Towards Data-Free Domain Generalization

Ahmed Frikha[*,†]
Siemens Technology
University of Munich

Haokun Chen[*,†]
Siemens Technology
Technical University of Munich

Denis Krompaß
Siemens Technology

Thomas Runkler
Siemens Technology
Technical University of Munich

Volker Tresp
Siemens Technology
University of Munich

Abstract

In this work, we investigate the unexplored intersection of domain generalization and data-free learning. In particular, we address the question: How can knowledge contained in models trained on different source data 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 for real-world scenarios with often changing data distributions. Prior domain generalization 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 ensemble and data-free knowledge distillation baselines.

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 often suffer from a performance degradation and fail to generalize to the out-of-distribution (OOD) data from the target domain [62, 66, 17, 21]. For instance, this domain shift problem is encountered when applying deep learning models on MRI data from different clinical centers that use different scanners [10]. Domain Adaptation (DA) approaches [71, 73] 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 sometimes be expensive, slow, or infeasible, e.g. self-driving cars have to generalize to a variety of weather conditions [80] and object poses [3] in urban and rural environments from different countries. In this work, we focus on the Domain Generalization (DG) [5, 48] setting, 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 [86]. Nevertheless, the assumption that access to source domain data can always be granted

[*]Equal contribution.
[†]Primary contact authors: ahmed.frikha@siemens.com and haokun.chen@siemens.com

Workshop on Distribution Shifts, 35th Conference on Neural Information Processing Systems (NeurIPS 2021).

does not hold in many cases. For instance, General Data Protection Regulation (GDPR) prohibits the access to sensitive data that might identify individuals, e.g. bio-metric data or other confidential information. 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 [39]. 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 [44, 7, 50, 37, 33, 28, 78, 1].

Although Data-Free Knowledge Distillation (DFKD) methods were developed to transfer knowledge from a teacher model to a student model without any access to the original data [39, 44, 7, 50, 78, 9], 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 [37, 33, 28, 70, 11] or multiple [1] 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. To the best of our knowledge, we are the first to address this problem setting. Works addressing related problems are discussed in Appendix A.

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 Approach

2.1 Problem statement

Let D_s^i 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 the Domain Generalization (DG) [5, 48] problem setting, 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 [22]. 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) [37, 33, 28] is the absence of target domain data D_t^j in DFDG.

The DFDG problem is a prototype for a practical use case where a model robust to domain shifts is needed and models trained on the same task but different data domains are available. This problem definition is motivated by the question: How can we amalgamate 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?

2.2 Domain Entanglement via Knowledge Amalgamation from Domain-Specific Networks

We propose Domain Entanglement via Knowledge Amalgamation from domain-specific Networks (DEKAN). Our approach tackles the challenges of DFDG in 3 stages: Knowledge extraction, fusion and transfer. In the first stage, *Intra-Domain Data-Free Knowledge Extraction*, we extract the knowledge from the different source domain teacher models separately. Hereby, we generate domain-specific synthetic datasets via inceptionism-style [46] image synthesis, i.e., we initialize random noise images $\hat{x}$ and optimize them to be recognized as a sample from a pre-defined class by a trained domain-specific model. In particular, we apply the data-free knowledge distillation method described in [78, 83] to invert each domain-specific teacher separately. In the second stage, *Cross-Domain Data-Free Knowledge Fusion*, DEKAN generates cross-domain synthetic data by leveraging all pairs of inter-domain model-dataset combinations. Here, the cross-domain examples are optimized to be recognizable by teacher models trained on different domains. In the final stage, *Multi-Domain Knowledge Distillation*, DEKAN transfers the extracted knowledge from the domain-specific teachers to a student model via knowledge distillation using the generated data. At test time, i.e., deployment phase, the resulting student model is evaluated on target domain data without any modification. Details about the first stage as well as DEKAN's complete algorithm can be found in Appendix B. In the following, we focus on the second and third stages.

