

Discovery of New Multi-Level Features for Domain Generalization via Knowledge Corruption

Ahmed Frikha
Siemens Technology
University of Munich (LMU)
ahmed.frikha@siemens.com

Denis Krompaß
Siemens Technology
denis.krompass@siemens.com

Volker Tresp
Siemens Technology
University of Munich (LMU)
volker.tresp@siemens.com

Abstract—Machine learning models that can generalize to unseen domains are essential when applied in real-world scenarios involving strong domain shifts. We address the challenging domain generalization (DG) problem, where a model trained on a set of source domains is expected to generalize well in unseen domains without any exposure to their data. The main challenge of DG is that the features learned from the source domains are not necessarily present in the unseen target domains, leading to performance deterioration. We assume that learning a richer set of features is crucial to improve the transfer to a wider set of unknown domains. For this reason, we propose COLUMBUS, a method that enforces new feature discovery via a targeted corruption of the most relevant input and multi-level representations of the data. We conduct an extensive empirical evaluation to demonstrate the effectiveness of the proposed approach which achieves new state-of-the-art results by outperforming 18 DG algorithms on multiple DG benchmark datasets in the DOMAINBED framework.

I. INTRODUCTION

Deep learning models have achieved tremendous success when applied to independent and identically distributed (i.i.d.) data. However, in real-world applications, distribution shifts between training and test data are commonly encountered. For instance, data distributions might differ from one hospital to another [1], and from one production plant to another. Similarly, the models in self-driving cars are exposed to different urban and rural environments in different countries with changing weather conditions [2] and object poses [3].

Approaches to make machine learning models resilient to such data distribution changes were studied for different domain shift settings. For example, several domain adaptation methods were developed to address the case where, besides the data from the source domain(s), a set of labeled [5] or unlabeled [6] data is available from a specific target domain. However, in real-world scenarios, collecting data from the target domain(s) is often slow, e.g., a new hospital or production site, expensive, or even infeasible, e.g., collecting images from every street of every country in the context of self-driving cars. Sometimes, the target domains cannot be known beforehand.

The Domain Generalization (DG) problem [7], [8] was introduced to address such cases. Specifically, a model trained on multiple source domains is expected to directly perform well in unseen target domains without requiring any exposure to its data. This problem setting can be interpreted as multi-source 0-shot domain adaptation.

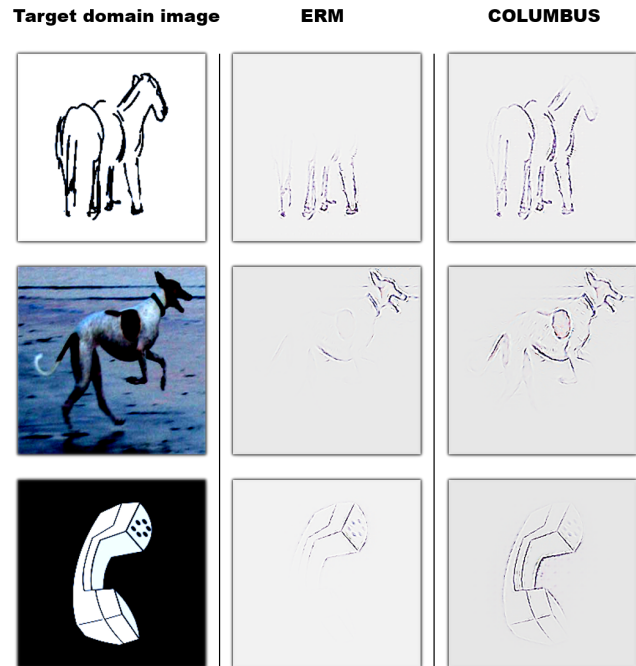


Fig. 1. Relevance maps computed with GuidedGrad-CAM [4] for ERM and COLUMBUS using images from the target domains, PACS *Sketch*, VLCS *VOC* and OfficeHome *Clipart*. COLUMBUS recognizes more features than ERM, including horse back and muzzle, dog legs and tail, and phone shape.

Training a model to generalize across several related but unseen data distributions remains arguably one of the most challenging open problems in machine learning. In the last decade, a plethora of widely different methods were developed to address the DG problem. We refer to [9] for an extensive overview of DG algorithms. Despite these efforts, [10] found that carefully tuning the baseline, which simply applies Empirical Risk Minimization (ERM) on the data of the source domains, achieves a high performance that is competitive with state-of-the-art methods.

One major challenge of DG is that the model can only observe and learn features from the source domains, which may not be present in the unseen target domains, limiting generalization. We presume that learning a wider set of different features would increase the chance of learning features that are useful for a larger set of unseen domains. Hence, we introduce

COLUMBUS, a training procedure for automated new feature discovery, which leads to a better feature recognition in unseen domains (Figure 1). During training on the source domains, COLUMBUS incentivizes the model to discover new features, even in data examples on which it already performs well. To achieve this, COLUMBUS prevents the model from using the features it deems most relevant for the source domains by corrupting them during training. To identify the most relevant features for a model, we leverage attribution methods [11] usually used for model explainability purposes.

We evaluate our approach on the recently proposed DOMAINBED framework [10] which includes several DG datasets and algorithm implementations to promote a fair and reproducible comparison of different approaches. Our method outperforms 18 DG algorithms evaluated on 3 datasets in the DOMAINBED framework, achieving new state-of-the-art results using 2 different model selection methods. Furthermore, our method achieves the highest performance when evaluated on unseen data from the source domains used for training (in-domain generalization), which confirms its effectiveness and ability to learn new features useful for unseen data.

II. RELATED WORK

A. Domain Generalization

This section presents an overview of domain generalization (DG) approaches. Methods to which we compare in our experiments (Section IV) are highlighted in bold. We refer to [9] for an extensive overview of DG algorithms. The simplest approach to DG is to train one model via Empirical Risk Minimization (**ERM**) [12] on the training datasets of all source domains. **GroupDRO** [13] additionally increases the importance of source domains where the model yields a lower performance. In the following, we broadly categorize DG approaches into three categories.

Domain alignment methods aim to learn domain-invariant representations of the data by aligning features across the source domains. The reduction of the representation distribution mismatch across source domains can be achieved by minimizing the maximum mean discrepancy criteria [14] combined with an adversarial autoencoder (**MMD**) [15], minimizing the difference between the means [16] or covariance matrices (**CORAL**) [17] in the embedding space across different domains, or minimizing a contrastive loss [18]–[20], e.g., **SelfReg** [21]. Domain alignment is also performed by aligning the loss gradients across source domains via inner product maximization (**Fish**) [22], or binary (**AND-mask**) [23], [24] or continuous gradient masking (**SAND-mask**) [24].

Another line of works optimizes for features that confuse a domain discriminator model [25]–[28], and includes **DANN** [29] and its class-conditional extension **C-DANN** [30]. Other works additionally involve the classifier in the representation alignment, either by optimizing for an embedding space such that the optimal linear classifier on top of it is the same across different domains (**IRM**) [31], or by passing a domain-specific mean embedding to the classifier as a second argument (**MTL**) [32]. **VREx** [33] is an approximation of IRM via a variance

penalty and **ARM** [34] is an extension of MTL that employs a separate embedding CNN.

