our LDSC framework on seven synthetic and real-world datasets, achieving percentage improvements of (41.75%, 20.12%, 52.05%, 8.27%, 22.21%, 3.55%) over state-of-the-art DG regression methods. Beyond achieving better generalization, we further show that LDSC can improve the fairness compared to existing SOTA DG methods in regression.

2. Related work

Invariant representations. In Kernel-based representation learning methods, independence of variable X from Z given Y can be enforced by leveraging the Hilbert-Schmidt Conditional Independence Criterion[21](HSIC) which facilitates the learning of generalized counterfactually invariant representations. The Generalized Covariance Measure[28](GCM) normalizes the covariance between residuals from kernel-ridge regressions of X on Y and Z on Y, making it particularly suitable for multivariate settings. Conditional Independence Regression Covariance[24](CIRCE) provides a measure of conditional independence that can be applied as a regularizer to enforce conditional independence between intermediate feature representations and pre-specified distractors. However, CIRCE demands specialized domain knowledge to explicitly specify distractors and does not account for spurious correlations that emerge during training. To address these limitations, our proposed LDSC framework automatically identifies and suppresses latent distractors, eliminating the need for predefined distractors and enhancing generalization performance across unseen domains.

Domain Generalization in Regression. Existing DG methods mostly focus on classification [34, 42], limiting their applicability to regression tasks. IRM [3] shows promise for DG in classification, yet its application on regression tasks remains largely unexplored. Domain adaptation methods like MMD [31] and DANN [8] align cross-domain distributions but have not been explored for DG in regression. Recent work C-Mixup [39] proposes a label-distance-based mixup technique to augment data for DG in regression. Existing DG methods fail to decorrelate irrelevant input variables from the learned feature representations, limiting its generalization performance on unseen domains.

Deep Imbalanced Regression. DIR approaches, such as RankSim [11], FDS [38], BalancedMSE [25], and ConR [15], address skewed label distributions by ensuring better representation of underrepresented labels. Unlike DG, which targets robustness to unseen distributions, DIR focuses on mitigating data imbalance within the training data.

3. Methodology

In this section, we introduce our proposed LDSC framework, which centers on the concept of **L**atent **D**istractor **S**ubspace **C**onsistency. The key intuition of our approach is that high-variance, label-irrelevant directions in the repre-

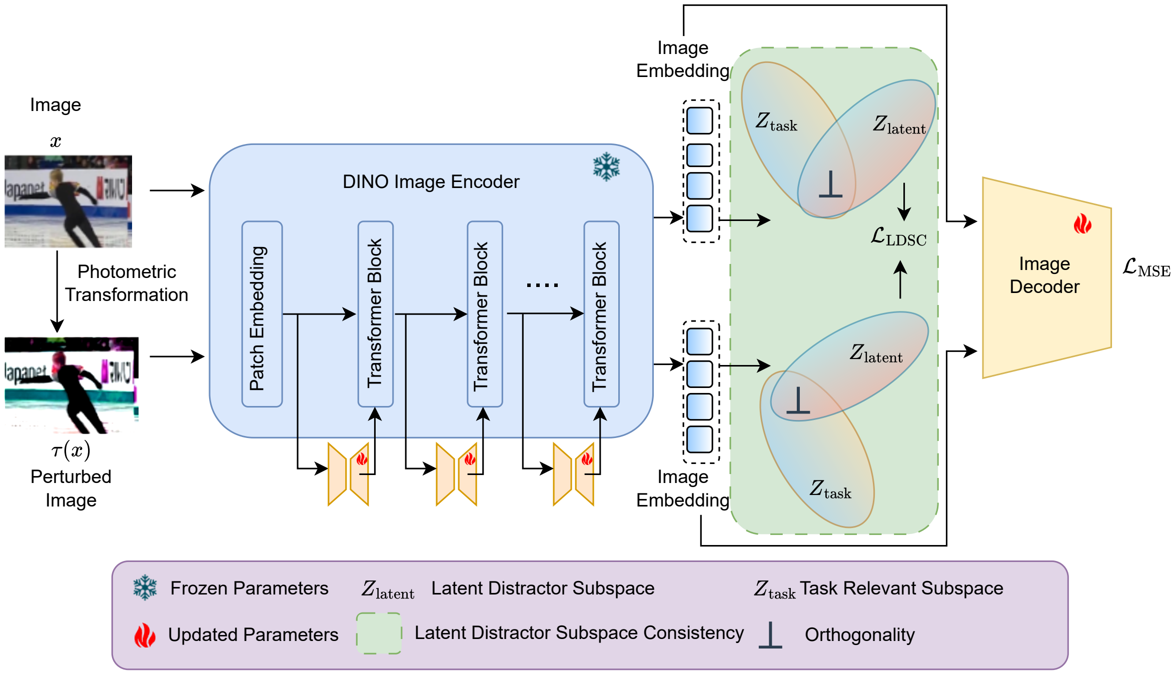

Figure 1. Architecture of the proposed Latent Distractor Subspace Consistency (LDSC) framework for domain-generalized visual regression tasks. An input image and its photometrically perturbed twin (e.g., via color jitter) are fed into a shared DINO vision-transformer encoder (with multiple transformer blocks) to produce high-dimensional feature embeddings. The encoder outputs are decomposed via orthonormal projections into two orthogonal latent subspaces: a task-relevant subspace aligned with the regression target, and a latent distractor subspace capturing high-variance style factors (e.g., illumination or blur) that are explicitly constrained to be orthogonal to the task features. The task-relevant features are passed through an image decoder to predict the continuous output, while the latent distractor subspace is regularized by enforcing consistency under photometric perturbations (penalizing style-sensitive covariance components).