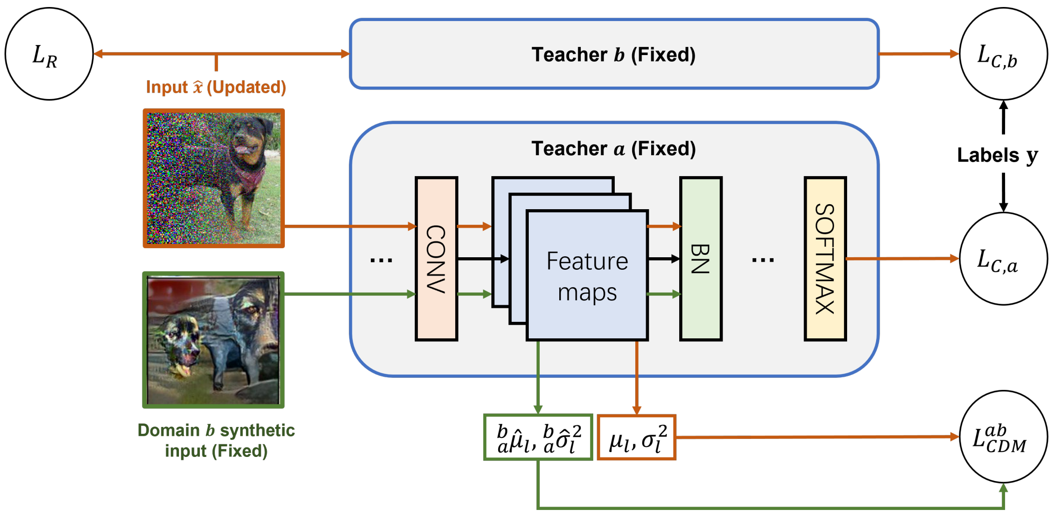

Figure 1: Overview of the Cross-Domain Data-Free Knowledge Fusion.

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 synthetic data generated in the first stage (Appendix B.0.1), 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

$$L_{CD}^{ab} = L_C(T_a(\hat{x}), y) + L_C(T_b(\hat{x}), y) + \alpha_1 L_R(\hat{x}) + \alpha_2 L_{CDM}^{ab}(\hat{x}), \tag{1}$$

where L_C denotes the classification loss, e.g., cross-entropy, L_R an image prior regularization, L_{CDM} the cross-domain feature moment matching loss, and α_1 and α_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 [42, 52, 46, 78]. 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 [78].

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,

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

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 through the teacher model 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 ${}_a^b\delta_l$ and ${}_a^b\gamma_l$ 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_g^b synthesized from the domain b teacher model in the first stage. Note that $\epsilon_{CD} = 100\%$ 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., $L_{CDM}^{ab} = 0$. This stage can be viewed as a domain augmentation, since the synthesized images D_g^{ab} do not belong neither to domain a nor to domain b. The synthesis of cross-domain data is applied to all possible domain pairs.

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 [22], i.e., we train the student model to mimic the predictions of the teachers for the synthetic data. As described in Equation 3, we minimize the Kullback-Leibler divergence D_{KL} between the predictions of the student S and the teacher(s) corresponding to the synthetic image $\hat{x}$. In particular, if the data example is domain-specific, i.e., it was 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.

$$L_{KD} = D_{KL}(S(\hat{x}) \mid\mid p) \quad \text{with} \quad p = \begin{cases} T_i(\hat{x}), & \text{if } \hat{x} \in D_g^i \quad \text{(domain-specific)} \\ \frac{1}{2}(T_i(\hat{x}) + T_j(\hat{x})), & \text{if } \hat{x} \in D_g^{ij} \quad \text{(cross-domain)} \end{cases} \quad (3)$$

3 Experiments and Results

The conducted experiments[3] aim to tackle the following key 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 [83] of DeepInversion (DI) [78] 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

[3]Code will be made public upon paper acceptance.

the original data from the source domains to train a single model via Empirical Risk Minimization (**ERM**) [68, 20], a common domain generalization baseline (Question (c)).

