Uncertainty-guided Model Generalization to Unseen Domains

Fengchun Qiao
University of Delaware
fengchun@udel.edu

Xi Peng
University of Delaware
xipeng@udel.edu

Abstract

We study a worst-case scenario in generalization: Out-of-domain generalization from a single source. The goal is to learn a robust model from a single source and expect it to generalize over many unknown distributions. This challenging problem has been seldom investigated while existing solutions suffer from various limitations. In this paper, we propose a new solution. The key idea is to augment the source capacity in both input and label spaces, while the augmentation is guided by uncertainty assessment. To the best of our knowledge, this is the first work to (1) access the generalization uncertainty from a single source and (2) leverage it to guide both input and label augmentation for robust generalization. The model training and deployment are effectively organized in a Bayesian meta-learning framework. We conduct extensive comparisons and ablation study to validate our approach. The results prove our superior performance in a wide scope of tasks including image classification, semantic segmentation, text classification, and speech recognition.

1. Introduction

Existing machine learning algorithms have achieved remarkable success under the assumption that training and test data are sampled from similar distributions. When this assumption no longer holds, even strong models (*e.g.,* deep neural networks) may fail to produce reliable predictions. In this paper, we study a worst-case scenario in generalization: *Out-of-domain generalization from a single source.* A model learned from a single source is expected to generalize over a series of unknown distributions. This problem is more challenging than *domain adaptation* [39, 42, 63, 34] which usually requires the assessment of target distributions during training, and *domain generalization* [41, 14, 33, 4, 9] which often assumes the availability of multiple sources. For example, there exists significant distribution difference in medical images collected across different hospitals. The intelligent diagnosis system is required to process images unexplored during training where model update is infeasible due to time or resource limitations.

Recently, [59] casts this problem in an ensemble framework. It learns a group of models each of which tackles an unseen test domain. This is achieved by performing *adversarial training* [15] on the source to mimic the unseen test distributions. Yet, its generalization capability is limited due to the proposed semantic constraint, which allows only a small amount of data augmentation to avoid semantic changes in the label space. To address this limitation, [45] proposes *adversarial domain augmentation* to relax the constraint. By maximizing the Wasserstein distance between the source and augmentation, the domain transportation is significantly enlarged in the input space.

However, existing data (domain) augmentation based methods [59, 44, 8, 6, 22] merely consider to increase the source capacity by perturbing the input space. Few of them investigate the possibility of label augmentation. An exception is Mixup [66] which pioneers label augmentation by randomly interpolating two data examples in both input and label spaces. However, Mixup can hardly address the out-of-domain generalization problem since it is restricted in creating in-domain generations due to the linear interpolation assumption. Besides, the interpolations are randomly sampled from a fixed distribution, which also largely restricts the flexibility of domain mixtures, yielding sub-optimal performance for unseen domain generalization.

Another limitation of existing work [41, 14, 33, 4, 9] is they usually overlook the potential risk of leveraging augmented data in tackling out-of-domain generalization. This raises serious safety and security concerns in mission-critical applications [11]. For instance, when deploying self-driving cars in unknown environments, it is crucial to be aware of the predictive uncertainty in risk assessment.

To tackle the aforementioned limitations, we propose uncertain out-of-domain generalization. The key idea is to increase the source capacity guided by uncertainty estimation in both input and label spaces. More specifically, in the input space, instead of directly augmenting raw data [59, 45], we apply uncertainty-guided perturbations to latent fea-

[1]The source code and pre-trained models are publicly available at: https://github.com/joffery/UMGUD.

tures, yielding a domain-knowledge-free solution for various modalities such as image, text, and audio. In the label space, we leverage the uncertainty associated with feature perturbations to augment labels via interpolation, improving generalization over unseen domains. Moreover, we explicitly model the domain uncertainty as a byproduct of feature perturbation and label mixup, guaranteeing fast risk assessment without repeated sampling. Finally, we organize the training and deployment in a Bayesian meta-learning framework that is specially tailored for single source generalization. To summarize, our contribution is multi-fold:

- To the best of our knowledge, we are the first to access the uncertainty from a single source. We leverage the uncertainty assessment to gradually improve the domain generalization in a curriculum learning scheme.
- For the first time, we propose learnable label mixup in addition to widely used input augmentation, further increasing the domain capacity and reinforcing generalization over unseen domains.
- We propose a Bayesian meta-learning method to effectively organize domain augmentation and model training. Bayesian inference is crucial in maximizing the posterior of domain augmentations, such that they can approximate the distribution of unseen domains.
- Extensive comparisons and ablation study prove our superior performance in a wide scope of tasks including image classification, semantic segmentation, text classification, and speech recognition.

