Multi-Domain Ensembles for Domain Generalization

Anonymous Author(s)
Affiliation
Address
email

Abstract

In this paper, we consider the challenging problem of multi-source zero shot domain generalization (MDG), where labeled training data from multiple source domains are available but with no access to data from the target domain. Many methods have been proposed to address this problem, but surprisingly the naïve solution of pooling all source data together and training a single ERM model is highly competitive. Constructing an ensemble of deep classifiers is a popular approach for building models that are calibrated under challenging distribution shifts. Hence, we propose MulDEns (Multi-Domain Deep Ensembles), a new approach for constructing deep ensembles in multi-domain problems that does not require to construct domain-specific models. Our empirical studies on multiple standard benchmarks show that MulDEns significantly outperforms ERM and existing ensembling solutions for MDG.

1 Introduction

Typical supervised machine learning models are developed with the assumption that the training and testing data are independent and identically distributed (*i.i.d.*). However, even models that produce high accuracies on the *i.i.d.* test set, can fail drastically when tested on a non *i.i.d.* test data [23]. This severe drop in performance indicates poor generalization capabilities of the learned models, and addressing this fundamental challenge has become an important topic of research [5, 7, 10]. Since it is often infeasible to know the distribution of test data ahead of time, zero-shot generalization is a more practical formulation. Furthermore, it is common to leverage data from multiple source domains to improve model generalization. Commonly referred to as zero-shot, multi-domain generalization (ZS-MDG), this formulation assumes that labeled training data from multiple source domains is available but with no access to the target domain.

The simplest solution to this problem is the naïve empirical risk minimization (ERM) [24] approach that minimizes an average loss computed on data pooled together from all available source domains. The inability of this approach to exploit statistical discrepancies between different domains has motivated the design of multi-domain learning techniques [26]. However, recently [8] reported a surprising finding that an appropriate model selection strategy can make ERMs highly competitive to sophisticated ZS-MDG methods on standard benchmarks. Since then, there is renewed interest in improving the performance of the ERM baseline. In this context, approaches that enforce ERM-based models to be consistent under appropriate data augmentations have become popular [18, 27]. Despite their effectiveness, choosing the most appropriate augmentation for a given dataset is challenging and hence can provide varying degrees of performance gains across datasets (see Figure 1(right)). Ensembling methods [13] form another important class of approaches for MDG [4, 20]. Existing ensembling solutions have focused extensively on combining domain-specific models and interestingly, we find that, these approaches do not fair competitively even when compared to the vanilla ERM performance from the recent DomainBed framework. In this paper, we design multi-domain ensembles that can leverage both within- and across-domain discrepancies to improve

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

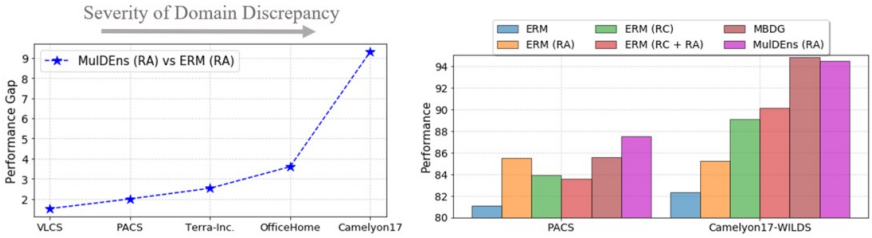


Figure 1: (left) Our proposed MulDEns significantly improves upon ERM as well as state-of-the-art ensemble construction methods in zero-shot multi-domain generalization, wherein we obtain larger performance gains as the domain discrepancy becomes more severe; (right) When compared to recent approaches that utilize advanced data augmentation strategies to improve generalization, MulDEns eliminates the need for choosing the suitable augmentations and is consistently effective for all benchmarks. Note, *RA*: RandAug, *RC*: RandConv, *MBDG*: model-based domain generalization.

generalization. We make the following contributions and key findings in regard to ZS-MDG that have not been reported before in the literature: (i) We propose MulDEns, a meta-optimization approach for constructing highly effective multi-domain ensembles; (ii) A gradient-matching based model assignment strategy that we find to be empirically superior to a variety of design choices; (iii) We investigate two new model selection strategies for multi-domain ensembles using only source domain validation sets similar to [8]; (iv) Using extensive empirical studies based on the DomainBed framework [8] with standard benchmarks, we find that MulDEns significantly improves over ERM as well as existing ensemble constructions of the same complexity (Figure 1(left)). (v) MulDEns works effectively on all datasets without the need to tweak the data augmentation strategy.