sentation space often capture style-related factors (e.g., illumination, blur, or contrast) rather than task-relevant content. If left unregularized, these irrelevant factors can induce spurious correlations that degrade performance on unseen domains. Our goal is therefore to explicitly identify and regularize such latent distractors, ensuring that they remain stable under photometric perturbations while being orthogonal to task-relevant features.

Problem Definition. We consider the domain generalization (DG) setting for regression. The input and label spaces are denoted by $\mathcal{X}$ and $\mathcal{Y}$, respectively. The label space $\mathcal{Y}$ has a continuous range and contains: $\mathcal{Y}_{\text{source}}$ and $\mathcal{Y}_{\text{target}}$, with no overlap between them. The source and target domain data are represented by $D_s = \{(\mathbf{x}, \mathbf{y}) \in \{\mathcal{X} \times \mathcal{Y}_{\text{source}}\}\}$ and $D_t = \{(\mathbf{x}, \mathbf{y}) \in \{\mathcal{X} \times \mathcal{Y}_{\text{target}}\}\}$, respectively. The model is trained only on D_s and then used to predict on D_t without any further adaptation.

Notations. are represented below:

- $X \in \mathbb{R}^{n \times p}$: batch of n images, where $p = H \times W \times C$ for height H, width W, and C channels. The i-th row corresponds to $x_i^\top$.
- $Y \in \mathbb{R}^{n \times 1}$: continuous target labels; the i-th entry is y_i.
- $\phi(\cdot) : \mathbb{R}^p \to \mathbb{R}^k$: feature extractor (image encoder outputs). The stacked features are $\Phi = \begin{bmatrix} \phi(x_1)^\top \\ \vdots \\ \phi(x_n)^\top \end{bmatrix} \in \mathbb{R}^{n \times k}$.
- $f(\cdot) : \mathbb{R}^p \to \mathbb{R}$: prediction function (regressor). For input x_i, the model prediction is $f(x_i)$.
- $d_1, d_2 \in \mathbb{N}$: target dimensions for the task-relevant and latent-distractor subspaces, with $d_1 + d_2 \leq k$.
- $U_1 \in \mathbb{R}^{k \times d_1}$, $U_2 \in \mathbb{R}^{k \times d_2}$: column-orthonormal projection matrices satisfying $U_1^\top U_1 = I_{d_1}$, $U_2^\top U_2 = I_{d_2}$, $U_1^\top U_2 = \mathbf{0}$.
- $Z_{\text{task}} = \Phi U_1 \in \mathbb{R}^{n \times d_1}$: task-relevant subspace (aligned with Y).
- $Z_{\text{latent}} = \Phi U_2 \in \mathbb{R}^{n \times d_2}$: latent distractor subspace (high-variance, orthogonal to Z_{task} and minimally aligned with Y).

3.1. Implicit Feature-Label Decoupling

In many real-world datasets, explicit distractors are either unavailable or require specialized domain knowledge to define, while spurious correlations often arise from hidden irrelevant factors during training. To mitigate this, we propose an implicit feature–label decoupling strategy that disentangles task-relevant features from latent distractors in the representation space.

Theoretical Framework. Consider the data matrix $X \in \mathbb{R}^{n \times p}$ with n samples and p features. Let $\phi(x) \in \mathbb{R}^k$ denote the intermediate feature representations (encoder outputs) of an input x. Stacking n samples yields the feature matrix $\Phi \in \mathbb{R}^{n \times k}$. Our objective is to learn two orthogonal subspaces: a task-relevant subspace $Z_{\text{task}} \in \mathbb{R}^{n \times d_1}$ aligned with the target label Y, and a latent distractor subspace $Z_{\text{latent}} \in \mathbb{R}^{n \times d_2}$ capturing high-variance directions orthogonal to Z_{task} and minimally aligned with target Y. These are defined as:

$$Z_{\text{task}} = \Phi U_1, \quad Z_{\text{latent}} = \Phi U_2, \tag{1}$$

where $U_1 \in \mathbb{R}^{k \times d_1}$ and $U_2 \in \mathbb{R}^{k \times d_2}$ are orthonormal projection matrices with $U_1^\top U_1 = I$, $U_2^\top U_2 = I$, and $U_1^\top U_2 = 0$.

Statistical Independence via HSIC. To measure dependence, we employ the Hilbert–Schmidt Independence Criterion (HSIC) [21]. For random variables A and B, the empirical HSIC is defined as:

$$\text{HSIC}(A, B) \approx \frac{1}{n^2} \operatorname{tr}\big(K_A H K_B H\big), \tag{2}$$

where K_A and K_B are Gram matrices with $K_A(i,j) = k_A(a_i, a_j)$ and $K_B(i,j) = k_B(b_i, b_j)$, and $H = I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ is the centering matrix and I_n is the identity matrix of size n, and $\mathbf{1}$ is the n-dimensional vector whose entries are all ones.