Meta-learning techniques were applied to DG by training a model in a bi-level optimization scheme on meta-train and meta-test sets sampled from the source domains. Hereby, **MLDG** [35] optimizes for parameters that can be quickly adapted to different domains, **MASF** [1] adds inter-class and intra-class losses to regularize the embedding space, and **MetaReg** meta-learns a regularizer for the output layer [36].

Data augmentation approaches were proposed to tackle DG and our method falls into this category. Some works use **Mixup** [37] to compute inter-domain examples to augment the training set [38]–[40]. **SagNets** [41] reduce the domain gap by randomizing the style of images while keeping their content. Another line of works generate images by using adversarial attacks [42] to perturb input images based on a class classifier [43]–[45] or a domain classifier [46], by training CNNs to generate images within the source domains [47]–[49] or novel domains [50]–[52]. Other works apply such perturbations on a feature level [53], [54].

Our approach corrupts the raw input data as well as the multi-level representations that the model learns in order to enforce new feature discovery. Instead of using visually undetectable adversarial attacks or highly parametrized generative models, we employ attribution methods, e.g., Guided-Grad-CAM [4], to identify and corrupt the most relevant features. Our approach shares similarities with **RSC** [53] which discards the most dominant features fed to the output layer to promote the activation of the remaining features. The key difference of our approach is that we corrupt features not only in the last high-level representation space, i.e., the input to the output layer, but also in the raw input space and other low-level representation spaces. We argue that by discarding the features only in the representation space (e.g., elephant trunk detector), as done in RSC, the same silenced feature detectors can be relearned as long as the model is exposed to the corresponding features in the input space (e.g., the pixels of the elephant trunk). We hypothesize that corrupting the features in the input space is crucial to enforce the discovery of new features. Our empirical results show that our method outperforms RSC by a significant margin (Section IV) on unseen data from source and target domains, hence confirming our hypothesis.

B. Relevance Attribution

In an attempt to explain and interpret the predictions of deep learning models, several attribution methods that assign relevance scores to input features have been developed [4], [55], [56]. In Saliency Maps [55] the relevance scores are given by the gradient of the output neuron corresponding to the ground truth w.r.t. the input. Better attributions were achieved by averaging these gradients over local neighborhood patches in SmoothGrad [57] and over brightness level interpolations in IntegratedGradients [58]. Another category of approaches modifies the backpropagation procedure by considering only positive gradients [59] or to satisfy the relevance conservation property through the layers [60], [61]. Class Activation Maps

(CAM) [62] leverages the activations in the last convolutional layer to produce a heatmap highlighting the relevance of each feature in the raw input. Gradient-weighted CAM (Grad-CAM) [4] generalizes CAM to a variety of CNNs by using the gradient information flowing into the last convolutional layer. This method can be combined with GuidedBP [59] to yield GuidedGrad-CAM [4]. IBA [56] approximates attribution scores by restricting the information flow via noise injection to intermediate feature maps during the forward pass.

While prior works used attribution methods to explain and interpret model predictions, we leverage them for training purposes. To the best of our knowledge, we are the first to incorporate attribution methods combined with data corruption into training to improve the model’s generalization ability. For a broader overview of attribution methods, we refer to [11].

III. METHOD

The proposed method improves the knowledge transfer to unknown data distributions by training a model to learn a rich set of features on several representation levels of the data via an automated new feature discovery.

Let F_s and F_t denote the sets of features learnable for the addressed classification task, which are present in the data of the source domains and in the target domain, respectively, and G their intersection. In the optimal case, the set of features L learned by the model on the source domains encompasses G fully. Since F_t is unknown at training time, our method maximizes the size of L by training the model to learn as many features as possible, resulting in a higher chance to capture features from G via the higher intersection between L and G . To achieve this, we propose COLUMBUS, a training procedure that enables automated new feature discovery. COLUMBUS prevents the model from using (a part of) the most relevant features for its current predictions during training. This is done in 3 major steps: identification of the most relevant features, their corruption, and training with the corrupted data representations. Figure 2 presents an overview of our approach. We apply this technique on several levels of representations of the data ranging from the raw input, e.g., pixels of the elephant trunk, to the high-level features fed to the output layers, e.g., elephant-trunk-detector, including the representations yielded by intermediate layers, hence fostering multi-level new feature discovery.

A. Identification

In each training iteration, we sample a method from a set of attribution methods A , and use it to compute an attribution map M that identifies the most relevant features. Any attribution method can be included in the set A . In this work, we use Saliency Maps [55] and GuidedGrad-CAM [4], since they are simple, fast, and model-architecture-agnostic. Other methods require modifications to support skip connections and batch normalization layers [60] or involve training additional parameters after each update [56]. Moreover, GuidedGrad-CAM was found to be competitive with the state-of-the-art attribution methods in the image degradation evaluation [56].

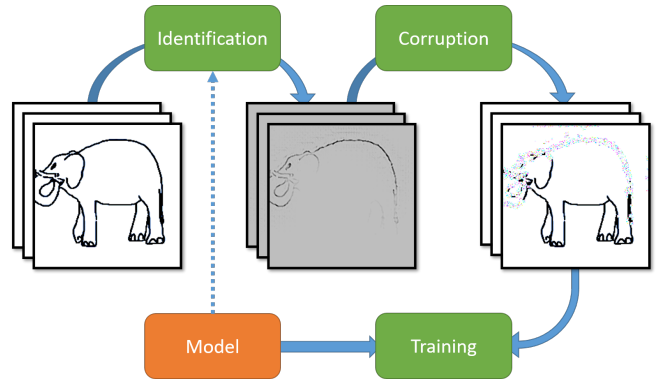


Fig. 2. Overview of the proposed COLUMBUS method. In the identification stage, the most class-discriminative features according to the current model are identified via a relevance attribution method, which in this case is applied to the raw input representation. In the corruption stage, the identified features, e.g., elephant trunk and back, are perturbed by using a corruption method, in this case a replacement by a random pixel. Finally, the model is trained with the batch of corrupted data, promoting the discovery of new features, e.g., elephant feet and toes. The image used belongs to the PACS Sketch domain.

While Saliency Maps and GuidedGrad-CAM were developed to assign relevance scores to features in the input space, we extend their usage to identify relevant features in representations extracted by intermediate layers. Let \mathbf{y} and $\hat{\mathbf{y}}$ denote the ground truth and the model prediction for a datapoint \mathbf{X} . The attribution map M_l yielded by Saliency Maps for a representation R_l yielded by layer l is given by

$$M_{l,Saliency} = \frac{\partial(\hat{\mathbf{y}} \odot \mathbf{y})}{\partial R_l}. \quad (1)$$

Note that the original Saliency Maps method [55] corresponds to the case where $l = 0$, i.e., R_0 is the raw input representation.