2. Related Work

Out-of-Domain Generalization. Domain generalization [14, 32, 18, 50, 4, 9, 58, 67] has been intensively studied in recent years. JiGen [4] proposed to generate jigsaw puzzles from source domains and leverage them as self-supervised signals. Wang *et al.* [61] leveraged both extrinsic relationship supervision and intrinsic self-supervision for domain generalization. Specially, GUD [59] proposed adversarial data augmentation to solve single domain generalization, and learned an ensemble model for stable training. M-ADA [45] extended it to create augmentations with large domain transportation, and designed an efficient meta-learning scheme within a single unified model. Both GUD [59] and M-ADA [45] fail to assess the uncertainty of augmentations and only augment the input, while our method explicitly model the uncertainty and leverage it to increase the augmentation capacity in both input and label spaces. Several methods [38, 60, 21] proposed to leverage adversarial training [15] to learn robust models, which can also be applied in single source generalization. PAR [60] proposed to learn robust global representations by penalizing the predictive power of local representations. [21] applied self-supervised learning to improve the model robustness.

Adversarial training. Szegedy *et al.* [55] discovered the intriguing weakness of deep neural networks to minor adversarial perturbations. Goodfellow *et al.* [15] proposed adversarial training to improve model robustness against adversarial samples. Madry *et al.* [38] illustrated that adversarial samples generated through projected gradient descent can provide robustness guarantees. Sinha *et al.* [52] proposed principled adversarial training with robustness guarantees through distributionally robust optimization. More recently, Stutz *et al.* [53] illustrated that on-manifold adversarial samples can improve generalization. Therefore, models with both robustness and generalization can be achieved at the same time. In our work, we leverage adversarial training to create feature perturbations for domain augmentation instead of directly perturbing raw data.

Meta-learning. Meta-learning [49, 56] is a long standing topic on learning models to generalize over a distribution of tasks. Model-Agnostic Meta-Learning (MAML) [10] is a recent gradient-based method for fast adaptation to new tasks. In this paper, we propose a modified MAML to make the model generalize over the distribution of domain augmentation. Several approaches [33, 1, 9] have been proposed to learn domain generalization in a meta-learning framework. Li *et al.* [33] firstly applied MAML in domain generalization by adopting an episodic training paradigm. Balaji *et al.* [1] proposed to meta-learn a regularization function to train networks which can be easily generalized to different domains. Dou *et al.* [9] incorporated global and local constraints for learning semantic feature spaces in a meta-learning framework. However, these methods cannot be directly applied for single source generalization since there is only one distribution available during training.

Uncertainty Assessment. Bayesian neural networks [23, 17, 3] have been intensively studied to integrate uncertainty into weights of deep networks. Instead, we apply Bayesian inference to assess the uncertainty of domain augmentations. Several Bayesian meta-learning frameworks [16, 11, 64, 30, 36] have been proposed to model the uncertainty of few-shot tasks. Grant *et al.* [16] proposed the first Bayesian variant of MAML [10] using the Laplace approximation. Yoon *et al.* [64] proposed a novel Bayesian MAML with a stein variational inference framework and chaser loss. Finn *et al.* [11] approximated MAP inference of the task-specific weights while maintain uncertainty only in the global weights. Lee *et al.* [30] proposed a Bayesian meta-learning framework to deal with class/task imbalance and out-of-distribution tasks. Lee *et al.* [31] proposed meta-dropout which generates learnable perturbations to regularize few-shot learning models. In this paper, instead of modelling the uncertainty of tasks, we propose a novel Bayesian meta-learning framework to maximize the posterior distribution

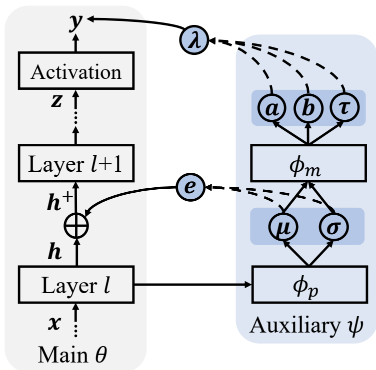

Figure 1: The main and auxiliary models.