Task-Relevant Subspace. For the task-relevant subspace Z_{task}, we maximize its dependence with Y using Hilbert–Schmidt Independence Criterion (HSIC):

$$\begin{aligned} U_1 &= \arg \max_{U,\, U^\top U = I} \text{HSIC}(\Phi U, Y) \\ &= \arg \max_{U} \operatorname{tr}\Big(U^\top \Phi^\top H K_Y H \Phi U\Big), \end{aligned} \tag{3}$$

where $K_Y(i,j) = \exp(-\|y_i - y_j\|^2 / 2\sigma^2)$ is a Gaussian kernel. This ensures Z_{task} captures discriminative features predictive of Y.

Latent Distractor Subspace. For the latent distractor subspace Z_{latent}, we maximize the captured variance in X while minimizing its dependence with target Y:

$$U_2 = \arg \max_{\substack{U^\top U = I \\ U^\top U_1 = 0}} \operatorname{tr}(U^\top \Phi^\top \Phi U) - \mu\, \text{HSIC}(\Phi U, Y). \tag{4}$$

where $\mu > 0$ is a HSIC regularization strength parameter balancing variance maximization and independence. Substituting the HSIC estimator yields

$$\begin{aligned} U_2 = \arg \max_{\substack{U^\top U = I \\ U^\top U_1 = 0}} & \operatorname{tr}\big(U^\top \Phi^\top \Phi U\big) \\ & - \mu \operatorname{tr}\big(U^\top \Phi^\top H K_Y H \Phi U\big). \end{aligned} \tag{5}$$

This formulation ensures Z_{latent} encapsulates high-variance directions orthogonal to Z_{task} and is minimally aligned with target Y.

3.2. Latent Distractor Subspace Consistency (LDSC)

To improve the robustness of intermediate representations, we introduce Latent Distractor Subspace Consistency (LDSC), a regularization objective that enforces stability of the latent distractor subspace $Z_{\text{latent}} = \Phi U_2$ under photometric perturbations (e.g., color jitter). Since this subspace captures high-variance, label-irrelevant factors such as illumination or blur, we regularize it to suppress spurious style-sensitive components. Following RobustNet [5], which shows that feature covariances reveal style versus content, we enforce consistency of Z_{latent} under photometric perturbations, thereby filtering out style-variant covariances while preserving invariant, task-relevant structure.

Step 1: Instance standardization. For an image x and its photometric twin $\tau(x)$, compute standardized latent features

$$\begin{aligned} Z^{\text{s}}_{\text{latent}}(x) = \big(\operatorname{diag}(\Sigma_Z(x))\big)^{-\frac{1}{2}} \odot \\ \big(Z_{\text{latent}}(x) - \mu_Z(x)\big), \\ Z^{\text{s}}_{\text{latent}}(\tau(x)) = \big(\operatorname{diag}(\Sigma_Z(\tau(x)))\big)^{-\frac{1}{2}} \odot \\ \big(Z_{\text{latent}}(\tau(x)) - \mu_Z(\tau(x))\big), \end{aligned} \tag{6}$$

where $\mu_Z(\cdot)$ and $\Sigma_Z(\cdot)$ denote the per-instance mean and covariance.

Step 2: Sensitivity estimation. From the covariances of the standardized representations, $\Sigma(Z^{\text{s}}_{\text{latent}}(x))$ and $\Sigma(Z^{\text{s}}_{\text{latent}}(\tau(x)))$, compute a variance matrix

$$V^Z = \tfrac{1}{2}\Big(\big(\Sigma(Z^{\text{s}}_{\text{latent}}(x)) - \mu^Z_\Sigma\big)^2 + \big(\Sigma(Z^{\text{s}}_{\text{latent}}(\tau(x))) - \mu^Z_\Sigma\big)^2\Big), \tag{7}$$

$$\mu^Z_\Sigma = \tfrac{1}{2}\big(\Sigma(Z^{\text{s}}_{\text{latent}}(x)) + \Sigma(Z^{\text{s}}_{\text{latent}}(\tau(x)))\big), \tag{8}$$

High-variance entries in V^Z identify style-sensitive covariances; we obtain a binary mask $\widetilde{M}^Z$ via clustering, following [5].

Step 3: LDSC loss. We penalize only the masked, style-sensitive entries:

$$\mathcal{L}_{\text{LDSC}} = \tfrac{1}{2}\left\|\Sigma^{\text{s}}_Z(x) \odot \widetilde{M}^Z\right\|_1 + \tfrac{1}{2}\left\|\Sigma^{\text{s}}_Z(\tau(x)) \odot \widetilde{M}^Z\right\|_1. \tag{9}$$

Our approach applies the consistency objective specifically to the latent distractor subspace. By attenuating covariance components that fluctuate under photometric perturbations, LDSC encourages Z_{latent} to eliminate spurious style-sensitive information, thereby enhancing robust generalization.