Algorithm	**Art Painting**	**Cartoon**	**Photo**	**Sketch**	**Average**
Ensemble - AvgPred	79.88	65.40	96.35	79.46	80.27
Ensemble - HighestConf	82.28	65.96	96.59	76.86	80.42
Multi-DI	82.13	72.14	95.57	73.75	80.90
DEKAN (ours)	83.01	75.94	96.29	80.17	**83.88**
BestTeacher (oracle)	75.24	62.80	96.41	69.76	76.05
ERM [20] (not data-free)	86.0	81.8	96.8	80.4	86.2

Algorithm	**MNIST**	**MNIST-M**	**SVHN**	**USPS**	**Average**
Ensemble - AvgPred	97.85	45.83	31.33	96.12	67.78
Ensemble - HighestConf	98.52	46.71	30.45	96.47	68.04
Multi-DI	93.31	54.04	36.72	96.53	70.15
DEKAN (ours)	94.64	55.86	39.15	96.77	**71.61**
BestTeacher (oracle)	99.27	48.33	38.11	97.73	70.86
ERM (not data-free)	98.22	55.18	50.13	96.54	75.02

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

We evaluate DEKAN and the baselines on two DG benchmark datasets, PACS [30] and Digits, which comprises images from MNIST [29], MNIST-M [15], SVHN [51] and USPS [24]. 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.

4 Conclusion

This work addressed the unstudied intersection of domain generalization and data-free learning, a practical setting where a model robust to domain shifts 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.

References

[1] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10103–10112, 2021.

[2] Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9163–9171, 2019.

[3] Michael A Alcorn, Qi Li, Zhitao Gong, Chengfei Wang, Long Mai, Wei-Shinn Ku, and Anh Nguyen. Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[4] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. *Advances in Neural Information Processing Systems*, 2018.

[5] Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. *Advances in neural information processing systems*, 2011.

[6] Francesco Cappio Borlino, Antonio D'Innocente, and Tatiana Tommasi. Rethinking domain generalization baselines. In *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021.

[7] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3514–3522, 2019.

[8] Ali Cheraghian, Shafin Rahman, Pengfei Fang, Soumava Kumar Roy, Lars Petersson, and Mehrtash Harandi. Semantic-aware knowledge distillation for few-shot class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2534–2543, 2021.

[9] Yoojin Choi, Jihwan Choi, Mostafa El-Khamy, and Jungwon Lee. Data-free network quantization with adversarial knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 710–711, 2020.

[10] Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. *Advances in Neural Information Processing Systems*, 2019.

[11] Cian Eastwood, Ian Mason, Christopher KI Williams, and Bernhard Schölkopf. Source-free adaptation to measurement shift via bottom-up feature restoration. *arXiv preprint arXiv:2107.05446*, 2021.

[12] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR, 2017.

[13] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1322–1333, 2015.

[14] Ahmed Frikha, Denis Krompaß, and Volker Tresp. Columbus: Automated discovery of new multi-level features for domain generalization via knowledge corruption. *arXiv preprint arXiv:2109.04320*, 2021.

[15] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

[16] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 2016.

[17] Robert Geirhos, Carlos R Medina Temme, Jonas Rauber, Heiko H Schütt, Matthias Bethge, and Felix A Wichmann. Generalisation in humans and deep neural networks. *arXiv preprint arXiv:1808.08750*, 2018.

[18] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[19] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 2012.

[20] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

[21] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

[22] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[23] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *Computer Vision–ECCV 2020: 16th European Conference*, 2020.

[24] Jonathan J. Hull. A database for handwritten text recognition research. *IEEE Transactions on pattern analysis and machine intelligence*, 16(5):550–554, 1994.

[25] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, 2015.

[26] Daehee Kim, Seunghyun Park, Jinkyu Kim, and Jaekoo Lee. Selfreg: Self-supervised contrastive regularization for domain generalization. *arXiv preprint arXiv:2104.09841*, 2021.