The class-discriminative relevance map M_l yielded by Grad-CAM [4] for a layer l is given by a sum over the channels of the representation R_l weighted by importance factors α_c for each channel c , resulting in

$$M_{l,GradCAM} = ReLU\left(\sum_c \alpha_c R_l^c\right). \quad (2)$$

Hereby, the importance factors are given by the gradient of the model prediction for the correct class w.r.t. the global-averaged-pooled representation R_l . Formally,

$$\alpha_c = \frac{1}{Z} \sum_i \sum_j \frac{\partial(\hat{\mathbf{y}} \odot \mathbf{y})}{\partial R_l^{c,i,j}}. \quad (3)$$

Grad-CAM is applied to the representation yielded by the last convolutional layer to obtain a relevance map M which is upsampled to the input size [4]. For intermediate representation R_l , we use the corresponding relevance map M_l (Eq. 2). We use GuidedGrad-CAM [4] which yields more fine-grained maps than Grad-CAM by multiplying $M_{l,GradCAM}$ with the relevance maps determined by Guided Backpropagation (Guided BP) [59], for the same representation R_l . Guided BP modifies Saliency Maps by removing negative gradients when backpropagating through ReLU layers.

B. Corruption

In each training iteration, we sample a method from the set of corruption methods C and use it to corrupt the identified features based on the relevance attribution map M . Any technique that perturbs the information contained in the identified features can be used. We use different corruption methods depending on the sampled representation level l .

To corrupt the raw input ($l = 0$), the most relevant input features according to the attribution map M_0 , e.g., the pixels corresponding to an elephant trunk, are perturbed using a corruption method. Hereby, we perturb the identified pixel values by setting them to a random value, to zero, i.e., black pixels, by applying the Fast Gradient Sign Method (FGSM) [42], or by applying Gaussian blurring. For an intermediate representation level $l > 0$, first the original input is fed through the model up to the corresponding layer l to yield the representation R_l . The latter is then corrupted based on the relevance attribution map M_l resulting from the identification stage and finally fed to the next layer. To corrupt intermediate embeddings, we drop the most relevant features, i.e., set their values to zero. This can be viewed as a targeted Dropout [63].

C. Training

The COLUMBUS training procedure is described by Algorithm 1. In each training iteration, a data batch B , a representation level l , an attribution method, and a corruption method are sampled. Subsequently, the aforementioned identification and corruption steps are performed. The model is trained on the corrupted data (representations). Hereby, the $p\%$ most relevant features are corrupted. To enable the model to learn some features at the beginning of training, p is set to 0, i.e., no corruption is applied. As training progresses, more features are corrupted in the data, forcing the model to discover and learn new features. Concretely, p is linearly increased throughout the training to reach p_{max} , a hyperparameter. Note that the resulting gradual learning of new features, independently from each other, promotes also feature disentanglement, which was found to be beneficial for visual reasoning [64]. We use different identification and corruption methods to increase the diversity of the corrupted datapoints used to train the model and prevent overfitting. We also found that sampling multiple methods leads to better empirical results. It should also be noted that COLUMBUS is adaptive to the model’s learning, since the identification step is model-state-specific. In other words, if the model *forgets* a set of features during training, these will not be identified (again), and hence will not be corrupted (again), which enables the model to relearn them.

The model parameters θ are updated by minimizing a loss function f using a gradient-based optimization algorithm, e.g., Adam [65]. In algorithm 1, SGD is used for simplicity of notation. The loss function f used is a weighted sum comprising a classification loss L_{cls} and a domain-alignment regularization loss term L_{DA} . Formally,

$$f = L_{cls} + \lambda \sum_{i=1}^{N_s} \sum_{j=i+1}^{N_s} L_{DA}(i, j), \quad (4)$$

where λ is a weighting factor and N_s the number of source domains. We use cross-entropy as classification loss L_{cls} . L_{DA} regularizes the embedding space by minimizing the l_2 -norm between the domain-specific embedding means and covariance matrices for each pair of source domains i and j , as in DDC [16] and CORAL [17] respectively. The normalization of the regularization loss by the number of source domain pairs is omitted for simplicity of notation. We note that, unlike prior works [10], [16], [17] that apply this loss on the representations of the original images, we use corrupted images and representations. This leads to an alignment in the embedding space not only across the domains, but also between class-specific features, e.g., by minimizing the difference between the embedding distribution of cartoon images of elephant feet and art painting images of elephant trunks.

Algorithm 1 The COLUMBUS training procedure

Require: D_s : Training data of all source domains

Require: f : Loss function

Require: α : Learning rate

Require: L : Set of representation levels including the raw input level

Require: A : Set of relevance attribution methods

Require: C : Set of corruption methods

Require: p_{max} : Max. % of representation to be corrupted

- 1: Initialize the model parameters θ randomly or from a pre-trained model
 - 2: Initialize the % of representation to be corrupted $p = 0$
 - 3: **while** not done **do**
 - 4: Sample a data batch $B = \{\mathbf{X}, \mathbf{y}\}$ from D_s
 - 5: Sample level of representation l from L
 - 6: Feed B through the model parametrized by θ
 - 7: Get the relevance attribution map M_l by applying a relevance attribution method randomly sampled from A on level l using the current model parameters θ
 - 8: **if** $l = 0$ **then**
 - 9: Corrupt the values in B corresponding to the $p\%$ highest values in M_0 by applying a corruption method randomly sampled from C , yielding the corrupted input B_c
 - 10: Feed B_c through the model to obtain the predictions $\hat{\mathbf{y}}$
 - 11: **else**
 - 12: Feed B through the model until level l to obtain the representation R_l
 - 13: Corrupt the values in R_l corresponding to the $p\%$ highest values in M_l , yielding the corrupted representation $R_{l,c}$
 - 14: Feed $R_{l,c}$ through the model starting from level $l + 1$ to obtain the predictions $\hat{\mathbf{y}}$
 - 15: **end if**
 - 16: Update θ : $\theta \leftarrow \theta - \alpha \nabla_{\theta} f(\mathbf{y}, \hat{\mathbf{y}})$
 - 17: Increase p linearly towards p_{max}
 - 18: **end while**
 - 19: **return** Learned model parameters θ
-

In our experiments, we corrupt $q\%$ of the sampled data batch in each iteration, and increase q linearly during the training until q_{max} is reached, as done for the representation percentage to be corrupted p . This is omitted in Algorithm 1 for simplicity of notation. During training, we alternate between sampling an intermediate representation and the raw input for corruption. The intermediate representations correspond to the outputs of each ResNet block in the used ResNet-50 model [66]. At test time, the model trained with COLUMBUS is applied to the data from the target domains without any corruption.

IV. EXPERIMENTS

A. Experimental Setup

We evaluate our approach empirically* on the recently proposed DOMAINBED framework [10] which includes several DG datasets, implementations of DG algorithms, and model selection methods. DOMAINBED promotes a fair and reproducible comparison of the different approaches by including a common automated hyperparameter search, i.e., a random search with the given seeds conducts the same experiments for all methods. For a fair comparison with the 18 DG algorithms, our experiments follow the same experimental setting adopted in DOMAINBED [10]: We use a ResNet-50 model [66] pre-trained on ImageNet [67] with frozen batch normalization [68] statistics as suggested in [69], the same optimization algorithm, data augmentation techniques and number of training iterations used in DOMAINBED. The COLUMBUS-specific hyperparameters p_{max} , q_{max} and λ are included in the hyperparameter search of DOMAINBED, and the intervals used can be found in the Appendix. We noticed that the published code [10] with the provided seeds does not enable the reproduction of the published results, since the resulting points in the hyperparameter search space are different from the ones used for the published results. Therefore, for a fairer comparison, we additionally rerun the experiments of the best performing DG method in DOMAINBED, i.e., CORAL, with the published code and seeds that we used for COLUMBUS.

