Towards Data-Free Domain Generalization

Ahmed Frikha[*]   AHMED.FRIKHA@SIEMENS.COM
Siemens Technology and University of Munich

Haokun Chen[*]   HAOKUN.CHEN@SIEMENS.COM
Siemens Technology and University of Munich

Denis Krompaß   DENIS.KROMPASS@SIEMENS.COM
Siemens Technology

Thomas Runkler   THOMAS.RUNKLER@SIEMENS.COM
Siemens Technology and Technical University of Munich

Volker Tresp   VOLKER.TRESP@SIEMENS.COM
Siemens Technology and University of Munich

Abstract

In this work, we investigate the unexplored intersection of domain generalization (DG) and data-free learning. In particular, we address the question: How can knowledge contained in models trained on different source domains be merged into a single model that generalizes well to unseen target domains, in the absence of source and target domain data? Machine learning models that can cope with domain shift are essential for real-world scenarios with often changing data distributions. Prior DG methods typically rely on using source domain data, making them unsuitable for private decentralized data. We define the novel problem of Data-Free Domain Generalization (DFDG), a practical setting where models trained on the source domains separately are available instead of the original datasets, and investigate how to effectively solve the domain generalization problem in that case. We propose DEKAN, an approach that extracts and fuses domain-specific knowledge from the available teacher models into a student model robust to domain shift. Our empirical evaluation demonstrates the effectiveness of our method which achieves first state-of-the-art results in DFDG by significantly outperforming data-free knowledge distillation and ensemble baselines.

Keywords: Domain Generalization; Privacy-Preserving ML; Data-free Knowledge Distillation

1. Introduction

Deep learning methods have achieved impressive performance in a wide variety of tasks where the data is independent and identically distributed. However, real-world scenarios usually involve a distribution shift between the training data used during development and the test data faced at deployment time. In such situations, deep learning models suffer from a performance degradation and fail to generalize to the out-of-distribution (OOD) data from the target domain (Torralba and Efros, 2011). For instance, this domain shift problem is encountered on MRI data from different clinical centers that use different scanners (Dou et al., 2019). Domain Adaptation (DA) approaches (Wilson and Cook, 2020) assume access to data from the source domain(s) for training as well as target domain data for model adaptation. However, data collection from the target domain can

---

[*]Equal contribution

sometimes be expensive, slow, or infeasible, e.g. self-driving cars have to generalize to a variety of weather conditions (Zhang et al., 2017b) in urban and rural environments from different countries. In this work, we focus on Domain Generalization (DG) (Muandet et al., 2013), where a model trained on multiple source domains is applied without any modification to unseen target domains.

In the last decade, a plethora of DG methods requiring only access to the source domains were proposed (Zhou et al., 2021a). Nevertheless, the assumption that access to source domain data can always be granted does not hold in many cases. For instance, General Data Protection Regulation (GDPR) prohibits the access to confidential information and sensitive data that might identify individuals, e.g. bio-metric data. Likewise, some commercial entities are not willing to share their original data to prevent competitive disadvantage. Furthermore, as datasets get larger, their release, transfer, storage and management can become prohibitively expensive (Lopes et al., 2017). To circumvent the concerns related to releasing the original dataset, the data owners might want to share a model trained on their data instead. In light of increasing data privacy concerns, this alternative has recently enjoyed a surge of interest (Micaelli and Storkey, 2019; Chen et al., 2019; Nayak et al., 2019; Liang et al., 2020; Li et al., 2020; Kundu et al., 2020; Yin et al., 2020; Ahmed et al., 2021).

Although Data-Free Knowledge Distillation (DFKD) methods were developed to transfer knowledge from a teacher model to a student model without access to the original data (Lopes et al., 2017; Micaelli and Storkey, 2019; Chen et al., 2019; Nayak et al., 2019; Yin et al., 2020), only single-teacher scenarios with no domain shift were studied. On the other hand, Source-Free Domain Adaptation (SFDA) approaches were proposed to tackle the domain shift problem setting where one (Liang et al., 2020; Li et al., 2020; Kundu et al., 2020) or multiple (Ahmed et al., 2021) models trained on source domain data are available instead of the original dataset(s). Nonetheless, they require access to data from the target domain. In this work, we investigate the unstudied intersection of Domain Generalization and Data-Free Learning. Data-Free Domain Generalization (DFDG) is a problem setting that assumes only access to models trained on the source domains, without requiring data from source or target domains. Hereby, the goal is to have a single model able to generalize to unseen domains without any modification or data exposure, as it is the case in DG.

Our contribution is threefold: Firstly, we introduce and define the novel and practical DFDG problem setting. Secondly, we tackle it by proposing a first and strong approach that merges the knowledge stored in the domain-specific models via the generation of synthetic data and distills it into a single model. Thirdly, we demonstrate the effectiveness of our method by empirically evaluating it on two DG benchmark datasets.

2. Related Work

To the best of our knowledge, we are the first to address the Data-Free Domain Generalization (DFDG) problem. In the following, we discuss approaches to related problem settings.