[27] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[28] Jogendra Nath Kundu, Naveen Venkat, R Venkatesh Babu, et al. Universal source-free domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4544–4553, 2020.

[29] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

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

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

[32] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[33] Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9641–9650, 2020.

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

[35] Yanghao Li, Naiyan Wang, Jianping Shi, Jiaying Liu, and Xiaodi Hou. Revisiting batch normalization for practical domain adaptation. *arXiv preprint arXiv:1603.04779*, 2016.

[36] Yuhang Li, Feng Zhu, Ruihao Gong, Mingzhu Shen, Fengwei Yu, Shaoqing Lu, and Shi Gu. Learning in school: Multi-teacher knowledge inversion for data-free quantization. *arXiv preprint arXiv:2011.09899*, 2020.

[37] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 6028–6039. PMLR, 2020.

[38] Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2604–2613, 2019.

[39] Raphael Gontijo Lopes, Stefano Fenu, and Thad Starner. Data-free knowledge distillation for deep neural networks. *arXiv preprint arXiv:1710.07535*, 2017.

[40] Liangchen Luo, Mark Sandler, Zi Lin, Andrey Zhmoginov, and Andrew Howard. Large-scale generative data-free distillation. *arXiv preprint arXiv:2012.05578*, 2020.

[41] Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain generalization using causal matching. *arXiv preprint arXiv:2006.07500*, 2020.

[42] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5188–5196, 2015.

[43] Fabio Maria Carlucci, Paolo Russo, Tatiana Tommasi, and Barbara Caputo. Hallucinating agnostic images to generalize across domains. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019.

[44] Paul Micaelli and Amos Storkey. Zero-shot knowledge transfer via adversarial belief matching. *arXiv preprint arXiv:1905.09768*, 2019.

[45] Asit Mishra and Debbie Marr. Apprentice: Using knowledge distillation techniques to improve low-precision network accuracy. *arXiv preprint arXiv:1711.05852*, 2017.

[46] Alexander Mordvintsev, Christopher Olah, and Mike Tyka. Inceptionism: Going deeper into neural networks. 2015.

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

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

[49] Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap via style-agnostic networks. *arXiv e-prints*, 2019.

[50] Gaurav Kumar Nayak, Konda Reddy Mopuri, Vaisakh Shaj, Venkatesh Babu Radhakrishnan, and Anirban Chakraborty. Zero-shot knowledge distillation in deep networks. In *International Conference on Machine Learning*, pages 4743–4751. PMLR, 2019.

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

[52] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436, 2015.

[53] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3967–3976, 2019.

[54] Antonio Polino, Razvan Pascanu, and Dan Alistarh. Model compression via distillation and quantization. *arXiv preprint arXiv:1802.05668*, 2018.

[55] Fengchun Qiao, Long Zhao, and Xi Peng. Learning to learn single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[56] Mohammad Mahfujur Rahman, Clinton Fookes, Mahsa Baktashmotlagh, and Sridha Sridharan. Multi-component image translation for deep domain generalization. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019.

[57] Jathushan Rajasegaran, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Mubarak Shah. Self-supervised knowledge distillation for few-shot learning. *arXiv preprint arXiv:2006.09785*, 2020.

[58] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

[59] 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 preprint arXiv:2106.02266*, 2021.

[60] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. *arXiv preprint arXiv:1804.10745*, 2018.

[61] Yuge Shi, Jeffrey Seely, Philip HS Torr, N Siddharth, Awni Hannun, Nicolas Usunier, and Gabriel Synnaeve. Gradient matching for domain generalization. *arXiv preprint arXiv:2104.09937*, 2021.

[62] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

[63] Aman Sinha, Hongseok Namkoong, Riccardo Volpi, and John Duchi. Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.

[64] Nathan Somavarapu, Chih-Yao Ma, and Zsolt Kira. Frustratingly simple domain generalization via image stylization. *arXiv preprint arXiv:2006.11207*, 2020.