2 Approach

2.1 Meta Optimization for MulDEns

We denote a multi-domain ensemble by $\mathcal{E} : \{f_{\theta_m}\}_{m=1}^{M}$ with parameters θ_m, where M is the ensemble size and all models are initialized randomly. Labeled data from each observed domain $\mathcal{D}_k$ is divided into three disjoint sets - train set $\mathcal{D}_k^t$, meta-validation set $\mathcal{D}_k^v$ and held-out validation set $\mathcal{D}_k^{hv}$. Note, in contrast to existing ensembling methods in ZS-MDG [16], M can be different from K in our setup. MulDEns has two main stages, both operating at the mini-batch level - (i) a meta-train stage, where we obtain ERM-style gradients for each constituent member of the ensemble using the collection $\bigcup_k \mathcal{D}_k^t$ from all K source domains; (ii) a meta-test stage, where a *model relevance score* (MRS) is used to determine the most appropriate model from $\mathcal{E}$ (f_{θ_i}), to apply for each of the meta-validation sets $\mathcal{D}_k^v$ and subsequently f_{θ_i} is updated only using meta-gradients from the subset of $\mathcal{D}_k^v$'s assigned to this model. This step enables the ensemble to capture both intra- and inter-diversity in the source domains, since the MRS is computed for each mini-batch separately. Furthermore, we also explore the use of synthetic augmentations to create additional meta-validation batches for better modeling the intra-diversity in the domains. Finally, the held-out validation sets $\{\mathcal{D}_k^{hv}\}$ are used for model selection (following standard practice). Figure 2 lists the steps involved in our algorithm.

Algorithm 1: MulDEns training

Input: Set of training domains $\mathcal{D} := \{\mathcal{D}_1 \dots \mathcal{D}_K\}$,
Model relevance scoring function h
Output: Ensemble $\mathcal{E} := \{f_{\theta_1}, \dots, f_{\theta_M}\}$
Initialization: Parameters $\{\theta_1, \dots, \theta_M\}$, meta-train sets $\{\mathcal{D}_1^t \dots \mathcal{D}_K^t\}$, meta-validation sets $\{\mathcal{D}_1^v \dots \mathcal{D}_K^v\}$, hyper-parameters α, λ, η ;
for *iter* ***in*** n_{iter} **do**
  //meta-train //
  for f_{θ_m} *in* $f_{\theta_1} \cdots f_{\theta_M}$ **do**
    $\mathcal{B} = [\mathcal{B}_1, \cdots \mathcal{B}_K]$ // pool the minibatches from $\{\mathcal{D}_k^t\}$ //
    Compute empirical risk $\mathcal{L}_{\theta_m}$ w.r.t. $\mathcal{B}$, using eq. (2);
    Update $\theta'_m = \theta_m - \alpha \nabla_{\theta_m} \mathcal{L}_{\theta_m}(\theta_m)$ eq. (3) //inner gradient update//
  end
  //model relevance score//
  for k *in* $1 \cdots K$ **do**
    for f_{θ_m} *in* $f_{\theta_1} \cdots f_{\theta_M}$ **do**
      sample a mini-batch $\mathcal{V}_k$ from $\mathcal{D}_k^v$;
      compute β_{km} using eq. (4);
    end
  end
  //meta-update//
  for f_{θ_m} *in* $f_{\theta_1} \cdots f_{\theta_M}$ **do**
    compute γ_m as the set of indices of meta-validation batches relevant to f_{θ_m};
    compute meta-test loss $\mathcal{G}_{\theta'_m}$ using eq. (5);
    perform meta-update using eq. (6)
  end
end

Figure 2: An outline of the proposed approach.