2.1. Domain Generalization

Domain Generalization (DG) approaches can be classified into three categories. Domain alignment methods attempt to learn a domain-invariant representation of the data from the source domains by regularizing the learning objective. Variants of such a regularization include the minimization across the source domains of the maximum mean discrepancy criteria (MMD) (Gretton et al., 2012; Li et al., 2018b), the minimization of a distance metric between the domain-specific means (Tzeng et al., 2014) or covariance matrices (Sun and Saenko, 2016), the minimization of a contrastive loss

(Motiian et al., 2017; Kim et al., 2021), or the maximization of loss gradient alignment (Shi et al., 2021; Shahtalebi et al., 2021). Other works use adversarial training with a domain discriminator model (Ganin et al., 2016; Li et al., 2018c) for the same purpose. Another category of works leverages meta-learning techniques, e.g., the bi-level optimization scheme proposed in (Finn et al., 2017), to optimize for adaptation. Even though gradient-based meta-learning methods (Finn et al., 2017) were initially developed to enable few-shot learning, they were leveraged to address a wide variety of problems, such as continual learning (Riemer et al., 2018; Frikha et al.), anomaly detection (Frikha et al., 2021a) and DG. In DG, meta-learning was employed to optimize for quick adaptation to different domains (Li et al., 2018a), to learn how to regularize the output layer (Balaji et al., 2018), and to regularize the embedding space (Dou et al., 2019). Another line of work augment the training data to tackle DG. Hereby, the source domain data is perturbed by computing inter-domain examples (Xu et al., 2020; Yan et al., 2020; Wang et al., 2020b) via Mixup (Zhang et al., 2017a), by randomizing the style of images (Nam et al., 2019), by computing adversarial examples (Goodfellow et al., 2014) using a class classifier (Sinha et al., 2017; Qiao et al., 2020) or a domain classifier (Shankar et al., 2018), or corrupting learned features to incentivize new feature discovery (Frikha et al., 2021b). Other works perturb intermediate representations of the data (Huang et al., 2020; Zhou et al., 2021b; Frikha et al., 2021b). Unlike standard DG approaches that require access to the source domain datasets, our method merges the domain-specific knowledge from models trained on the source domains into a single model resilient to domain shift, while preserving data privacy (Figure 1).

2.2. Knowledge Distillation

Knowledge distillation (KD) (Hinton et al., 2015) was originally proposed to compress the knowledge of a large teacher network into a smaller student network. Several KD extensions and improvements enabled its application to a variety of scenarios including quantization (Mishra and Marr, 2017), domain adaptation (Zhao et al., 2020), and few-shot learning (Rajasegaran et al., 2020). While these methods rely on the original data, Data-Free Knowledge Distillation (DFKD) methods were recently developed (Lopes et al., 2017; Micaelli and Storkey, 2019; Nayak et al., 2019; Chen et al., 2019; Liu et al., 2021b). Hereby, knowledge is transferred from one (Micaelli and Storkey, 2019; Nayak et al., 2019; Chen et al., 2019; Choi et al., 2020; Yin et al., 2020; Luo et al., 2020; Zhang et al., 2021b) or multiple (Li et al., 2021b) teacher(s) to the student model via the generation of synthetic data, either by optimizing random noise examples (Nayak et al., 2019; Yin et al., 2020; Zhang et al., 2021b) or by training a generator (Micaelli and Storkey, 2019; Chen et al., 2019; Choi et al., 2020; Luo et al., 2020). Nevertheless, the aforementioned DFKD methods focus on scenarios without any domain shift, i.e. the student is evaluated on examples from the same data distribution used for training the teacher. In the DFDG problem setting we address, the student is trained from multiple teachers that are trained on different source domains in a way that enables generalization to data from unseen target domains. We propose a baseline that extends the usage of a recent DFKD method (Zhang et al., 2021b) to the DFDG setting, and compare it to our approach (Section 4).

2.3. Source-free domain adaptation

The recently addressed Source-Free Domain Adaptation problem (Liang et al., 2020; Li et al., 2020; Kundu et al., 2020) assumes access to one or multiple model(s) trained on the source domains, as well as data examples from a specific target domain. Proposed approaches to tackle it include the

combination of generative models with a regularization loss (Li et al., 2020), a feature alignment mechanism (Yeh et al., 2021), or a weighting of the target domain samples by their similarity to the source domain (Kundu et al., 2020). SHOT (Liang et al., 2020) employs an information maximization loss along with a self-supervised pseudo-labeling, and is extended to the multi-source scenario via source model weighting (Ahmed et al., 2021). BUFR (Eastwood et al., 2021) aligns the target domain feature distribution with the one from the source domain. Another line of works leverage Batch Normalization (BN) (Ioffe and Szegedy, 2015) layers by replacing the BN-statistics computed on the source domain with those computed on the target domain (Li et al., 2016), or by training the BN-parameters on the target domain via entropy minimization (Wang et al., 2020a). While these approaches rely on the availability of data from a known target domain, we address the DFDG scenario where the model is expected to generalize to *a priori unknown* target domain(s) without any modification or exposure to their data. We also note that some methods (Kundu et al., 2020; Liang et al., 2020; Eastwood et al., 2021) modify the training procedure on the source domain, which would not be possible in cases where the data is not accessible anymore.

2.4. Federated Learning

Federated Learning (FL) (McMahan et al., 2017) is a decentralized learning paradigm where multiple clients train deep learning models without centralizing their local data. While most FL works (Yin et al., 2021) focus on single-domain applications, i.e., the clients have the same data distribution, concurrent works addressed FL settings involving distribution shift (Zhang et al., 2021a; Liu et al., 2021a; Li et al., 2021a; Chen et al., 2022). We propose DFDG as an alternative privacy-preserving DG setting, which enables cross-domain knowledge sharing in a static, permanent and communication-efficient way, i.e., no client interaction overhead. Moreover, in scenarios where the data cannot be accessed, e.g., it was lost or deleted, or only a trained model was made public, FL is not applicable as it requires access to the data. However, DFDG methods would be applicable.

3. Approach

3.1. Problem statement

Let D_s and D_t^j denote the datasets available from the source and target domains respectively with $i = 1, .., I$ and $j = 1, .., J$. Hereby, I and J denote the number of source and target domains respectively. In Domain Generalization (DG), the goal is to train a model on the source domain data D_s^i in a way that enables generalization to *a priori* unavailable target domain data D_t^j, without any model modification at test time. We consider the source-data-free scenario of this problem where the source domain datasets D_s^i are not accessible, e.g., due to privacy, security, safety or commercial concerns, and models trained on these domain-specific datasets separately are available instead.