[65] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, 2016.

[66] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011.

[67] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

[68] Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 1999.

[69] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. *arXiv preprint arXiv:1805.12018*, 2018.

[70] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.

[71] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 2018.

[72] Yufei Wang, Haoliang Li, and Alex C Kot. Heterogeneous domain generalization via domain mixup. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[73] Garrett Wilson and Diane J Cook. A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2020.

[74] Minghao Xu, Jian Zhang, Bingbing Ni, Teng Li, Chengjie Wang, Qi Tian, and Wenjun Zhang. Adversarial domain adaptation with domain mixup. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[75] Zheng Xu, Yen-Chang Hsu, and Jiawei Huang. Training shallow and thin networks for acceleration via knowledge distillation with conditional adversarial networks. *arXiv preprint arXiv:1709.00513*, 2017.

[76] Shen Yan, Huan Song, Nanxiang Li, Lincan Zou, and Liu Ren. Improve unsupervised domain adaptation with mixup training. *arXiv preprint arXiv:2001.00677*, 2020.

[77] Hao-Wei Yeh, Baoyao Yang, Pong C Yuen, and Tatsuya Harada. Sofa: Source-data-free feature alignment for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 474–483, 2021.

[78] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8715–8724, 2020.

[79] Chris Yoon, Ghassan Hamarneh, and Rafeef 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.

[80] Xiangyu Yue, Yang Zhang, Sicheng Zhao, Alberto Sangiovanni-Vincentelli, Kurt Keutzer, and Boqing Gong. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

[81] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.

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

[83] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15658–15667, 2021.

[84] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12975–12983, 2020.

[85] Brady Zhou, Nimit Kalra, and Philipp Krähenbühl. Domain adaptation through task distillation. In *European Conference on Computer Vision*, pages 664–680. Springer, 2020.

[86] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *arXiv preprint arXiv:2103.02503*, 2021.

[87] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Deep domain-adversarial image generation for domain generalisation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[88] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. *arXiv preprint arXiv:2104.02008*, 2021.

A Related Work

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

A.1 Domain Generalization

Domain Generalization (DG) approaches can be broadly 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) [19, 32], the minimization of a distance metric between the domain-specific means [67] or covariance matrices [65], the minimization of a contrastive loss [47, 79, 41, 26], or the maximization of loss gradient alignment [61, 59]. Other works use adversarial training with a domain discriminator model [16, 34] for the same purpose. Another category of works leverages meta-learning techniques, e.g., the bi-level optimization scheme proposed in [12], to optimize for quick adaptation to different domains [31], or to learn how to regularize the output layer [4]. A combination of meta-learning and embedding space regularization is proposed in [10]. Another line of works augment the training data to tackle DG. On the one hand, some approaches perturb the source domain data by computing inter-domain examples [74, 76, 72] via Mixup [82], by randomizing the style of images [49], by computing adversarial examples [18] using a class classifier [63, 69, 55] or a domain classifier [60], or corrupting learned features to incentivize new feature discovery [14]. On the other hand, CNNs are trained to generate new images from the source domains [56, 64, 6] or from novel domains [43, 87]. Other works perturb intermediate representations of the data [23, 88, 14]. We refer to [86] for a more extensive overview of DG approaches.

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.

A.2 Knowledge Distillation

Knowledge distillation (KD) [22] was originally proposed to compress the knowledge of a large teacher network into a smaller student network. Several KD extensions and improvements [58, 81, 75, 2, 53] enabled its application to a variety of scenarios including quantization [45, 54], domain adaptation [84, 85], semantic segmentation [38], and few-shot learning [57, 8]. While these methods rely on the original data, Data-Free Knowledge Distillation (DFKD) methods were recently developed [39, 44, 50, 7]. Hereby, knowledge is transferred from one [44, 50, 7, 9, 78, 40, 83] or multiple [36] teacher(s) to the student model via the generation of synthetic data, either by optimizing random noise examples [50, 78, 83] or by training a generator network [44, 7, 9, 40]. 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 [83] to the DFDG setting, and compare it to our approach (Section 3).