Meta-train Stage - Produce ERM style gradients. In every iteration, K mini-batches $\{\mathcal{B}_k^t\}$ are randomly sampled from the K training sets $\{\mathcal{D}_k^t\}$, which are then pooled to form the data batch $\mathcal{B}$ ($\mathcal{B} = \bigcup_k \mathcal{B}_k^t \subset \mathcal{D}_k^t$) and passed as input to all M models. The empirical risk $\mathcal{L}_{\theta_m} = \frac{1}{|\mathcal{B}|} \sum_{(x_i,y_i)\in\mathcal{B}} \ell(f_{\theta_m}(x_i), y_i), \forall m \in (1, \cdots, M)$ (1) and the corresponding gradients for

each of the models $\nabla_{\theta_m}(\mathcal{L}_{\theta_m})$ are computed independently w.r.t. $\mathcal{B}$. MulDEns then takes one gradient step for each of the models f_{θ_m} to obtain $f_{\theta'_m}$.

Meta-test Stage - Model relevance, synthetic augmentation and meta-update. Here, we systematically regulate the gradient flow from the meta-validation data batches to each of the constituent models based on a model relevance score. We denote a generic, model relevance scoring function by $h\colon \mathcal{V}_k \times f_{\theta_m} \rightarrow \mathbb{R}^+[0,1]$ which scores the model f_{θ_m} for a mini-batch $\mathcal{V}_k \subset \mathcal{D}_k^v$ from the meta-validation dataset. We denote by β_{km} the resulting score *i.e.*,$\beta_{km} = h(\mathcal{V}_k, f_{\theta_m})$. We describe the various design choices of h and explain in detail our proposed gradient-matching based MRS. Intuitively, when the MRS is high, one expects that taking a gradient step for θ_m based on $\mathcal{B}$ is highly likely to improve the performance on $\mathcal{V}_k$. We compute this pair-wise relevance score between every pair of K meta-validation mini-batches and the M models to obtain the matrix $\boldsymbol{\beta} \in \mathbb{R}^{K\times M}$. The final step is to compute the meta-gradients for $\theta_m, \forall m$ w.r.t. to their "relevant" domains and perform a gradient-through-gradient update. We denote the indices of meta-validation batches that are assigned to f_{θ_m} by $\gamma_m = \{j \in (1, \cdots, K)\}$, such that for each j, model f_{θ_m} provides the largest MRS. The meta-validation loss $\mathcal{G}_{\theta'_m}$ of $f_{\theta'_m}$ using the relevant validation batches, γ_m, can be written as

$$\mathcal{G}_{\theta'_m} = \sum_{\forall (x_i,y_i)\in\{\mathcal{V}_j\}, j\in\gamma_m} \ell(f_{\theta'_m}(x_i), y_i), \tag{2}$$

where the definition of θ'_m comes from meta-train update. The final meta-update of f_{θ_m} using a gradient-through-gradient optimization is

$$\hat{\theta}_m = \theta_m - \lambda \frac{\partial(\mathcal{L}_{\theta_m} + \eta\mathcal{G}_{\theta'_m})}{\partial\theta_m}. \tag{3}$$

MulDEns Inference. We evaluate by averaging the predictions from all M models in the ensemble, to obtain labels for a sample $x \in \mathcal{D}^\dagger$, $\hat{y} = \frac{1}{M}\sum_{m=1}^{M} f_{\theta_m}(x)$.

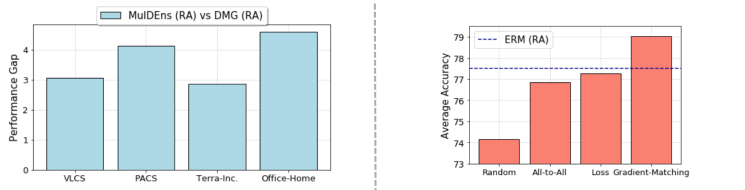
MRS Design. We have considered the following choices for h (i) *Random Assignment:* Here h assigns randomly assigns each mini-batch to one of the models; (ii) *All-to-All assignment*: Here every validation mini-batch $\mathcal{V}_k$ is assigned to all members of members. (iii) *Loss-based assignment*: Here, we directly use the empirical loss to determine the member assignment i.e., $\beta_{km} = 1 - \frac{1}{|\mathcal{V}_k|}\sum_{(x_i,y_i)\in\mathcal{V}_k} \ell(f_{\theta_m}(x_i), y_i)$; and (iv) *Gradient-matching based assignment*: While empirical loss-based assignment is a reasonable choice, we take inspiration from [2] and design MRS through gradient-matching. [2] showed that gradient embedding-based sample selection outperforms loss-based selection and gradient embeddings implicitly capture the model uncertainties.