We refer to the source domain models as teacher models T_i as in the knowledge distillation literature (Hinton et al., 2015). We assume that the teacher models were trained without the prior knowledge that they would be used in a DFDG setting, i.e., their training does not involve any domain shift robustness mechanism. Hence, the application scenarios where the source domain data is not accessible anymore, e.g., was deleted, are also considered. We refer to this novel learning scenario as *Data-Free Domain Generalization (DFDG)*. The major difference with Source-Free Domain Adaptation (SFDA) (Liang et al., 2020; Li et al., 2020; Kundu et al., 2020) is the absence of target domain data D_t^j available for training in DFDG (Figure 1).
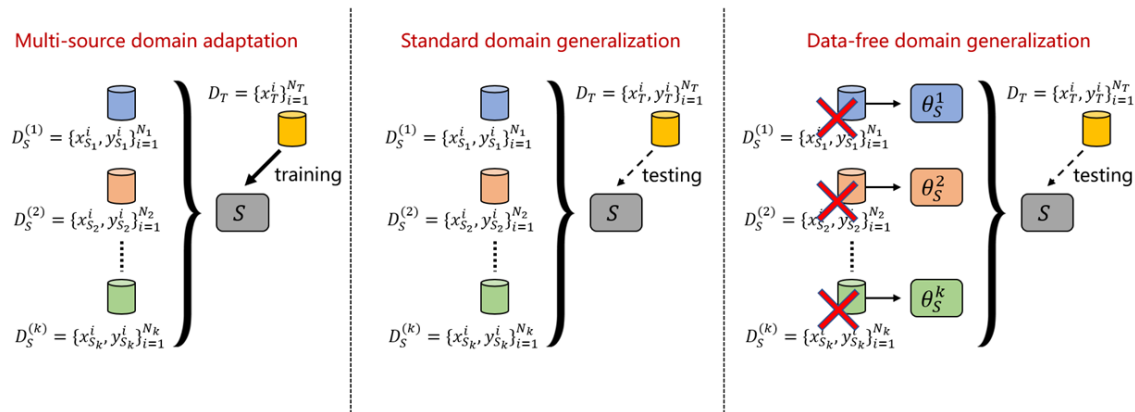
Figure 1: Comparison between the domain generalization, the multi-source domain adaptation and the proposed data-free domain generalization problems.

The DFDG problem is a prototype for a practical use case where a model robust to domain shift is needed and models trained on the same task but different domains are available. This problem definition is motivated by the question: How can we merge the knowledge from multiple models trained on different domains into a single model that is able to generalize to unseen target domains without any data exposure?

Applications of the proposed DFDG include every DG application that involves inaccessible data, e.g., due to privacy concerns, every DFKD application (see Table 4 in Liu et al. (2021b) for an overview) that involves domain shift, and every SFDA application with data-scarce or *a priori* unknown target domains. DFDG addresses all real-world scenarios where different entities have the same application with different data distributions, are unwilling (or not allowed) to share their original data, and would benefit from a model robust to domain shift. In the healthcare sector, the entities could be clinical centers (CCs) using different scanners or data acquisition protocols, such as in (Dou et al., 2019). In the latter, the MRI datasets are not made public, probably due to patient privacy concerns. With DFDG promoting a privacy-preserving way of data sharing, the CCs might consider releasing models. Analogously, in industrial manufacturing, companies are usually unwilling to share their raw production data, due to concerns of intellectual property infringement and reverse engineering. In this sector, examples of DFDG applications include image classification of gas turbine deficiencies and powder bed anomalies in additive manufacturing. Hereby, the images are taken in diverse production environments and conditions specific to the data owner. Note that a model robust to domain shift would be desirable for the machine users, i.e., data owners, and/or for the machine manufacturer (of the gas turbine or additive manufacturing machine).

3.2. DEKAN: Domain Entanglement via Knowledge Amalgamation from Domain-Specific Networks

To address the DFDG problem, we propose Domain Entanglement via Knowledge Amalgamation from domain-specific Networks (DEKAN). Our approach tackles the different challenges of DFDG in 3 stages: Knowledge extraction, fusion and transfer. In the first stage, we extract the knowledge from the different source domain teacher models separately by generating domain-specific syn-

thetic datasets. Thereafter, DEKAN generates cross-domain synthetic data by leveraging all pairs of inter-domain model-dataset combinations. Hereby, the cross-domain examples are optimized to be recognizable by teacher models trained on different domains. In the final stage, DEKAN transfers the extracted knowledge from the domain-specific teachers to a student model via knowledge distillation using the generated data. At test time, the resulting student model is evaluated on target domain data without any modification. In the following, we introduce the stages in more detail.

3.2.1. INTRA-DOMAIN DATA-FREE KNOWLEDGE EXTRACTION

In this stage, we extract the domain-specific knowledge from the available teacher models T_i separately by generating domain-specific synthetic datasets D_g^i. For this, we apply an improved version (Zhang et al., 2021b) of the data-free knowledge distillation method DeepInversion (DI) (Yin et al., 2020) that enables the generation of more diverse images. Hereby, we use inceptionism-style (Mahendran and Vedaldi, 2015) image synthesis, also called DeepDream or inversion, i.e., we initialize random noise images $\hat{x}$ and optimize them to be recognized as samples from pre-defined classes by a trained model. We uniformly sample labels y and optimize the corresponding random images $\hat{x}$ by minimizing the domain-specific inversion loss L_{DS} given by

$$L_{DS} = L_C(T(\hat{x}), y) + \lambda_1 L_R(\hat{x}) + \lambda_2 L_M(\hat{x}), \tag{1}$$

where L_C denotes the classification loss, e.g., cross-entropy, L_R an image prior regularization, L_M a feature moment matching loss, and λ_1 and λ_2 weighting coefficients. L_R penalizes the l_2-norm and the total variation of the image to ensure the convergence to valid natural images (Mahendran and Vedaldi, 2015; Yin et al., 2020). L_M, also called moment matching loss, optimizes the synthetic images so that their feature distributions captured by batch normalization (BN) layers match those of the real data used to train the teacher model. Formally,

$$L_M(\hat{x}) = \sum_l max(\|\mu_l(\hat{x}) - \hat{\mu}_l\|_2 - \delta_l, 0) + \sum_l max(\|\sigma_l^2(\hat{x}) - \hat{\sigma}_l^2\|_2 - \gamma_l, 0). \tag{2}$$