Total objective. The final training loss is

$$\mathcal{L}_{\text{total}} = \underbrace{\frac{1}{n}\sum_{i} \|f(\mathbf{x}_i) - y_i\|_2^2}_{\mathcal{L}_{\text{MSE}}} + \lambda_{\text{LDSC}}\,\mathcal{L}_{\text{LDSC}}, \qquad (10)$$

where λ_{LDSC} regulates the trade-off between accurate prediction and consistency in the latent distractor subspace.

4. Experiments

Implementation Details. We implement our approach in PyTorch [22] and initialize model parameters with standard Xavier initialization. We use the Adam optimizer with a learning rate of 1×10^{-3}. We conduct experiments on an NVIDIA H100 80GB GPU. We use λ_{LDSC}, as 1×10^{-1}. We choose μ, d_1, d_2 as 0.4, 5 , 5 respectively. We employ Sis-PCA [30] to identify and extract the orthogonal subspaces referred to in Equation (1). We choose the number of iterations as 250, with early stopping criteria applied when the loss difference between two consecutive iterations falls below 1×10^{-5}. We use color jitter as the photometric transformation with the default setting in Pytorch. Each run on average requires ∼5.8 hours of training for In-distribution datasets and ∼4.2 hours for Out-of-Distribution datasets. The remaining hyperparameters (e.g., kernel bandwidth σ) and training details are reported in the supplementary material. We use DINOv2-L as the image encoder with a LoRA rank of 8, and use ViT-Base as the image decoder, which outputs regression predictions. For evaluation, we benchmark against existing DG, DIR methods (e.g., RankSim, FDS) and obtain state-of-the-art performance on the task setting outlined in Tab. 1.

Evaluation criteria. We use the average Root Mean Square Error (average RMSE) and worst Root Mean Square Error (worst RMSE) as evaluation metrics. We perform ten random runs with random seeds and report each dataset's average RMSE and worst RMSE.

4.1. In-Distribution Generalization results

In-Distribution datasets. We consider three In-Distribution datasets: BIWI Kinect [6, 17], Rendered Hand Pose (RHD) [43], and dSprites [18].

BIWI Kinect [7, 17]. an RGB-D dataset of ∼15K frames from 20 subjects (4 recorded twice) captured with a Microsoft Kinect (640×480). Each frame provides depth and RGB images with ground-truth 3D head location and rotation (yaw/pitch/roll), spanning wide pose ranges (about ±75° yaw, ±60° pitch, ±50° roll). The prediction is the head center's position, which is used for the performance analysis. We report the average of M→F and F→M evaluations.

Rendered Hand Pose (RHD) [43]. is a fully synthetic RGB dataset for hand pose estimation, containing 41,258 training and 2,728 test images at 320×320 resolution. It provides both 2D and 3D annotations for 21 hand keypoints, along with segmentation masks and depth maps. The dataset includes a wide variety of viewpoints, backgrounds, lighting conditions, and self-occlusions.

dSprites [10, 18]. is a synthetic dataset of 2D shapes, organized into three domains: Color (C), Noise (N), and Scream (S). Each domain enumerates the full combinatorial grid, resulting in 737,280 images at 64 × 64 resolution. The ground-truth generative factors include shape, scale, rotation, and (x, y) positions. In our study, we formulate regression tasks on $\{\text{scale}, x, y\}$. Performance is reported as the average over six domain transfer settings: C→N, C→S, N→C, N→S, S→C, and S→N.

Performance on In-Distribution Datasets. We compare the performance of our LDSC framework on the In-distribution datasets, mentioned in the Sec. 4.1. We compare with prior existing methods namely Manifold Mixup [33], Local Mixup [4], k-Mixup [12], Mixup [41], and MixRL [14] which were originally proposed for the classification task. For a fair comparison, we adapt DINOv2-L + ViT-Base with LoRA rank as 8 and report its performance without our proposed method (i.e., with only MSE loss). Prior methods Mixup [39], RankSim [11], and FDS [38] are proposed for regression tasks. We report the average RMSE and worst RMSE by considering ten random seeds. Results are shown in Tab. 2. The proposed LDSC framework significantly increases the performance of current SOTA C-Mixup [39] by 20.12% on the Rendered Hand Pose dataset. We demonstrate that our proposed LDSC framework can demonstrate SOTA performance compared to ERM and C-Mixup methods across different datasets.

4.2. Out-of-Distribution Generalization results

To evaluate the out-of-distribution (OOD) generalization ability of the existing methods, we perform experiments on three different datasets. These include three datasets, namely PovertyMap [40], one synthetic dataset, RCF-MNIST [39], and another dataset containing both synthetic and real images, MPI3D [9].

RCF-MNIST [39]. is a dataset with 60,000 images built on FashionMNIST [36] and includes spurious correlations between colors and rotation angles.

Poverty Map dataset [40]. is an image regression dataset that contains satellite images used to estimate wealth in countries not present in the training data.

MPI3D[9]. is a benchmark image dataset of 1,036,800 images with three distributions to predict intrinsic factors. Our experiments only focus on predicting the rotation around a vertical and horizontal axis.