Definition 2.1 (Gradient-matching based model relevance score)**.** $\beta_{km} = \sum \nabla_{\theta_m}(\mathcal{L}_{\theta_m}).\nabla_{\theta_m}(\mathcal{G}^k_{\theta_m})$. where $\mathcal{L}_{\theta_m}$ and $\mathcal{G}^k_{\theta_m}$ are the empirical risks computed using the model f_m on the meta-train ($\mathcal{B}$) and meta-validation ($\mathcal{V}_k$) batches respectively. The summation is over all parameters in θ_m, and this score computes the dot product between parameter sensitivities of θ_m w.r.t. the train and validation batches.

Model Selection Strategies A crucial component of any ZS-MDG algorithm is the specification of a model selection criterion. We propose two different model selections (i) Overall Avg: Here, we choose the checkpoint in which each individual model produces high accuracy on each of the K domains, on average. (ii) Overall Ens: we choose the checkpoint in which the ensemble model produces on average the highest accuracy for each of the K domains.

3 Experiments

Experimental Setup. We evaluate MulDEns using five standard visual MDG benchmarks (i) PACS [15], (ii) VLCS [6],(iii) OfficeHome [25] (iv) OfficeHome [25] (v) Camelyon17-WILDS [3,12]. Except for Camelyon17-WILDS, we run experiments by leaving out one of K domains for testing while using the $K-1$ domains for training. We use ResNet-50 [9], pre-trained on ImageNet [19] as the backbone feature extractor for all experiments. For MulDEns, we use a random 80-20 split from each of the source domains to obtain the train and validation sets, while the train set itself is further subdivided (80-20) to construct meta-train and meta-validation data. We report the mean and standard deviation of performance, obtained across three trials with different random seeds, for each experiment similar to [8]. In MulDEns ,the training mini-batches are augmented using a composition of the following augmentation choices: random horizontal flip, random color jitter and grayscaling



a) Comparison of MulDEns vs DMG

b) Effect of different MRS choices on VLCS dataset

Figure 3: (a) Our proposed MulDEns significantly improves upon DMG,a state-of-the-art ensemble construction method in ZS-MDG. We obtain larger performance gains as the domain discrepancy becomes more severe; (b) When compared against different design choices of MRS function h, we evidence that the proposed gradient-matching based h performs best.

Table 1: Summary performance of popular ZS-MDG baselines (obtained from [8]) on all 4 benchmarks. MulDEns with Overall (Avg.) model selection consistently generalizes well to novel domains. † denotes results from our implementation of baselines. We highlight the best performing method with bold face and next best with bold italics.

Methods	PACS	VLCS	OfficeHome	Terra Incognita
ERM [24]	85.5 $\pm$ 0.2	77.5 $\pm$ 0.4	66.5 $\pm$ 0.3	46.1 $\pm$ 1.8
IRM [1]	83.5 $\pm$ 0.8	78.5 $\pm$ 0.5	64.3 $\pm$ 2.2	47.6 $\pm$ 0.8
MLDG [14]	84.9 $\pm$ 1.0	77.2 $\pm$ 0.4	66.8 $\pm$ 0.6	47.7 $\pm$ 0.9
CORAL [22]	86.2 $\pm$ 0.3	***78.8 $\pm$ 0.6***	***68.7 $\pm$ 0.3***	47.6 $\pm$ 1.0
SagNet [17]	86.3 $\pm$ 0.2	77.8 $\pm$ 0.5	68.1 $\pm$ 0.1	***48.6 $\pm$ 1.0***
RSC [11]	85.2 $\pm$ 0.9	77.1 $\pm$ 0.5	65.5 $\pm$ 0.9	46.6 $\pm$ 1.0
FISH [21]	85.5 $\pm$ 0.3	77.8 $\pm$ 0.6	68.6 $\pm$ 0.4	45.1 $\pm$ 1.3
DSON [20]	***86.64***	-	-	-
DMG [4]	83.37	75.95† $\pm$ 0.2	66.6† $\pm$0.8	45.8† $\pm$0.3
MulDEns (Avg.)	**87.35 $\pm$ 0.2**	**79.02 $\pm$ 0.2**	69.76 $\pm$ 0.5	**48.66 $\pm$ 0.8**
MulDEns (Ens.)	87.21 $\pm$ 0.9	78.40 $\pm$ 0.2	**70.06 $\pm$ 0.2**	48.48 $\pm$ 1.0