L_M minimizes the l_2-norm between the BN-statistics of the synthetic data, i.e., mean $\mu_l(\hat{x})$ and variance $\sigma_l^2(\hat{x})$, and those stored in the trained teacher model, $\hat{\mu}_l$ and $\hat{\sigma}_l^2$, at each BN layer l (Yin et al., 2020). In order to increase the diversity of the generated images, we relax this optimization by allowing the BN-statistics computed on the synthetic images to deviate from those stored in the model within certain margins, as introduced in (Zhang et al., 2021b). These deviation margins are defined by relaxation constants for mean and variance, denoted by δ_l and γ_l respectively. The latter are computed as the ϵ_{DS} percentile of the distribution of differences between the stored BN-statistics and those computed using random images, as proposed in (Zhang et al., 2021b). We note that the higher the value of the hyperparameter ϵ_{DS}, the higher the relaxation.

We apply this data-free inversion to each model T_i separately, yielding datasets D_g^i that are correctly classified by the respective model and match the distribution of features extracted by it.

3.2.2. CROSS-DOMAIN DATA-FREE KNOWLEDGE FUSION

In the second stage, we propose a technique to merge the knowledge from two domains by generating cross-domain synthetic images that capture class-discriminative features present in the two domains, and match the distribution of intermediate features extracted by a domain-specific model from images of another domain. Let T_a and T_b denote the teacher models, and D_g^a and D_g^b the
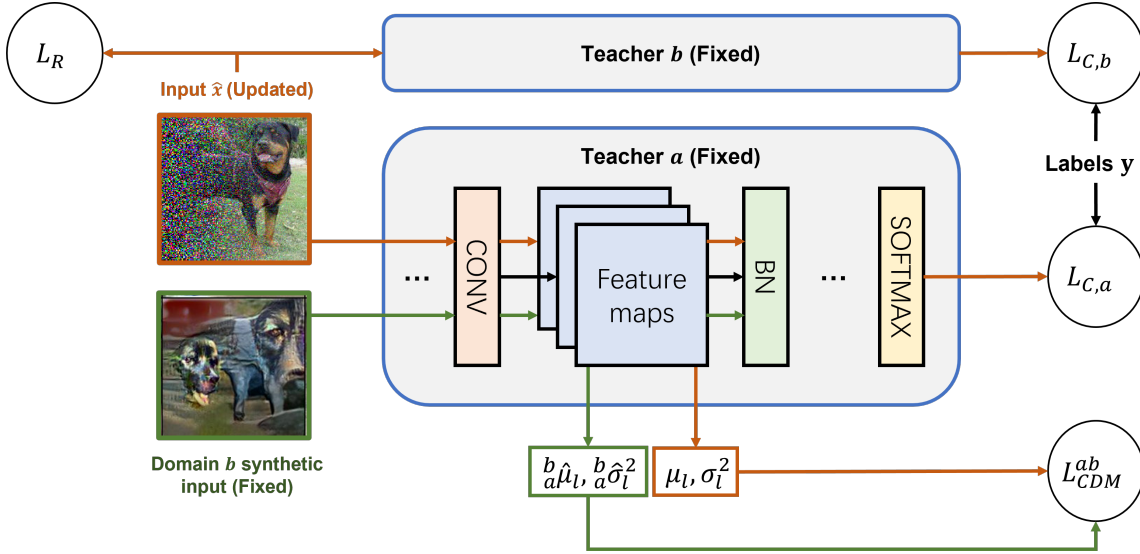
Figure 2: Overview of the Cross-Domain Data-Free Knowledge Fusion.

synthetic data generated in the first stage, specific to two domains a and b. We generate synthetic images D_g^{ab} by minimizing the cross-domain inversion loss L_{CD}^{ab}, that we formulate as

$$L_{CD}^{ab} = L_C(T_a(\hat{x}), y) + L_C(T_b(\hat{x}), y) + \alpha_1 L_R(\hat{x}) + \alpha_2 L_{CDM}^{ab}(\hat{x}), \tag{3}$$

where L_C denotes the classification loss, e.g., cross-entropy, L_R the aforementioned image prior regularization, L_{CDM} the cross-domain feature moment matching loss, and α_1 and α_2 weighting coefficients. We incentivize the generated images to contain class-discriminative features from both domains by minimizing the classification loss using both teachers. We hypothesize that images that can be recognized by models trained on different domains capture more domain-agnostic semantic features than those generated by inverting a single domain-specific model as done in prior works. In addition, the cross-domain feature distribution matching loss L_{CDM}^{ab} optimizes the cross-domain synthetic images D_g^{ab} so that their feature distribution matches the distribution of the features extracted by T_a, the model trained on domain a, for images D_g^b synthesized from domain b. Note that $L_{CDM}^{ab} \neq L_{CDM}^{ba}$ and that using the model T_b and the data generated by inverting T_a in the first stage, i.e., D_g^a, would yield the cross-domain images D_g^{ba} that are different from D_g^{ab}. Formally,

$$L_{CDM}^{ab}(\hat{x}) = \sum_l max(\|\mu_l(\hat{x}) - {}_a^b\hat{\mu}_l\|_2 - {}_a^b\delta_l, 0) + \sum_l max(\|\sigma_l^2(\hat{x}) - {}_a^b\hat{\sigma}_l^2\|_2 - {}_a^b\gamma_l, 0). \tag{4}$$