Performance on Out-of-Distribution Datasets. We compare the performance of the LDSC framework on the OOD

	BIWI Kinect		Rendered Hand Pose (RHD)		dSprites	
Method	Avg RMSE	Worst RMSE	Avg RMSE	Worst RMSE	Avg RMSE	Worst RMSE
Manifold Mixup [33]	0.532	0.565	7.422	7.645	0.652	0.687
k-Mixup [12]	0.848	0.917	6.583	6.872	0.465	0.512
Local Mixup [4]	0.612	0.655	7.842	8.326	0.782	0.839
MixRL [14]	0.557	0.612	6.187	6.463	0.583	0.624
Mixup [41]	0.667	0.719	5.783	6.184	0.698	0.724
RankSim [11]	0.450	0.598	5.237	5.511	0.554	0.589
FDS [38]	0.378	0.418	4.532	4.987	0.665	0.778
ERM	0.487	0.558	5.671	6.239	0.768	0.872
C-Mixup [39]	0.388	0.418	3.659	3.962	0.559	0.652
DINOv2-L + ViT-Base [20]	0.388	0.416	4.238	4.576	0.489	0.532
LDSC (ours)	**0.226**	**0.248**	**2.923**	**3.451**	**0.263**	**0.314**

Table 2. Performance comparison of existing methods with LDSC framework for in-distribution datasets BIWI Kinect, Rendered Hand Pose (RHD), and dSprites (see Sec. 4.1). We report the average RMSE and worst RMSE of ten runs with random seeds. Results of our proposed LDSC framework are compared with SOTA DG regression methods.

	RCF-MNIST		PovertyMap	
	Avg RMSE	Worst RMSE	Avg RMSE	Worst RMSE
IRM [3]	0.154	0.172	1.115	1.329
IB-IRM [2]	0.178	0.195	1.276	1.563
CORAL [16]	0.169	0.184	1.874	1.932
GroupDRO [27]	0.182	0.207	1.563	1.788
Mixup [41]	0.177	0.192	1.687	1.982
RankSim [11]	0.192	0.208	1.334	1.487
FDS [38]	0.187	0.211	1.221	1.335
ERM	0.155	0.173	1.348	1.562
C-Mixup [39]	0.133	0.146	0.995	1.113
DINOv2-L + ViT-Base [20]	0.145	0.158	0.815	0.838
LDSC (ours)	**0.122**	**0.134**	**0.774**	**0.789**

Table 3. Performance comparison of existing methods with LDSC framework for out-of-distribution datasets RCF-MNIST and PovertyMap (see Sec. 4.2). We report the average RMSE and worst RMSE of ten different runs with random seeds. Results of our proposed LDSC framework are compared with SOTA DG regression methods.

	rc	rl	t	Average
ERM	$0.08132 \pm 9.6e^{-6}$	$0.09819 \pm 6.2e^{-5}$	$0.007004 \pm 5.4e^{-9}$	0.06217
C-Mixup	$0.07112 \pm 2.1e^{-6}$	$0.08344 \pm 4.3e^{-4}$	$0.006489 \pm 5.4e^{-3}$	0.05311
DINOv2-L + ViT-Base [20]	$0.09226 \pm 4.2e^{-5}$	$0.10495 \pm 1.8e^{-4}$	$0.014453 \pm 5.9e^{-8}$	0.07055
LDSC (ours)	$\mathbf{0.07812 \pm 2.6e^{-5}}$	$\mathbf{0.04844 \pm 0.1e^{-4}}$	$\mathbf{0.001072 \pm 2.1e^{-8}}$	**0.05122**

Table 4. Performance comparison of existing methods with LDSC framework on the MPI3D [9] dataset. We report the Average RMSE performance for ten different runs. Results of our proposed LDSC framework are compared with ERM and C-Mixup methods.

datasets reported in the Sec. 4.2. We compare with prior existing methods such as ERM, IRM [3], IB-IRM [2], CORAL [16], GroupDRO [27], and Mixup [41]. For a fair comparison, we adapt DINOv2-L + ViT-Base with LoRA rank as 8 and report its performance without our proposed method (i.e., with only MSE loss).

We demonstrate the performance comparison of LDSC framework with existing prior methods. Results are shown in Tab. 3, Tab. 4 and Tab. 5 respectively. We report that proposed LDSC framework improves the performance of IB-IRM [2] by 31.46% on the RCF-MNIST dataset. Our approach consistently improves across both In-distribution and OOD datasets. Additional information regarding hyper-parameters can be found in the supplementary material.

Performance on MPI3D dataset. We used the MPI3D domain generalization dataset settings used in [9] and com-

	rc	rl	t	Average
ERM	$0.3163 \pm 3.3 \times 10^{-5}$	$0.3511 \pm 3.2 \times 10^{-4}$	$0.0922 \pm 6.7 \times 10^{-7}$	0.2532
C-Mixup [39]	$0.3011 \pm 1.4 \times 10^{-6}$	$0.2489 \pm 1.8 \times 10^{-5}$	$0.0644 \pm 1.1 \times 10^{-6}$	0.2381
DINOv2-L + ViT-Base [20]	$0.3328 \pm 1.6 \times 10^{-4}$	$0.3648 \pm 5.5 \times 10^{-4}$	$0.1287 \pm 5.1 \times 10^{-6}$	0.2772
LDSC (ours)	$\mathbf{0.2911 \pm 4.3 \times 10^{-3}}$	$\mathbf{0.2843 \pm 2.1 \times 10^{-3}}$	$\mathbf{0.1022 \pm 2.8 \times 10^{-5}}$	**0.2243**

Table 5. Performance comparison of existing methods with our LDSC approach on the MPI3D [9] dataset. We report the average Mean Average Error (MAE) performance for ten different runs. Results of our proposed LDSC framework are compared with ERM and C-Mixup methods.