We conduct experiments on 3 challenging multi-domain datasets commonly used as DG benchmarks: VLCS [70], OfficeHome [71] and PACS [72]. VLCS contains images belonging to 5 classes from 4 photographic domains: VOC2007 (V), LabelMe (L), Caltech101 (C), and SUN09 (S). OfficeHome consists of images of 65 classes from the domains Art (A), Clipart (C), Product (P), and Real (R). PACS comprises images belonging to 7 classes from the domains Art-painting (A), Cartoon (C), Photo (P), and Sketch (S). DOMAINBED splits each source domain data into 80% for training and 20% for validation. Each experiment is run with the provided 3 seeds.

B. Results

Tables I, II and III show the results averaged over the 3 seeds pre-determined by DOMAINBED, on VLCS, PACS and OfficeHome respectively. Hereby, the unseen target domain is defined by the column name, i.e., the 3 other domains are

used as source domains for training. The test accuracy is computed on the test set of the target domain. We provide results including standard deviations in the appendix. The average results over the domains of each dataset can be seen in Table IV. We select the model with the highest source-domain validation performance for the evaluation on the target domain.

TABLE I
DOMAIN GENERALIZATION RESULTS ON VLCS.

Algorithm	C	L	S	V	Avg
ERM	97.7	64.3	73.4	74.6	77.5
IRM	98.6	64.9	73.4	77.3	78.5
GroupDRO	97.3	63.4	69.5	76.7	76.7
Mixup	98.3	64.8	72.1	74.3	77.4
MLDG	97.4	65.2	71.0	75.3	77.2
CORAL	98.3	66.1	73.4	77.5	78.8
MMD	97.7	64.0	72.8	75.3	77.5
DANN	99.0	65.1	73.1	77.2	78.6
CDANN	97.1	65.1	70.7	77.1	77.5
MTL	97.8	64.3	71.5	75.3	77.2
SagNet	97.9	64.5	71.4	77.5	77.8
ARM	98.7	63.6	71.3	76.7	77.6
VREx	98.4	64.4	74.1	76.2	78.3
RSC	97.9	62.5	72.3	75.6	77.1
CORAL [†]	97.3	65.2	71.5	75.6	77.4
COLUMBUS	98.9	65.0	75.0	77.9	79.2

TABLE II
DOMAIN GENERALIZATION RESULTS ON PACS.

Algorithm	A	C	P	S	Avg
ERM	84.7	80.8	97.2	79.3	85.5
IRM	84.8	76.4	96.7	76.1	83.5
GroupDRO	83.5	79.1	96.7	78.3	84.4
Mixup	86.1	78.9	97.6	75.8	84.6
MLDG	85.5	80.1	97.4	76.6	84.9
CORAL	88.3	80.0	97.5	78.8	86.2
MMD	86.1	79.4	96.6	76.5	84.6
DANN	86.4	77.4	97.3	73.5	83.6
CDANN	84.6	75.5	96.8	73.5	82.6
MTL	87.5	77.1	96.4	77.3	84.6
SagNet	87.4	80.7	97.1	80.0	86.3
ARM	86.8	76.8	97.4	79.3	85.1
VREx	86.0	79.1	96.9	77.7	84.9
RSC	85.4	79.7	97.6	78.2	85.2
CORAL [†]	87.4	79.4	97.5	73.9	84.5
COLUMBUS	88.7	78.7	97.2	81.5	86.5

COLUMBUS achieves the highest results on all datasets on average, advancing the state-of-the-art by 1.6% and 1.2% compared to ERM and CORAL respectively. We note an impressive 5.5% improvement on OfficeHome’s most challenging domain *Clipart* (C) compared to ERM and 3.3% compared to CORAL, on this 65-class classification task. Likewise, on the *Art* (A) domain of PACS, substantial 4% and 1.3% increases are observed compared to ERM and CORAL respectively. A significant performance increase is achieved

*Code under <https://github.com/AhmedFrikha/columbus-domainbed>.

[†]Results yielded by using published code [10] with the provided seeds.

TABLE III
DOMAIN GENERALIZATION RESULTS ON OFFICEHOME.

Algorithm	A	C	P	R	Avg
ERM	61.3	52.4	75.8	76.6	66.5
IRM	58.9	52.2	72.1	74.0	64.3
GroupDRO	60.4	52.7	75.0	76.0	66.0
Mixup	62.4	54.8	76.9	78.3	68.1
MLDG	61.5	53.2	75.0	77.5	66.8
CORAL	65.3	54.4	76.5	78.4	68.7
MMD	60.4	53.3	74.3	77.4	66.3
DANN	59.9	53.0	73.6	76.9	65.9
CDANN	61.5	50.4	74.4	76.6	65.8
MTL	61.5	52.4	74.9	76.8	66.4
SagNet	63.4	54.8	75.8	78.3	68.1
ARM	58.9	51.0	74.1	75.2	64.8
VREx	60.7	53.0	75.3	76.6	66.4
RSC	60.7	51.4	74.8	75.1	65.5
CORAL [†]	64.8	54.6	76.8	78.4	68.6
COLUMBUS	62.8	57.9	75.5	77.9	68.5

TABLE IV
AVERAGE DOMAIN GENERALIZATION RESULTS.

Algorithm	VLCS	PACS	OfficeHome	Avg
ERM	77.5 ± 0.4	85.5 ± 0.2	66.5 ± 0.3	76.5
IRM	78.5 ± 0.5	83.5 ± 0.8	64.3 ± 2.2	75.5
GroupDRO	76.7 ± 0.6	84.4 ± 0.8	66.0 ± 0.7	75.7
Mixup	77.4 ± 0.6	84.6 ± 0.6	68.1 ± 0.3	76.7
MLDG	77.2 ± 0.4	84.9 ± 1.0	66.8 ± 0.6	76.3
CORAL	78.8 ± 0.6	86.2 ± 0.3	68.7 ± 0.3	77.9
MMD	77.5 ± 0.9	84.6 ± 0.5	66.3 ± 0.1	76.2
DANN	78.6 ± 0.4	83.6 ± 0.4	65.9 ± 0.6	76.0
CDANN	77.5 ± 0.1	82.6 ± 0.9	65.8 ± 1.3	75.3
MTL	77.2 ± 0.4	84.6 ± 0.5	66.4 ± 0.5	76.1
SagNet	77.8 ± 0.5	86.3 ± 0.2	68.1 ± 0.1	77.4
ARM	77.6 ± 0.3	85.1 ± 0.4	64.8 ± 0.3	75.8
VREx	78.3 ± 0.2	84.9 ± 0.6	66.4 ± 0.6	76.5
RSC	77.1 ± 0.5	85.2 ± 0.9	65.5 ± 0.9	75.9
SelfReg	77.8 ± 0.9	85.6 ± 0.4	67.9 ± 0.7	77.1
Fish	77.8 ± 0.3	85.5 ± 0.3	68.6 ± 0.4	77.3
AND-mask	78.1 ± 0.9	84.4 ± 0.9	65.6 ± 0.4	76.0
SAND-mask	77.4 ± 0.2	84.6 ± 0.9	65.8 ± 0.4	75.9
CORAL [†]	77.4 ± 0.3	84.5 ± 0.5	68.6 ± 0.2	76.9
COLUMBUS	79.2 ± 0.2	86.5 ± 0.4	68.5 ± 0.4	78.1