Similarly to L_M (Eq. 2) in the first stage, L_{CDM}^{ab} minimizes the l_2-norm between the BN-statistics of the synthetic data, $\mu_l(\hat{x})$ and $\sigma_l^2(\hat{x})$, and target statistics, at each BN layer l. Here, the target statistics, ${}_a^b\hat{\mu}_l$ and ${}_a^b\hat{\sigma}_l^2$, are computed in a way that involves knowledge from different domains. In particular, they result from feeding the synthetic data specific to domain b to the teacher trained on data from domain a, and computing the first two feature moments, i.e., mean and variance, for each BN layer. The intention behind this is to synthesize images that capture the features learned by the model on domain a that are activated and recognized when exposed to images from domain

b. We hypothesize that such images would encompass domain-agnostic semantic information that would be useful for training a single model resilient to domain shift in the next stage.

We relax L_{CDM} by allowing the BN-statistics of the synthetic input to fluctuate within a certain interval. Here, we compute the relaxation constants ${}^b_a\delta_l$ and ${}^b_a\gamma_l$ as the ϵ_{CD} percentile of the distribution of differences between the stored BN-statistics, i.e., computed on the original domain a images, and those computed using the images D^b_g synthesized from the domain b teacher in the first stage. $\epsilon_{CD} = 100\%$ corresponds to synthesized images $\hat{x}$ yielding the BN-statistics from domain a, i.e., stored in model T_a, would not be penalized, i.e., $L^{ab}_{CDM} = 0$. This stage can be viewed as a domain augmentation, since the synthesized images D^{ab}_g do not belong neither to domain a nor to domain b. The synthesis of cross-domain data is applied to all possible domain pairs.

Algorithm 1 Domain Entanglement via Knowledge Amalgamation from domain-specific Networks

---

Require: $T_{1..I}$: I Domain-specific teacher models
    // First stage: Intra-Domain Knowledge Extraction
  1: **for** $i \leftarrow 1$ to I **do**
  2:     Initialize the domain-specific synthetic data D^i_g: Images $\hat{x} \sim \mathcal{N}(0, I)$ and arbitrary labels
  3:     **while** not converged **do**
  4:         Update D^i_g by minimizing the domains-specific inversion loss L_{DS} (Eq. 1) using T_i
  5:     **end while**
  6: **end for**
    // Second stage: Cross-Domain Knowledge Fusion
  7: **for** $i \leftarrow 1$ to I **do**
  8:     **for** $j \leftarrow 1$ to I and $i \neq j$ **do**
  9:         Initialize the cross-domain synthetic data D^{ij}_g: Images $\hat{x} \sim \mathcal{N}(0, I)$ and arbitrary labels
10:         **while** not converged **do**
11:            Update D^{ij}_g by minimizing the cross-domain loss L^{ij}_{CD} (Eq. 3) using T_i, T_j and D^j_g
12:         **end while**
13:     **end for**
14: **end for**
    // Third stage: Multi-Domain Knowledge Distillation
15: Initialize the student model S_θ randomly or from a pre-trained model
16: Concatenate the domain-specific and cross-domain synthetic datasets into one dataset D_g
17: **while** not converged **do**
18:     Randomly sample a mini-batch $B = \{\hat{x}, y\}$ from D_g
19:     Update θ by minimizing the knowledge distillation loss L_{KD} (Eq. 5) using B and $T_{1..I}$
20: **end while**
21: **return** Domain-generalized student model S_θ

3.2.3. MULTI-DOMAIN KNOWLEDGE DISTILLATION

In the final DEKAN stage, the domain-specific and cross-domain knowledge, which is captured in the synthetic data generated in the first and second stages respectively, is transferred to a single student model S. To this end, we use knowledge distillation (Hinton et al., 2015), i.e., we train the student model to mimic the predictions of the teachers for the synthetic data.

$$L_{KD} = D_{KL}(S(\hat{x}) \,||\, p) \text{ with } p = \begin{cases} T_i(\hat{x}), & \text{if } \hat{x} \in D_g^i \text{ (domain-specific)} \\ \frac{1}{2}(T_i(\hat{x}) + T_j(\hat{x})), & \text{if } \hat{x} \in D_g^{ij} \text{ (cross-domain)} \end{cases} \quad (5)$$

As described in Equation 5, we minimize the Kullback-Leibler divergence D_{KL} between the predictions of the student S and the teacher(s) corresponding to the synthetic images $\hat{x}$. In particular, if the data examples are domain-specific, i.e., they were generated in the first DEKAN stage, the predictions of the corresponding teacher are used as soft labels to train the student. For the cross-domain synthetic images that were generated in the second stage, the average predictions of the two corresponding teachers is used instead. The aggregation of the prediction distributions of two domain-specific teacher models contributes to the knowledge amalgamation across domains.

Algorithm 1 summarizes the 3 stages of the DEKAN's training procedure. We note that the updates of the syntehtic data and the student model parameters θ are performed using gradient-based optimization, specifically Adam (Kingma and Ba, 2014) in our case. Explicit update rule formulas and iteration over the synthetic data batches are omitted for simplicity of notation.

4. Experiments and Results

The conducted experiments aim to tackle the following questions: (a) How does DEKAN compare to leveraging the domain specific models directly to make predictions on data from unseen domains? (b) How does our approach compare to data-free knowledge distillation methods applied to each domain separately? (c) How much does the unavailability of data cost in terms of performance?