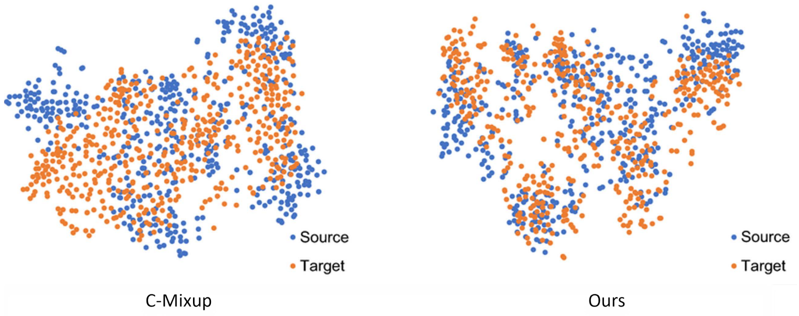

Figure 2. t-SNE [32] Visualization of the source and target feature representations on RCF-MNIST dataset.

pared our method on three generalization tasks: rl, rc → t; t, rc → rl; and rl, t → rc. To choose the best model, we used the test subsets of source distributions as validation sets and ran all experiments ten times with different random seeds, following the approach for random seed and hyper-parameter seed selection. For a fair comparison, we adapt DINOv2-L + ViT-Base with LoRA rank as 8 and report its performance without our proposed method (i.e., with only MSE loss). To measure the effectiveness of our domain generalization LDSC framework on the MPI3D dataset, we use the evaluation metrics Average Mean Square Error (Average MSE) and Average Mean Absolute Error (Average MAE). Performance comparison of existing methods with LDSC framework are reported in Tab. 4 and Tab. 5 respectively. Among the existing methods, only C-Mixup has shown results on the MPI3D dataset. Therefore, we only compare our approach with C-Mixup.

4.3. Additional Results: Improvement on Fairness

To demonstrate the effectiveness of the LDSC framework in improving fairness in regression tasks, we consider two datasets, namely, the PovertyMap dataset [40] and figure skating FisV dataset [37]. PovertyMap dataset is an image dataset, further details are specified in Sec. 4.2. FisV dataset consists of 500 figure skating videos, with an average length of 2 minutes and 50 seconds. The regression task assigns a target score based on the athlete's performance in a figure skating video. The ground truth score is based on human marking. Performance comparison of existing methods with LDSC framework is reported in Tab. 7. We use the ratio of separation ($\hat{r}_{sep}$) to measure the fairness of the regression model as defined in the [13]. The closer $\hat{r}_{sep}$ is to 1, the better the regression model's fairness. We consider the model described in LSTM model[37] and train it with and without our LDSC framework. Results are shown in Tab. 7. Results demonstrate that LDSC consistently improves the fairness ($\hat{r}_{sep}$ closer to 1) without compromising on Average RMSE.

5. Ablation Study

Ablation Study on Loss Components. We analyze the impact of HSIC-based subspace separation and latent distractor subspace consistency loss on the average RMSE performance. For this experiment, we consider the PovertyMap dataset and results as average RMSE. Results are reported in Tab. 6.

Visualising the Feature Representations. To illustrate the impact of our proposed LDSC framework on the embedding space in regression tasks, we visualize the distribution of source and target feature representations with and without LDSC framework. More specifically, we consider the output of the encoder of the DINO encoder on the RCF-MNIST dataset. We consider the LDSC and plot the source and target feature representations. Visualization is shown in Fig. 2. Our proposed LDSC approach is aimed at minimizing both MSE loss and decoupling-based regularization loss

Method	Average RMSE	Worst RMSE
C-Mixup [39]	0.995	1.113
LDSC (ours)	**0.774**	**0.789**
w/o HSIC-based subspace separation	0.889	0.997
w/o LDSC loss (no subspace consistency)	0.991	1.004

Table 6. Ablation study of our LDSC framework on the PovertyMap dataset. We analyze the impact of (i) HSIC-based subspace separation for disentangling task-relevant and distractor features, and (ii) the latent distractor subspace consistency (LDSC) loss that enforces stability under perturbations. Removing either component degrades both the average RMSE and worst RMSE, highlighting their complementary contributions.

Dataset	Method	Average RMSE	$\hat{r}_{\text{sep}}$
PovertyMap dataset	C-Mixup	0.995	1.544
	LDSC (ours)	0.774	1.241
FisV dataset	C-Mixup	20.322	1.849
	LDSC (ours)	18.781	1.344

Table 7. Fairness metric comparison of existing DG methods with our LDSC approach on the PovertyMap dataset and FisV dataset. We report the Average Root Mean Square Error (Average RMSE) and $\hat{r}_{\text{sep}}$ performance for ten different runs. Results of LDSC are compared with C-Mixup method.