on PACS’s challenging *Sketch (S)* domain as well. On all target domains, COLUMBUS consistently outperforms all the baselines or yields a competitive performance. The fact that COLUMBUS outperforms RSC [53] confirms our hypothesis, that corrupting the learned features in the raw input is crucial to prevent relearning the same high-level features, and hence enforce new feature discovery.

We also evaluate our approach using the oracle selection method [10], where the model is evaluated on a held-out validation set from the target domain. In order to limit access to the target domain, this evaluation is performed only once at the end of each training, disallowing early stopping. The average results are presented in Table V. We find that the performance advantage of COLUMBUS is increased when better proxies for model selection, e.g., a held-out set from the target domain, are available, further confirming the effectiveness of

TABLE V
DOMAIN GENERALIZATION RESULTS USING THE TEST-DOMAIN VALIDATION SET (ORACLE) AS A SELECTION METHOD.

Algorithm	VLCS	PACS	OfficeHome	Avg
ERM	77.6 ± 0.3	86.7 ± 0.3	66.4 ± 0.5	76.9
IRM	76.9 ± 0.6	84.5 ± 1.1	63.0 ± 2.7	74.8
GroupDRO	77.4 ± 0.5	87.1 ± 0.1	66.2 ± 0.6	76.9
Mixup	78.1 ± 0.3	86.8 ± 0.3	68.0 ± 0.2	77.6
MLDG	77.5 ± 0.1	86.8 ± 0.4	66.6 ± 0.3	77.0
CORAL	77.7 ± 0.2	87.1 ± 0.5	68.4 ± 0.2	77.7
MMD	77.9 ± 0.1	87.2 ± 0.1	66.2 ± 0.3	77.1
DANN	79.7 ± 0.5	85.2 ± 0.2	65.3 ± 0.8	76.8
CDANN	79.9 ± 0.2	85.8 ± 0.8	65.3 ± 0.5	77.0
MTL	77.7 ± 0.5	86.7 ± 0.2	66.5 ± 0.4	77.0
SagNet	77.6 ± 0.1	86.4 ± 0.4	67.5 ± 0.2	77.2
ARM	77.8 ± 0.3	85.8 ± 0.2	64.8 ± 0.4	76.1
VREx	78.1 ± 0.2	87.2 ± 0.6	65.7 ± 0.3	77.0
RSC	77.8 ± 0.6	86.2 ± 0.5	66.5 ± 0.6	76.8
AND-mask	76.4 ± 0.4	86.4 ± 0.4	66.1 ± 0.2	76.3
SAND-mask	76.2 ± 0.5	85.9 ± 0.4	65.9 ± 0.5	76.0
CORAL [†]	77.4 ± 0.6	85.6 ± 0.8	68.4 ± 0.4	77.1
COLUMBUS	77.7 ± 0.4	88.2 ± 0.2	69.6 ± 0.4	78.5

our approach. Our results on DOMAINBED using both model selection methods show that the additional features learned thanks to the corruption of the most relevant features are useful for generalization to unseen domains. This is backed by Figure 1, where COLUMBUS recognizes more features in examples from the unseen target domain than ERM.

Finally, we investigate whether the richer set of features learned by COLUMBUS leads to a better in-domain generalization, i.e., whether a performance boost is also yielded on unseen source domain data. We evaluate COLUMBUS and the DG baselines on the held-out validation sets of the source domains and report the maximal mean validation accuracy across domains in the Appendix. COLUMBUS consistently achieves the highest validation performance on the training domains compared to the DG baselines. This shows that the richer set of learned features improves generalization to unseen in-distribution datapoints, suggesting that COLUMBUS might also be suitable for applications without domain shift.

V. CONCLUSION

In this work, we proposed COLUMBUS, a novel and strong domain generalization (DG) approach that enforces new feature discovery to improve the transfer to a wider set of unseen domains. During training, COLUMBUS corrupts the input and multi-level representations of the data most relevant for the model. For the identification of such features, relevance attribution methods that are usually used for model explainability purposes are leveraged. Our extensive empirical evaluation on DOMAINBED demonstrates the effectiveness of the proposed method, which outperforms 18 DG algorithms and achieves new state-of-the-art results on multiple DG benchmarks. Our results show that the richer set of learned features improves the generalization to unseen data from both seen and unseen domains, suggesting the suitability of our approach for applications beyond domain generalization to include scenarios without domain shift.

Experimental Setting Details In this section we provide further details about the experiments conducted. The experiments were conducted on computing instances that include a Tesla T4 NVIDIA GPU, 8 custom Intel Cascade Lake CPUs and 32 Gb of memory. The operating system used is Ubuntu 20.04 LTS. The libraries PyTorch [73] and TorchVision were used with the versions 1.7.1 and 0.8.2, respectively.

In our experiments, the percentage of representation corrupted p and the percentage of the batch corrupted q are increased linearly towards p_{max} and q_{max} , respectively, during the first half of the training. In the second half of the training, the maximum values are used.

For a fair comparison, we used the automated hyperparameter search from DOMAINBED [10] for each domain and dataset. Hereby, each hyperparameter search involves 20 random search experiments, i.e., the hyperparameters are randomly sampled from the specified intervals. To distribute the hyperparameter search experiments over multiple devices (each experiment runs on a single GPU), we used the Ray Tune package [74], [75]. Our experiments follow the experimental setting: We use a ResNet-50 model [66] pretrained on ImageNet [67] with frozen batch normalization [68] statistics as suggested in [69], as well as the same optimization algorithm, ADAM [65], data augmentation techniques, and number of training iterations. An overview of the hyperparameter-specific intervals we used for COLUMBUS can be seen in Table VI. The algorithm-specific hyperparameter intervals used for the other DG algorithms can be found in [10]. Depending on whether the corruption is applied on the input or an intermediate representation, different value intervals were used for the percentage of the representation corrupted p and the percentage of the batch corrupted q . For the hyperparameters related to intermediate representations, i.e., $p_{max,intermediate}$ and $q_{max,intermediate}$, the interval upper bounds were chosen based on the results of RSC [53], which discards the most dominant features fed to the output layer, i.e., the last representation level. We used the same intervals used in DOMAINBED for the other algorithms for all hyperparameters.