We design baseline methods to address the novel DFDG problem, and compare them with DEKAN. The first category of baselines applies the available domain-specific models on the data from the target domains (Question (a)). We consider two ensemble baselines that aggregate the predictions of these models, e.g., by taking the average of the model predictions (**AvgPred**), or by taking the prediction of the most confident model, i.e., the model with the lowest entropy (**HighestConf**). Besides, we implement oracle methods that evaluate each of the domain-specific models separately on the target domain and then report the results of the best model (**BestTeacher**). Furthermore, we propose a baseline that applies an improved version (Zhang et al., 2021b) of Deep-Inversion (DI) (Yin et al., 2020) on each of the models separately to generate domain-specific synthetic images used to then train a student model via knowledge distillation (**Multi-DI**; Question (b)). Note that Multi-DI is equivalent to the application of DEKAN's first and third stage. Finally, we compare DEKAN to an upper-bound baseline that uses the original data from the source domains to train a single model via Empirical Risk Minimization (**ERM**) (Gulrajani and Lopez-Paz, 2020), a common domain generalization baseline (Question (c)).

We evaluate DEKAN and the baselines on two DG benchmark datasets. PACS (Li et al., 2017) includes images that belong to 7 classes from the domains Art-Painting, Cartoon, Photo and Sketch. Digits comprises images four digits datasets: MNIST (LeCun et al., 1998), MNIST-M (Ganin and Lempitsky, 2015), SVHN (Netzer et al., 2011) and USPS (Hull, 1994). We use the same model architecture for the teacher and the student models: ResNet-50 (He et al., 2016) and ResNet-18 pretrained on ImageNet (Russakovsky et al., 2015), for PACS and Digits respectively. In both synthetic input generation stages, we augment the optimized images before feeding them to the teacher(s). In particular, we apply random horizontal flipping and jitter as done in (Yin et al., 2020), as well as cutout (DeVries and Taylor, 2017), which was found to be essential for data-free object

Algorithm	Art	Cartoon	Photo	Sketch	Avg
Ens-AvgPred	79.9	65.4	96.4	79.5	80.3
Ens-HighestConf	82.3	66.0	96.6	76.9	80.4
Multi-DI	82.1	72.1	95.6	73.8	80.9
DEKAN (ours)	83.0	75.9	96.3	80.2	**83.9**
BestTeacher (oracle)	75.2	62.8	96.4	69.8	76.1
ERM (not data-free)	86.0	81.8	96.8	80.4	86.2

Algorithm	MNIST	M-M	SVHN	USPS	Avg
Ens-AvgPred	97.9	45.8	31.3	96.1	67.8
Ens-HighestConf	98.5	46.7	30.5	96.5	68.0
Multi-DI	93.3	54.0	36.7	96.5	70.2
DEKAN (ours)	94.6	55.9	39.2	96.8	**71.6**
BestTeacher (oracle)	99.3	48.3	38.1	97.7	70.9
ERM (not data-free)	98.2	55.2	50.1	96.5	75.0

Table 1: Domain Generalization results on PACS (top) and Digits (bottom).

detection (Chawla et al., 2021). For the Digits dataset, random horizontal flipping was not used, as it leads to images of invalid digits. Table 1 shows the results of DEKAN and the baselines. Hereby, the column name refers to the unseen target domain, i.e., the 3 other domains are the source domains used to train the teacher models. The test accuracy is computed on the test set of the target domain.

DEKAN outperforms all data-free baselines on both datasets on average, setting a first state-of-the-art performance for the novel DFDG problem. We find that generative approaches, i.e., Multi-DI and DEKAN, outperform the ensemble methods on average, suggesting that training a single model on data from different domains enables a better aggregation of knowledge than the aggregation of domain-specific model predictions. Most importantly, DEKAN substantially outperforms Multi-DI, highlighting the importance of the synthesized cross-domain images. This is especially the case for the challenging domains, i.e., the domains where all the methods yield the lowest performance. In particular, the generation of cross-domain synthetic data leads to performance improvements of 6.4% and 3.8% on the Sketch and Cartoon PACS domains respectively, as well as a 2.4% increase on the SVHN domain of Digits. Additionally, we note the positive knowledge transfer across domains on the PACS dataset, as all the multi-domain methods outperform the oracle BestTeacher baseline that uses a single domain-specific teacher model, i.e., the teacher that achieves the highest performance on a validation set from the target domain. Finally, it is worth noting that while DEKAN significantly reduces the gap between the best data-free baseline and the upper-bound baseline that uses the original data, there is still potential for improvement.