Backbone(s)	Trainable Params (M)	Avg. RMSE
DINOv2-S/14 + ViT-B/16	86.6	2.210
DINOv2-L/14 + ViT-S/16	304.7	2.196
DINOv2-G/14 + ViT-L/16	1017.0	2.194
DINOv2-B/14 + ViT-B/16	173.2	1.205
DINOv2-L/14 + ViT-B/16	391.3	**0.774**
DINOv2-L/14 + ViT-L/16	609.4	1.934
DINOv3-L/14 + ViT-L/16	1280.1	2.441

Table 8. Ablation on DINO and ViT combinations of LDSC on PovertyMap dataset. We report the average RMSE over ten different runs.

to decorrelate feature representations from distractors. Results demonstrate that our approach is more aligned with the source and target feature distribution pattern.

Ablation Study on Model Backbones. We analyze the impact of combining DINO backbones with standard ViT-Base models on the PovertyMap dataset. While DINO variants provide strong pretrained representations, their performance can be complementary when combined with lighter ViT backbones (e.g., feature fusion or ensemble). Tab. 8 compares different DINO+ViT combinations. Our preliminary experiments with DINOv3 did not yield comparable improvements, and we plan to further investigate the underlying reasons in a future study for its limited performance.

Ablation Study on λ_{LDSC}, which controls the impact of the latent distractor subspace consistency loss, is reported in the supplementary material.

Ablation Study on μ, which controls HSIC regularization strength, is reported in the supplementary material.

Ablation Study on d_1 d_2 representing the dimensions of the orthogonal subspaces referred in Sec. 3, is provided in the supplementary material.

Ablation Study on LoRA rank. is reported in the supplementary material.

Ablation Study on learning rate. is reported in the supplementary material.

6. Conclusion

In this paper, we introduced LDSC, an implicit feature-label decoupling for Domain Generalization (DG) in regression. The method learns to disentangle representations into two orthogonal subspaces: a task-relevant subspace aligned with the continuous target label and a latent distractor subspace that captures high variance, a label-irrelevant subspace minimally aligned with the target. Second, we introduced a novel regularization method, Latent Distractor Subspace Consistency (LDSC), to enhance robustness. LDSC enforces the latent distractor subspace to be consistent under photometric perturbations like color jitter. By penalizing covariance components that vary with style, it enforces the model to focus on invariant content, effectively suppressing spurious style-based cues that often hinder generalization performance. As the first work to successfully adapt the DINO Vision Foundation Model for domain-generalized visual regression, we demonstrate that LDSC sets a new state-of-the-art, outperforming existing methods in both ID and OOD settings.

Acknowledgment:. We thank UQ-IIT Delhi Research Academy (UQIDRA) for providing a contingency grant.

References

[1] Ahmed A Abdelrahman, Thorsten Hempel, Aly Khalifa, Ayoub Al-Hamadi, and Laslo Dinges. L2cs-net: Fine-grained gaze estimation in unconstrained environments. In *2023 8th International Conference on Frontiers of Signal Processing (ICFSP)*, pages 98–102. IEEE, 2023. 1

[2] Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Jean-Christophe Gagnon-Audet, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. Invariance principle meets information bottleneck for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 34: 3438–3450, 2021. 1, 6

[3] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2020. 1, 2, 6

[4] Raphael Baena, Lucas Drumetz, and Vincent Gripon. Preventing manifold intrusion with locality: Local mixup. *arXiv preprint arXiv:2201.04368*, 2022. 5, 6

[5] Sungha Choi, Sanghun Jung, Huiwon Yun, Joanne T Kim, Seungryong Kim, and Jaegul Choo. Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11580–11590, 2021. 4

[6] Computer Vision Lab, ETH Zurich. Biwi kinect head pose database. `https://vision.ee.ethz.ch/datsets.html`. Dataset page, accessed 2025-09-10. 5

[7] Gabriele Fanelli, Matthias Dantone, Juergen Gall, Andrea Fossati, and Luc Van Gool. Random forests for real time 3d face analysis. *International Journal of Computer Vision*, 101(3):437–458, 2013. 5

[8] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016. 2

[9] Muhammad Waleed Gondal, Manuel Wuthrich, Djordje Miladinovic, Francesco Locatello, Martin Breidt, Valentin Volchkov, Joel Akpo, Olivier Bachem, Bernhard Schölkopf, and Stefan Bauer. On the transfer of inductive bias from simulation to the real world: a new disentanglement dataset. *Advances in Neural Information Processing Systems*, 32, 2019. 5, 6, 7

[10] Muhammad W. Gondal, Manuel Wüthrich, Djordje Miladinovic, Francesco Locatello, Michael Tschannen, Valentin Volchkov, Jonathan Akpo, Olivier Bachem, Bernhard Schölkopf, and Stefan Bauer. On the transfer of inductive bias from simulation to the real world: A new disentanglement dataset. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. See Table/Section detailing Color-, Noisy-, and Scream-dSprites; also available as arXiv:1906.03292. 5