Source Domain Generalization

In this section, we investigate whether the richer set of features learned by COLUMBUS leads to a better in-domain generalization, i.e., whether a performance boost is also yielded on unseen data from the source domains used for training. We evaluate COLUMBUS and the DG baselines on the held-out validation sets of the source domains and report the maximal average validation accuracy across domains in Table VII*.

COLUMBUS consistently achieves the highest validation performance on the training domains compared to the DG

baselines, on every dataset. This shows that the richer set of learned features improves generalization to unseen in-distribution data examples as well, suggesting that COLUMBUS might be suitable for applications beyond domain generalization to include scenarios without domain shift.

Results including standard deviations We present the domain generalization results of COLUMBUS and the baselines, including the standard deviations computed over the 3 runs with the seeds provided by DOMAINBED in Tables VIII, IX and X. Hereby, for model selection, the *training-domain validation-set* from DOMAINBED is used.

*For the baselines, we computed the results using the logs made public in <https://drive.google.com/file/d/16VFQWTble6-nB5AdXBtQpQFwjEC7CChM/view?usp=sharing>.

TABLE VI
HYPERPARAMETER INTERVALS USED FOR THE HYPERPARAMETER SEARCH CONDUCTED WITH DOMAINBED

Hyperparameter	Random Distribution
Weighting coefficient λ	$10\text{Uniform}(-1,1)$
Max. corruption % for input representation $p_{max,input}$	$\text{Uniform}(0.2, 0.5)$
Max. corruption % for intermediate representation $p_{max,intermediate}$	$\text{Uniform}(0.01, 0.333)$
Max. batch corruption % for input representation $q_{max,input}$	$\text{Uniform}(0.2, 1.0)$
Max. batch corruption % for intermediate representation $q_{max,intermediate}$	$\text{Uniform}(0.1, 0.5)$

TABLE VII
SOURCE DOMAIN VALIDATION PERFORMANCE.

Algorithm	VLCS	PACS	OfficeHome	Avg
ERM	86.4 \pm 0.0	97.0 \pm 0.1	82.1 \pm 0.2	88.5
IRM	85.8 \pm 0.2	96.5 \pm 0.4	79.9 \pm 2.0	87.4
GroupDRO	86.4 \pm 0.0	96.9 \pm 0.1	81.6 \pm 0.2	88.3
Mixup	86.6 \pm 0.1	97.4 \pm 0.1	83.2 \pm 0.3	89.0
MLDG	86.4 \pm 0.1	97.1 \pm 0.1	82.4 \pm 0.3	88.6
CORAL	86.5 \pm 0.0	97.1 \pm 0.1	83.7 \pm 0.2	89.1
MMD	86.4 \pm 0.1	96.9 \pm 0.0	82.0 \pm 0.1	88.4
DANN	86.3 \pm 0.0	96.4 \pm 0.3	80.4 \pm 0.9	87.7
CDANN	86.4 \pm 0.1	96.4 \pm 0.3	80.5 \pm 0.9	87.8
MTL	86.3 \pm 0.0	97.0 \pm 0.0	81.7 \pm 0.2	88.3
SagNet	86.4 \pm 0.0	97.0 \pm 0.2	82.9 \pm 0.4	88.8
ARM	86.3 \pm 0.0	96.5 \pm 0.1	80.2 \pm 0.2	87.7
VREx	86.2 \pm 0.1	96.9 \pm 0.1	81.8 \pm 0.4	88.3
RSC	86.4 \pm 0.3	96.8 \pm 0.2	81.5 \pm 0.3	88.2
CORAL [†]	86.6 \pm 0.1	96.8 \pm 0.2	83.6 \pm 0.0	89.0
COLUMBUS	86.6 \pm 0.1	97.3 \pm 0.0	83.4 \pm 0.1	89.1

TABLE VIII
DOMAIN GENERALIZATION RESULTS ON VLCS, INCLUDING STANDARD DEVIATION.

Algorithm	C	L	S	V	Avg
ERM	97.7 \pm 0.4	64.3 \pm 0.9	73.4 \pm 0.5	74.6 \pm 1.3	77.5
IRM	98.6 \pm 0.1	64.9 \pm 0.9	73.4 \pm 0.6	77.3 \pm 0.9	78.5
GroupDRO	97.3 \pm 0.3	63.4 \pm 0.9	69.5 \pm 0.8	76.7 \pm 0.7	76.7
Mixup	98.3 \pm 0.6	64.8 \pm 1.0	72.1 \pm 0.5	74.3 \pm 0.8	77.4
MLDG	97.4 \pm 0.2	65.2 \pm 0.7	71.0 \pm 1.4	75.3 \pm 1.0	77.2
CORAL	98.3 \pm 0.1	66.1 \pm 1.2	73.4 \pm 0.3	77.5 \pm 1.2	78.8
MMD	97.7 \pm 0.1	64.0 \pm 1.1	72.8 \pm 0.2	75.3 \pm 3.3	77.5
DANN	99.0 \pm 0.3	65.1 \pm 1.4	73.1 \pm 0.3	77.2 \pm 0.6	78.6
CDANN	97.1 \pm 0.3	65.1 \pm 1.2	70.7 \pm 0.8	77.1 \pm 1.5	77.5
MTL	97.8 \pm 0.4	64.3 \pm 0.3	71.5 \pm 0.7	75.3 \pm 1.7	77.2
SagNet	97.9 \pm 0.4	64.5 \pm 0.5	71.4 \pm 1.3	77.5 \pm 0.5	77.8
ARM	98.7 \pm 0.2	63.6 \pm 0.7	71.3 \pm 1.2	76.7 \pm 0.6	77.6
VREx	98.4 \pm 0.3	64.4 \pm 1.4	74.1 \pm 0.4	76.2 \pm 1.3	78.3
RSC	97.9 \pm 0.1	62.5 \pm 0.7	72.3 \pm 1.2	75.6 \pm 0.8	77.1
CORAL [†]	97.3 \pm 0.3	65.2 \pm 0.5	71.5 \pm 0.6	75.6 \pm 0.9	77.4
COLUMBUS	98.9 \pm 0.2	65.0 \pm 1.3	75.0 \pm 0.2	77.9 \pm 0.9	79.2

TABLE IX
DOMAIN GENERALIZATION RESULTS ON PACS, INCLUDING STANDARD DEVIATION.