with 10% probability, which we refer to as RandAug (RA). As described earlier, we also create additional meta-validation batches by augmenting each batch $\mathcal{V}_k$ using subsets of augmentations used during training. We set η in eq. (3) to 1.0 and study the sensitivity of this hyper-parameter as part of our ablation study. We report results for both the proposed model selection strategies and our rigorous empirical study shows that the Avg. strategy provides a small margin of improvement over Ens.

Key Findings. (i) MulDEns is a significantly improved baseline over ERM. As can be seen from Figure 1 (left) MulDEns outperforms ERM on all five widely-adopted benchmarks. Overall, across the benchmarks, MulDEns improves over ERM by a large margin (between 1.33% and 10.1%), in terms of average generalization performance; (ii) We perform a comparative analysis of MulDEns to state-of-the-art multi-domain ensembling methods such as DMG [4] and DSON [20]. In Figure 3(a), the superiority of MulDEns over DMG is clearly evident across all the benchmarks,with an average improvement of around 3.2%; (iii) As showed in Figure 3(b) for the VLCS dataset, the proposed gradient-matching based MRS performs the best, when compared to other design choices; (iv) MulDEns provides non-trivial improvements over SoTA DG methods, which rely on a variety of strategies to leverage cross-domain discrepancies. Interestingly, our approach while producing state-of-the-art results on PACS and OfficeHome, on Camelyon17-WILDS achieves an overall accuracy of 94.6% matching the best reported performance which is 94.8% by MBDG [18].

4 Conclusion

Through MulDEns we explored the design of deep ensembles for zero-shot MDG. While our approach builds upon ERM, in terms of avoiding the need for training domain-specific models, we leverage the intra-diversity within a domain and inter-diversity between domains to infer the constituent models of a deep ensemble. Using rigorous empirical studies on standard benchmarks, we find that MulDEns consistently outperforms the highly effective ERM baseline by a significant margin.

References

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[2] Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. In *International Conference on Learning Representations*, 2020.

[3] Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE transactions on medical imaging*, 38(2):550–560, 2018.

[4] Prithvijit Chattopadhyay, Yogesh Balaji, and Judy Hoffman. Learning to balance specificity and invariance for in and out of domain generalization. In *European Conference in Computer Vision (ECCV)*, 2020.

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

[6] Chen Fang, Ye Xu, and Daniel N Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1657–1664, 2013.

[7] 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*, 17(1):2096–2030, 2016.

[8] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[10] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. PMLR, 2018.

[11] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. *arXiv preprint arXiv:2007.02454*, 2, 2020.

[12] Pang Wei Koh, Shiori Sagawa, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

[13] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *arXiv preprint arXiv:1612.01474*, 2016.

[14] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[15] 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*, pages 5542–5550, 2017.

[16] Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M Hospedales. Episodic training for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1446–1455, 2019.

[17] Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap via style-agnostic networks. *arXiv preprint arXiv:1910.11645*, 2019.

[18] Alexander Robey, George J. Pappas, and Hamed Hassani. Model-based domain generalization. *ArXiv*, abs/2102.11436, 2021.

[19] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[20] Seonguk Seo, Yumin Suh, Dongwan Kim, Geeho Kim, Jongwoo Han, and Bohyung Han. Learning to optimize domain specific normalization for domain generalization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 68–83. Springer, 2020.

[21] 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.

[22] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450. Springer, 2016.

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

[24] Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

[25] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5018–5027, 2017.

[26] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, and Tao Qin. Generalizing to unseen domains: A survey on domain generalization. *CoRR*, abs/2103.03097, 2021.

[27] Zhenlin Xu, Deyi Liu, Junlin Yang, Colin Raffel, and Marc Niethammer. Robust and generalizable visual representation learning via random convolutions. In *International Conference on Learning Representations*, 2021.