A.3 Source-free domain adaptation

The recently addressed Source-Free Domain Adaptation problem [37, 33, 28] 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 [33], a feature alignment mechanism [77], or a weighting of the target domain samples by their similarity to the source domain [28]. SHOT [37] employs an information maximization loss along with a self-supervised pseudo-labeling, and is extended to the multi-source scenario via source model weighting [1]. BUFR [11] aligns the target domain feature distribution with the one from the source domain. Another line of works leverage Batch Normalization (BN) [25] layers by replacing the BN-statistics computed on the source domain with those computed on the target domain [35], or by training the BN-parameters on the target domain via entropy minimization [70]. 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 [28, 37, 11] modify the training procedure on the source domain, which would not be possible in cases where the data is not accessible anymore.

B More details about DEKAN

In the following, we introduce the first DEKAN stage in more detail. The second and third stages are described in Section 2.2. DEKAN's training procedure is described in Algorithm 1.

B.0.1 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 [83], an improved version of the data-free knowledge distillation method DeepInversion (DI) [78] that enables the generation of more diverse images. Hereby, we use inceptionism-style [46] image synthesis, also called DeepDream, i.e., we initialize random noise images $\hat{x}$ and optimize them to be recognized as a sample from a pre-defined class by a trained model. This process is also referred to as Inversion [13, 78]. Following [78, 83], uniformly sample labels y and optimize the corresponding random images $\hat{x}$ by minimizing the domain-specific inversion loss L_{DS} given by

$$L_{DS} = L_C(T(\hat{x}), y) + \lambda_1 L_R(\hat{x}) + \lambda_2 L_M(\hat{x}), \tag{4}$$

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 [42, 52, 46, 78]. L_M, also called moment matching loss [40], 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,

$$L_M(\hat{x}) = \sum_l max(\|\mu_l(\hat{x}) - \hat{\mu_l}\|_2 - \delta_l, 0) + \sum_l max(\|\sigma_l^2(\hat{x}) - \hat{\sigma_l}^2\|_2 - \gamma_l, 0). \tag{5}$$

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 [78]. 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 [83]. 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 [83]. We note that the higher the value of the hyperparameter ϵ_{DS}, the higher the relaxation.

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

B.0.2 Algorithm

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 [27] in our case. Explicit update rule formulas and iteration over the synthetic data batches are omitted for simplicity of notation.

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

for $i \leftarrow 1$ to I **do**

    Initialize the domain-specific synthetic dataset D_g^i: Images $\hat{x} \sim \mathcal{N}(0, I)$ and arbitrary labels

    while not converged **do**

        Update D_g^i by minimizing the domains-specific inversion loss L_{DS} (Eq. 4) using T_i

    end while

end for

// Second stage: Cross-Domain Knowledge Fusion

for $i \leftarrow 1$ to I **do**

    for $j \leftarrow 1$ to I and $i \neq j$ **do**

        Initialize the cross-domain synthetic dataset D_g^{ij}: Images $\hat{x} \sim \mathcal{N}(0, I)$ and arbitrary labels

        while not converged **do**

            Update D_g^{ij} by minimizing the cross-domain inversion loss L_{CD}^{ij} (Eq. 1) using T_i, T_j and D_g^j

        end while

    end for

end for

// Third stage: Multi-Domain Knowledge Distillation

Initialize the student model S_θ randomly or from a pre-trained model

Concatenate the domain-specific and cross-domain synthetic datasets into one dataset D_g

while not converged **do**

    Randomly sample a mini-batch $B = \{\hat{x}, y\}$ from D_g

    Update θ by minimizing the knowledge distillation loss L_{KD} (Eq. 3) using B and $T_{1..I}$

end while

return Domain-generalized student model S_θ