Algorithm	A	C	P	S	Avg
ERM	84.7 ± 0.4	80.8 ± 0.6	97.2 ± 0.3	79.3 ± 1.0	85.5
IRM	84.8 ± 1.3	76.4 ± 1.1	96.7 ± 0.6	76.1 ± 1.0	83.5
GroupDRO	83.5 ± 0.9	79.1 ± 0.6	96.7 ± 0.3	78.3 ± 2.0	84.4
Mixup	86.1 ± 0.5	78.9 ± 0.8	97.6 ± 0.1	75.8 ± 1.8	84.6
MLDG	85.5 ± 1.4	80.1 ± 1.7	97.4 ± 0.3	76.6 ± 1.1	84.9
CORAL	88.3 ± 0.2	80.0 ± 0.5	97.5 ± 0.3	78.8 ± 1.3	86.2
MMD	86.1 ± 1.4	79.4 ± 0.9	96.6 ± 0.2	76.5 ± 0.5	84.6
DANN	86.4 ± 0.8	77.4 ± 0.8	97.3 ± 0.4	73.5 ± 2.3	83.6
CDANN	84.6 ± 1.8	75.5 ± 0.9	96.8 ± 0.3	73.5 ± 0.6	82.6
MTL	87.5 ± 0.8	77.1 ± 0.5	96.4 ± 0.8	77.3 ± 1.8	84.6
SagNet	87.4 ± 1.0	80.7 ± 0.6	97.1 ± 0.1	80.0 ± 0.4	86.3
ARM	86.8 ± 0.6	76.8 ± 0.5	97.4 ± 0.3	79.3 ± 1.2	85.1
VREx	86.0 ± 1.6	79.1 ± 0.6	96.9 ± 0.5	77.7 ± 1.7	84.9
RSC	85.4 ± 0.8	79.7 ± 1.8	97.6 ± 0.3	78.2 ± 1.2	85.2
CORAL [†]	87.4 ± 0.3	79.4 ± 0.3	97.5 ± 0.1	73.9 ± 1.8	84.5
COLUMBUS	88.7 ± 0.8	78.7 ± 1.0	97.2 ± 0.1	81.5 ± 1.5	86.5

TABLE X
DOMAIN GENERALIZATION RESULTS ON OFFICEHOME, INCLUDING STANDARD DEVIATION.

Algorithm	A	C	P	R	Avg
ERM	61.3 ± 0.7	52.4 ± 0.3	75.8 ± 0.1	76.6 ± 0.3	66.5
IRM	58.9 ± 2.3	52.2 ± 1.6	72.1 ± 2.9	74.0 ± 2.5	64.3
GroupDRO	60.4 ± 0.7	52.7 ± 1.0	75.0 ± 0.7	76.0 ± 0.7	66.0
Mixup	62.4 ± 0.8	54.8 ± 0.6	76.9 ± 0.3	78.3 ± 0.2	68.1
MLDG	61.5 ± 0.9	53.2 ± 0.6	75.0 ± 1.2	77.5 ± 0.4	66.8
CORAL	65.3 ± 0.4	54.4 ± 0.5	76.5 ± 0.1	78.4 ± 0.5	68.7
MMD	60.4 ± 0.2	53.3 ± 0.3	74.3 ± 0.1	77.4 ± 0.6	66.3
DANN	59.9 ± 1.3	53.0 ± 0.3	73.6 ± 0.7	76.9 ± 0.5	65.9
CDANN	61.5 ± 1.4	50.4 ± 2.4	74.4 ± 0.9	76.6 ± 0.8	65.8
MTL	61.5 ± 0.7	52.4 ± 0.6	74.9 ± 0.4	76.8 ± 0.4	66.4
SagNet	63.4 ± 0.2	54.8 ± 0.4	75.8 ± 0.4	78.3 ± 0.3	68.1
ARM	58.9 ± 0.8	51.0 ± 0.5	74.1 ± 0.1	75.2 ± 0.3	64.8
VREx	60.7 ± 0.9	53.0 ± 0.9	75.3 ± 0.1	76.6 ± 0.5	66.4
RSC	60.7 ± 1.4	51.4 ± 0.3	74.8 ± 1.1	75.1 ± 1.3	65.5
CORAL [†]	64.8 ± 0.2	54.6 ± 0.7	76.8 ± 0.6	78.4 ± 0.3	68.6
COLUMBUS	62.8 ± 0.3	57.9 ± 0.8	75.5 ± 0.1	77.9 ± 0.5	68.5