[11] Yu Gong, Greg Mori, and Frederick Tung. Ranksim: Ranking similarity regularization for deep imbalanced regression. *arXiv preprint arXiv:2205.15236*, 2022. 1, 2, 5, 6

[12] Kristjan Greenewald, Anming Gu, Mikhail Yurochkin, Justin Solomon, and Edward Chien. k-mixup regularization for deep learning via optimal transport. *arXiv preprint arXiv:2106.02933*, 2021. 1, 5, 6

[13] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016. 7

[14] Seonghyeon Hwang and Steven Euijong Whang. Mixrl: Data mixing augmentation for regression using reinforcement learning. *ArXiv*, abs/2106.03374, 2021. 5, 6

[15] Mahsa Keramati, Lili Meng, and R. David Evans. Conr: Contrastive regularizer for deep imbalanced regression. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 2

[16] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5400–5409, 2018. 1, 6

[17] K. Scott Mader. Biwi kinect head pose database. `https://www.kaggle.com/datasets/kmader/biwi-kinect-head-pose-database`. Kaggle, accessed 2025-09-10. 5

[18] Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. `https://github.com/google-deepmind/dsprites-dataset`, 2017. DeepMind, accessed 2025-09-10. 5

[19] Ismail Nejjar, Qin Wang, and Olga Fink. Dare-gram: Unsupervised domain adaptation regression by aligning inverse gram matrices. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11744–11754, 2023. 2

[20] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1, 6, 7

[21] Junhyung Park and Krikamol Muandet. A measure-theoretic approach to kernel conditional mean embeddings. *Advances in neural information processing systems*, 33:21247–21259, 2020. 2, 4

[22] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019. 5

[23] Pedro O Pinheiro. Unsupervised domain adaptation with similarity learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8004–8013, 2018. 2

[24] Roman Pogodin, Namrata Deka, Yazhe Li, Danica J Sutherland, Victor Veitch, and Arthur Gretton. Efficient conditionally invariant representation learning. *arXiv preprint arXiv:2212.08645*, 2022. 1, 2

[25] Jiawei Ren, Mingyuan Zhang, Cunjun Yu, and Ziwei Liu. Balanced mse for imbalanced visual regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1, 2

[26] Noviyanti T M Sagala and Laura Hestia Cendriawan. House price prediction using linier regression. In *2022 IEEE 8th International Conference on Computing, Engineering and Design (ICCED)*, pages 1–5, 2022. 1

[27] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019. 1, 6

[28] Rajen D Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. 2020. 1, 2

[29] Vikas Sheoran, Shreyansh Joshi, and Tanisha R Bhayani. Age and gender prediction using deep cnns and transfer learning. In *Computer Vision and Image Processing: 5th International Conference, CVIP 2020, Prayagraj, India, December 4-6, 2020, Revised Selected Papers, Part II 5*, pages 293–304. Springer, 2021. 1

[30] Jiayu Su, David A Knowles, and Raul Rabadan. Disentangling interpretable factors with supervised independent subspace principal component analysis. *Advances in Neural Information Processing Systems*, 37:37408–37438, 2024. 5

[31] Ilya O Tolstikhin, Bharath K Sriperumbudur, and Bernhard Schölkopf. Minimax estimation of maximum mean discrepancy with radial kernels. *Advances in Neural Information Processing Systems*, 29, 2016. 2

[32] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 7

[33] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *International conference on machine learning*, pages 6438–6447. PMLR, 2019. 5, 6

[34] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 2022. 2

[35] Xin Wang, Jielong Yang, and Yixing Gao. Universal domain alignment framework for classification and regression tasks. *IEEE Transactions on Artificial Intelligence*, pages 1–14, 2025. 2

[36] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017. 5

[37] Chengming Xu, Yanwei Fu, Bing Zhang, Zitian Chen, Yu-Gang Jiang, and Xiangyang Xue. Learning to score figure skating sport videos. *IEEE transactions on circuits and systems for video technology*, 30(12):4578–4590, 2019. 7

[38] Yuzhe Yang, Kaiwen Zha, Yingcong Chen, Hao Wang, and Dina Katabi. Delving into deep imbalanced regression. In *International Conference on Machine Learning*, pages 11842–11851. PMLR, 2021. 1, 2, 5, 6

[39] Huaxiu Yao, Yiping Wang, Linjun Zhang, James Zou, and Chelsea Finn. C-mixup: Improving generalization in regression, 2022. 1, 2, 5, 6, 7, 8

[40] Christopher Yeh, Anthony Perez, Anne Driscoll, George Azzari, Zhongyi Tang, David Lobell, Stefano Ermon, and Marshall Burke. Using publicly available satellite imagery and deep learning to understand economic well-being in africa. *Nature communications*, 11(1):2583, 2020. 5, 7

[41] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 1, 5, 6

[42] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2

[43] Christian Zimmermann and Thomas Brox. Rendered hand pose dataset (rhd). https://lmb.informatik.uni-freiburg.de/resources/datasets/RenderedHandposeDataset.en.html. University of Freiburg, accessed 2025-09-10. 5