We use UMAP (McInnes et al., 2018) to analyze the embeddings extracted by the student model trained by DEKAN (Figure 3). Despite being trained exclusively on synthetic data, the student model achieves a good class separation of the embeddings extracted for the original PACS dataset (Figure 3 (a)). This indicates that the generated data captures class-discriminative features useful for the classification of original examples. We also observe a good separability for most classes
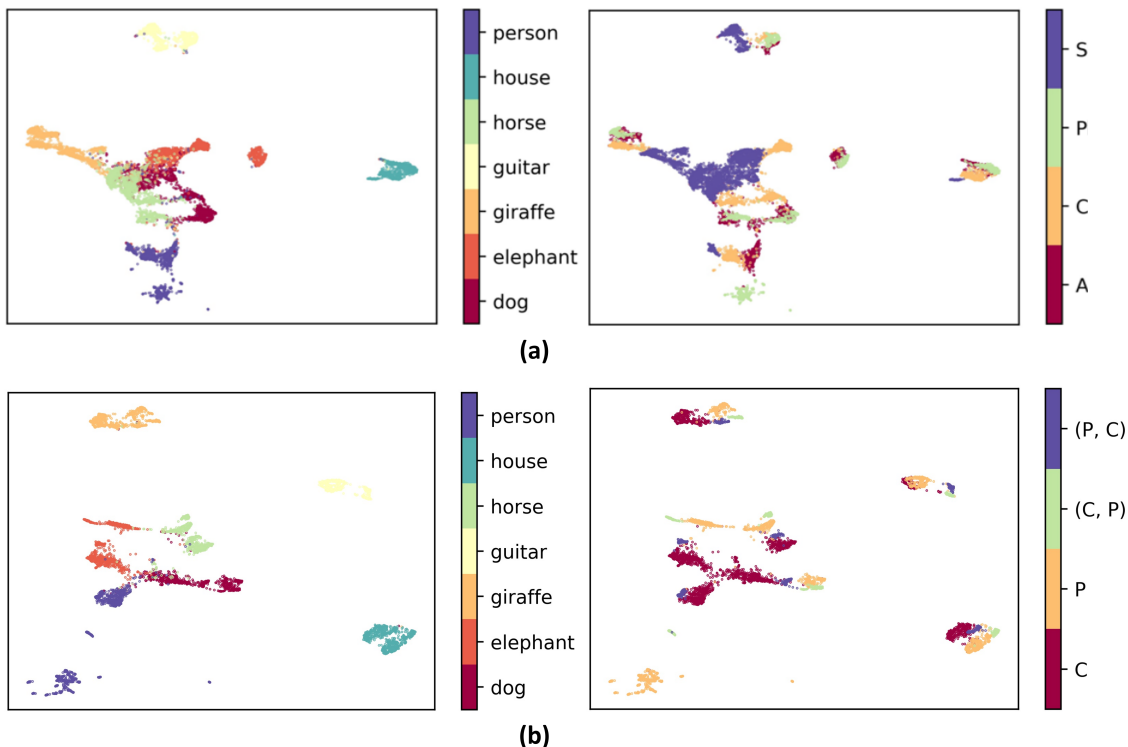
Figure 3: UMAP visualizations of the representations extracted by the last ResNet block, i.e., the input of the linear classifier, of student models trained by DEKAN. Different colors indicate different classes or different domains (S=Sketch, P=Photo, C=Cartoon, A=Art painting). (a) Embeddings of the *original* PACS dataset extracted by the DEKAN-trained student in the setting where the target domain is Sketch. Note that the student is trained exclusively with *synthetic* examples. (b) Embeddings of the original examples from the Photo and Cartoon domains as well as the generated cross-domain images extracted by the DEKAN-trained student in the setting where the target domain is Art painting. (P, C) and (C, P) denote the generated cross-domain synthetic datasets D_g^{PC} and D_g^{CP}, respectively.

of the target domain (Sketch), which is unseen during training. This emphasizes the high domain generalization ability of the student trained with DEKAN, without any access to original data.

Finally, we investigate the embeddings extracted for the generated cross-domain examples by the student model (Figure 3 (b)). We find that the embeddings extracted for the generated cross-domain examples lie between the embeddings of the original domain-specific examples, for most classes, e.g., house, person and dog. We hypothesize that the substantial increase in DG performance between DEKAN and Multi-DI (Table 1) is due to the fact that the generated cross-domain images bridge the gap between the source domains in the student's embedding space.

5. Conclusion

This work addresses the unstudied intersection of domain generalization and data-free learning, a practical setting where a model robust to domain shift is needed and the available models were trained on the same task but with data from different domains. We proposed DEKAN, an approach

that fuses domain-specific knowledge from the available teacher models into a single student model that can generalize to data from *a priori* unknown domains. Our empirical evaluation demonstrated the effectiveness of our method which outperformed ensemble and data-free knowledge distillation baselines, hence achieving first state-of-the-art results in the novel and challenging data-free domain generalization problem. An interesting avenue for future works could be the integration of differential privacy mechanism into DEKAN.

References

Sk Miraj Ahmed, Dripta S Raychaudhuri, Sujoy Paul, Samet Oymak, and Amit K Roy-Chowdhury. Unsupervised multi-source domain adaptation without access to source data. In *CVPR*, 2021.

Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Towards domain generalization using meta-regularization. *NeurIPS*, 2018.

Akshay Chawla, Hongxu Yin, Pavlo Molchanov, and Jose Alvarez. Data-free knowledge distillation for object detection. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021.

Hanting Chen, Yunhe Wang, Chang Xu, Zhaohui Yang, Chuanjian Liu, Boxin Shi, Chunjing Xu, Chao Xu, and Qi Tian. Data-free learning of student networks. In *ICCV*, 2019.

Haokun Chen, Ahmed Frikha, Denis Krompass, and Volker Tresp. Fraug: Tackling federated learning with non-iid features via representation augmentation. *arXiv preprint arXiv:2205.14900*, 2022.

Yoojin Choi, Jihwan Choi, Mostafa El-Khamy, and Jungwon Lee. Data-free network quantization with adversarial knowledge distillation. In *CVPR Workshops*, 2020.

Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *Arxiv*, 2017.

Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. *NeurIPS*, 2019.

Cian Eastwood, Ian Mason, Christopher KI Williams, and Bernhard Schölkopf. Source-free adaptation to measurement shift via bottom-up feature restoration. *Arxiv*, 2021.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017.

Ahmed Frikha, Denis Krompaß, and Volker Tresp. Arcade: A rapid continual anomaly detector. In *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE.

Ahmed Frikha, Denis Krompaß, Hans-Georg Köpken, and Volker Tresp. Few-shot one-class classification via meta-learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021a.

Ahmed Frikha, Denis Krompaß, and Volker Tresp. Columbus: Automated discovery of new multi-level features for domain generalization via knowledge corruption. *Arxiv*, 2021b.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2015.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 2016.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *Arxiv*, 2014.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *JMLR*, 2012.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *Arxiv*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *Arxiv*, 2015.

Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *ECCV*, 2020.

Jonathan J. Hull. A database for handwritten text recognition research. *IEEE Transactions on pattern analysis and machine intelligence*, 1994.

Sergey Ioffe and Christian Szegedy. Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.

Daehee Kim, Seunghyun Park, Jinkyu Kim, and Jaekoo Lee. Selfreg: Self-supervised contrastive regularization for domain generalization. *Arxiv*, 2021.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Arxiv*, 2014.