REFERENCES

- [1] Q. Dou, D. Coelho de Castro, K. Kamnitsas, and B. Glocker, "Domain generalization via model-agnostic learning of semantic features," *NeurIPS*, 2019.
- [2] X. Yue, Y. Zhang, S. Zhao, A. Sangiovanni-Vincentelli, K. Keutzer, and B. Gong, "Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data," in *IEEE/CVF ICCV*, 2019.
- [3] M. A. Alcorn, Q. Li, Z. Gong, C. Wang, L. Mai, W.-S. Ku, and A. Nguyen, "Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects," in *IEEE/CVF CVPR*, 2019.
- [4] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *IEEE ICCV*, 2017.
- [5] M. Wang and W. Deng, "Deep visual domain adaptation: A survey," *Neurocomputing*, 2018.
- [6] G. Wilson and D. J. Cook, "A survey of unsupervised deep domain adaptation," *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2020.
- [7] G. Blanchard, G. Lee, and C. Scott, "Generalizing from several related classification tasks to a new unlabeled sample," *NeurIPS*, 2011.
- [8] K. Muandet, D. Balduzzi, and B. Schölkopf, "Domain generalization via invariant feature representation," in *ICML*, 2013.
- [9] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," *arXiv:2103.02503*, 2021.
- [10] I. Gulrajani and D. Lopez-Paz, "In search of lost domain generalization," *arXiv:2007.01434*, 2020.
- [11] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (xai): Toward medical xai," *IEEE Transactions on Neural Networks and Learning Systems*, 2020.
- [12] V. N. Vapnik, "An overview of statistical learning theory," *IEEE transactions on neural networks*, 1999.
- [13] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang, "Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization," *arXiv:1911.08731*, 2019.
- [14] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *JMLR*, 2012.
- [15] H. Li, S. J. Pan, S. Wang, and A. C. Kot, "Domain generalization with adversarial feature learning," in *IEEE CVPR*, 2018.
- [16] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," *arXiv:1412.3474*, 2014.
- [17] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *ECCV*, 2016.
- [18] S. Motiian, M. Piccirilli, D. A. Adjeroh, and G. Doretto, "Unified deep supervised domain adaptation and generalization," in *IEEE ICCV*, 2017.
- [19] C. Yoon, G. Hamarneh, and R. Garbi, "Generalizable feature learning in the presence of data bias and domain class imbalance with application to skin lesion classification," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2019.
- [20] D. Mahajan, S. Tople, and A. Sharma, "Domain generalization using causal matching," *arXiv:2006.07500*, 2020.
- [21] D. Kim, S. Park, J. Kim, and J. Lee, "Selfreg: Self-supervised contrastive regularization for domain generalization," *arXiv:2104.09841*, 2021.
- [22] Y. Shi, J. Seely, P. H. Torr, N. Siddharth, A. Hannun, N. Usunier, and G. Synnaeve, "Gradient matching for domain generalization," *arXiv:2104.09937*, 2021.
- [23] G. Parascandolo, A. Neitz, A. Orvieto, L. Gresele, and B. Schölkopf, "Learning explanations that are hard to vary," *arXiv:2009.00329*, 2020.
- [24] S. Shahtalebi, J.-C. Gagnon-Audet, T. Laleh, M. Faramarzi, K. Ahuja, and I. Rish, "Sand-mask: An enhanced gradient masking strategy for the discovery of invariances in domain generalization," *arXiv:2106.02266*, 2021.
- [25] I. Albuquerque, J. Monteiro, M. Darvishi, T. H. Falk, and I. Mitliagkas, "Generalizing to unseen domains via distribution matching," *arXiv:1911.00804*, 2019.
- [26] R. Shao, X. Lan, J. Li, and P. C. Yuen, "Multi-adversarial discriminative deep domain generalization for face presentation attack detection," in *IEEE/CVF CVPR*, 2019.
- [27] M. M. Rahman, C. Fookes, M. Baktashmotlagh, and S. Sridharan, "Correlation-aware adversarial domain adaptation and generalization," *Pattern Recognition*, 2020.
- [28] Z. Deng, F. Ding, C. Dwork, R. Hong, G. Parmigiani, P. Patil, and P. Sur, "Representation via representations: Domain generalization via adversarially learned invariant representations," *arXiv:2006.11478*, 2020.
- [29] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *JMLR*, 2016.
- [30] Y. Li, X. Tian, M. Gong, Y. Liu, T. Liu, K. Zhang, and D. Tao, "Deep domain generalization via conditional invariant adversarial networks," in *ECCV*, 2018.
- [31] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," *arXiv:1907.02893*, 2019.
- [32] G. Blanchard, A. A. Deshmukh, U. Dogan, G. Lee, and C. Scott, "Domain generalization by marginal transfer learning," *arXiv:1711.07910*, 2017.
- [33] D. Krueger, E. Caballero, J.-H. Jacobsen, A. Zhang, J. Binias, D. Zhang, R. Le Priol, and A. Courville, "Out-of-distribution generalization via risk extrapolation (rex)," in *ICML*, 2021.
- [34] M. Zhang, H. Marklund, N. Dhawan, A. Gupta, S. Levine, and C. Finn, "Adaptive risk minimization: A meta-learning approach for tackling group distribution shift," *arXiv:2007.02931*, 2020.
- [35] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Learning to generalize: Meta-learning for domain generalization," in *AAAI Conference on Artificial Intelligence*, 2018.
- [36] Y. Balaji, S. Sankaranarayanan, and R. Chellappa, "Metareg: Towards domain generalization using meta-regularization," *NeurIPS*, 2018.
- [37] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv:1710.09412*, 2017.
- [38] M. Xu, J. Zhang, B. Ni, T. Li, C. Wang, Q. Tian, and W. Zhang, "Adversarial domain adaptation with domain mixup," in *AAAI Conference on Artificial Intelligence*, 2020.
- [39] S. Yan, H. Song, N. Li, L. Zou, and L. Ren, "Improve unsupervised domain adaptation with mixup training," *arXiv:2001.00677*, 2020.
- [40] Y. Wang, H. Li, and A. C. Kot, "Heterogeneous domain generalization via domain mixup," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.
- [41] H. Nam, H. Lee, J. Park, W. Yoon, and D. Yoo, "Reducing domain gap via style-agnostic networks," *arXiv e-prints*, 2019.
- [42] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv:1412.6572*, 2014.
- [43] A. Sinha, H. Namkoong, R. Volpi, and J. Duchi, "Certifying some distributional robustness with principled adversarial training," *arXiv:1710.10571*, 2017.
- [44] R. Volpi, H. Namkoong, O. Sener, J. Duchi, V. Murino, and S. Savarese, "Generalizing to unseen domains via adversarial data augmentation," *arXiv:1805.12018*, 2018.
- [45] F. Qiao, L. Zhao, and X. Peng, "Learning to learn single domain generalization," in *IEEE/CVF CVPR*, 2020.
- [46] S. Shankar, V. Piratla, S. Chakrabarti, S. Chaudhuri, P. Jyothi, and S. Sarawagi, "Generalizing across domains via cross-gradient training," *arXiv:1804.10745*, 2018.
- [47] M. M. Rahman, C. Fookes, M. Baktashmotlagh, and S. Sridharan, "Multi-component image translation for deep domain generalization," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019.
- [48] N. Somavarapu, C.-Y. Ma, and Z. Kira, "Frustratingly simple domain generalization via image stylization," *arXiv:2006.11207*, 2020.
- [49] F. C. Borlino, A. D'Innocente, and T. Tommasi, "Rethinking domain generalization baselines," in *International Conference on Pattern Recognition (ICPR)*, 2021.
- [50] F. Maria Carlucci, P. Russo, T. Tommasi, and B. Caputo, "Hallucinating agnostic images to generalize across domains," in *IEEE/CVF ICCV Workshops*, 2019.
- [51] K. Zhou, Y. Yang, T. Hospedales, and T. Xiang, "Deep domain-adversarial image generation for domain generalisation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 13 025–13 032.
- [52] —, "Learning to generate novel domains for domain generalization," in *European conference on computer vision*. Springer, 2020, pp. 561–578.
- [53] Z. Huang, H. Wang, E. P. Xing, and D. Huang, "Self-challenging improves cross-domain generalization," in *Computer Vision—ECCV 2020: 16th European Conference*, 2020.
- [54] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang, "Domain generalization with mixstyle," *arXiv:2104.02008*, 2021.

- [55] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv:1409.1556*, 2014.
- [56] K. Schulz, L. Sixt, F. Tombari, and T. Landgraf, "Restricting the flow: Information bottlenecks for attribution," *arXiv:2001.00396*, 2020.
- [57] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "Smoothgrad: removing noise by adding noise," *arXiv:1706.03825*, 2017.
- [58] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *ICML*, 2017.
- [59] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv:1412.6806*, 2014.
- [60] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS one*, 2015.
- [61] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, "Explaining nonlinear classification decisions with deep taylor decomposition," *Pattern Recognition*, 2017.
- [62] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *IEEE CVPR*, 2016.
- [63] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *JMLR*, 2014.
- [64] S. van Steenkiste, F. Locatello, J. Schmidhuber, and O. Bachem, "Are disentangled representations helpful for abstract visual reasoning?" *arXiv:1905.12506*, 2019.
- [65] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980*, 2014.
- [66] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE CVPR*, 2016.
- [67] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, 2015.
- [68] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *ICML*, 2015.
- [69] S. Seo, Y. Suh, D. Kim, G. Kim, J. Han, and B. Han, "Learning to optimize domain specific normalization for domain generalization," in *ECCV*, 2020.
- [70] C. Fang, Y. Xu, and D. N. Rockmore, "Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias," in *IEEE ICCV*, 2013.
- [71] H. Venkateswara, J. Eusebio, S. Chakraborty, and S. Panchanathan, "Deep hashing network for unsupervised domain adaptation," in *IEEE CVPR*, 2017.
- [72] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Deeper, broader and artier domain generalization," in *IEEE ICCV*, 2017.
- [73] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, pp. 8026–8037, 2019.
- [74] P. Moritz, R. Nishihara, S. Wang, A. Tumanov, R. Liaw, E. Liang, M. Elibol, Z. Yang, W. Paul, M. I. Jordan *et al.*, "Ray: A distributed framework for emerging {AI} applications," in *13th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 18)*, 2018, pp. 561–577.
- [75] R. Liaw, E. Liang, R. Nishihara, P. Moritz, J. E. Gonzalez, and I. Stoica, "Tune: A research platform for distributed model selection and training," *arXiv preprint arXiv:1807.05118*, 2018.