Jogendra Nath Kundu, Naveen Venkat, R Venkatesh Babu, et al. Universal source-free domain adaptation. In *CVPR*, 2020.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *IEEE*, 1998.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *ICCV*, 2017.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI*, 2018a.

Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *CVPR*, 2018b.

Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *CVPR*, 2020.

Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. Fedbn: Federated learning on non-iid features via local batch normalization. *arXiv preprint arXiv:2102.07623*, 2021a.

Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *ECCV*, 2018c.

Yanghao Li, Naiyan Wang, Jianping Shi, Jiaying Liu, and Xiaodi Hou. Revisiting batch normalization for practical domain adaptation. *Arxiv*, 2016.

Yuhang Li, Feng Zhu, Ruihao Gong, Mingzhu Shen, Xin Dong, Fengwei Yu, Shaoqing Lu, and Shi Gu. Mixmix: All you need for data-free compression are feature and data mixing. In *ICCV*, 2021b.

Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *ICML*, 2020.

Quande Liu, Cheng Chen, Jing Qin, Qi Dou, and Pheng-Ann Heng. Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1013–1023, 2021a.

Yuang Liu, Wei Zhang, Jun Wang, and Jianyong Wang. Data-free knowledge transfer: A survey. *ArXiv*, 2021b.

Raphael Gontijo Lopes, Stefano Fenu, and Thad Starner. Data-free knowledge distillation for deep neural networks. *Arxiv*, 2017.

Liangchen Luo, Mark Sandler, Zi Lin, Andrey Zhmoginov, and Andrew Howard. Large-scale generative data-free distillation. *Arxiv*, 2020.

Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *CVPR*, 2015.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *Arxiv*, 2018.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

Paul Micaelli and Amos Storkey. Zero-shot knowledge transfer via adversarial belief matching. *Arxiv*, 2019.

Asit Mishra and Debbie Marr. Apprentice: Using knowledge distillation techniques to improve low-precision network accuracy. *Arxiv*, 2017.

Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *ICCV*, 2017.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *ICML*, 2013.

Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap via style-agnostic networks. *Arxiv*, 2019.

Gaurav Kumar Nayak, Konda Reddy Mopuri, Vaisakh Shaj, Venkatesh Babu Radhakrishnan, and Anirban Chakraborty. Zero-shot knowledge distillation in deep networks. In *ICML*, 2019.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

Fengchun Qiao, Long Zhao, and Xi Peng. Learning to learn single domain generalization. In *CVPR*, 2020.

Jathushan Rajasegaran, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Mubarak Shah. Self-supervised knowledge distillation for few-shot learning. *Arxiv*, 2020.

Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro. Learning to learn without forgetting by maximizing transfer and minimizing interference. *ArXiv*, 2018.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 2015.

Soroosh Shahtalebi, Jean-Christophe Gagnon-Audet, Touraj Laleh, Mojtaba Faramarzi, Kartik Ahuja, and Irina Rish. Sand-mask: An enhanced gradient masking strategy for the discovery of invariances in domain generalization. *Arxiv*, 2021.

Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. *Arxiv*, 2018.

Yuge Shi, Jeffrey Seely, Philip HS Torr, N Siddharth, Awni Hannun, Nicolas Usunier, and Gabriel Synnaeve. Gradient matching for domain generalization. *Arxiv*, 2021.

Aman Sinha, Hongseok Namkoong, Riccardo Volpi, and John Duchi. Certifying some distributional robustness with principled adversarial training. *Arxiv*, 2017.

Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV*, 2016.

Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR*, 2011.

Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *Arxiv*, 2014.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Fully test-time adaptation by entropy minimization. *Arxiv*, 2020a.

Yufei Wang, Haoliang Li, and Alex C Kot. Heterogeneous domain generalization via domain mixup. In *ICASSP*, 2020b.

Garrett Wilson and Diane J Cook. A survey of unsupervised deep domain adaptation. *ACM TIST*, 2020.

Minghao Xu, Jian Zhang, Bingbing Ni, Teng Li, Chengjie Wang, Qi Tian, and Wenjun Zhang. Adversarial domain adaptation with domain mixup. In *AAAI*, 2020.

Shen Yan, Huan Song, Nanxiang Li, Lincan Zou, and Liu Ren. Improve unsupervised domain adaptation with mixup training. *Arxiv*, 2020.

Hao-Wei Yeh, Baoyao Yang, Pong C Yuen, and Tatsuya Harada. Source-data-free feature alignment for unsupervised domain adaptation. In *WACV*, 2021.

Hongxu Yin, Pavlo Molchanov, Jose M Alvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niraj K Jha, and Jan Kautz. Dreaming to distill: Data-free knowledge transfer via deepinversion. In *CVPR*, 2020.

Xuefei Yin, Yanming Zhu, and Jiankun Hu. A comprehensive survey of privacy-preserving federated learning: A taxonomy, review, and future directions. *ACM Computing Surveys (CSUR)*, 54 (6):1–36, 2021.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *Arxiv*, 2017a.

Liling Zhang, Xinyu Lei, Yichun Shi, Hongyu Huang, and Chao Chen. Federated learning with domain generalization. *arXiv preprint arXiv:2111.10487*, 2021a.

Xiangguo Zhang, Haotong Qin, Yifu Ding, Ruihao Gong, Qinghua Yan, Renshuai Tao, Yuhang Li, Fengwei Yu, and Xianglong Liu. Diversifying sample generation for accurate data-free quantization. In *CVPR*, 2021b.

Yang Zhang, Philip David, and Boqing Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. In *ICCV*, 2017b.

Sicheng Zhao, Guangzhi Wang, Shanghang Zhang, Yang Gu, Yaxian Li, Zhichao Song, Pengfei Xu, Runbo Hu, Hua Chai, and Kurt Keutzer. Multi-source distilling domain adaptation. In *AAAI*, 2020.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *Arxiv*, 2021a.

Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. *Arxiv*, 2021b.