Algorithm 1: Unseen Domain Generalization.

Input: Source domain $\mathcal{S}$, # of MC samples K.
Output: Learned backbone θ and auxiliary ψ.
while *not converged* **do**
  Meta-train: Compute θ^* on $\mathcal{S}$ using Eq. 4
  Generate $\mathcal{S}^+$ from $\mathcal{S}$ using Eq. 1
  for $k = 1, ..., K$ **do**
    Sample feature perturbation $\mathbf{h}_k^+$ using Eq. 2
    Generate label mixup $\mathbf{y}_k^+$ using Eq. 3
    Meta-test: Evaluate $\mathcal{L}(\theta^*; \mathcal{S}^+)$ w.r.t. $\mathcal{S}^+$
  end
  Meta-update: Update θ and ψ using Eq. 6
end

of domain augmentations.

3. Method

We first describe our problem setting and overall framework design. The goal is to learn a robust model from a *single* domain $\mathcal{S}$ and we expect the model to generalize over an *unknown* domain distribution $\{\mathcal{T}_1, \mathcal{T}_2, \cdots\} \sim p(\mathcal{T})$. This problem is more challenging than *domain adaptation* (assuming $p(\mathcal{T})$ is given) and *domain generalization* (assuming multiple source domains $\{\mathcal{S}_1, \mathcal{S}_2, \cdots\}$ are available). We create a series of domain augmentations $\{\mathcal{S}_1^+, \mathcal{S}_2^+, \cdots\} \sim p(\mathcal{S}^+)$ to approximate $p(\mathcal{T})$, from which the backbone θ can learn to generalize over unseen domains.

Uncertainty-guided domain generalization. We assume that $\mathcal{S}^+$ should integrate uncertainty assessment for efficient domain generalization. To achieve it, we introduce the auxiliary $\psi = \{\phi_p, \phi_m\}$ to explicitly model the uncertainty with respect to θ and leverage it to create $\mathcal{S}^+$ by increasing the capacity in both input and label spaces. In input space, we introduce ϕ_p to create feature augmentations $\mathbf{h}^+$ via adding perturbation $\mathbf{e}$ sampled from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})$. In label space, we integrate the same uncertainty encoded in $(\boldsymbol{\mu}, \boldsymbol{\sigma})$ into ϕ_m and propose learnable mixup to generate $\mathbf{y}^+$ (together with $\mathbf{h}^+$) through three variables (a, b, τ), yielding consistent augmentation in both input and output spaces. To effectively organize domain augmentation and model training, we propose a Bayesian meta-learning framework to *maximizing a posterior* of $p(\mathcal{S}^+)$ by jointly optimizing the backbone θ and the auxiliary ψ. The overall framework is shown in Fig. 1 and full algorithm is summarized in Alg. 1.

Merits of uncertainty assessment. Assessing the uncertainty of $\mathcal{S}^+$ plays a key role in our design. First, it provides consistent guidance to the augmentation in both input and label spaces when inferring $\mathcal{S}^+$, which has never been studied before. Second, we can gradually enlarge the domain transportation by increasing the uncertainty of $\mathcal{S}^+$ in a curriculum learning scheme [2]. Last, we can easily assess the domain uncertainty by checking the value of $\boldsymbol{\sigma}$, which measures how unsure it is when deploying on unseen domains $\mathcal{T}$ (Sec. 3.3).

3.1. Uncertainty-Guided Input Augmentation

The goal is to create $\mathcal{S}^+$ from $\mathcal{S}$ such that $p(\mathcal{S}^+)$ can approximate the out-of-domain distribution of $\mathcal{S}$. One the one hand, we expect a large domain transportation from $\mathcal{S}$ to $\mathcal{S}^+$ to best accommodate the unseen testing distribution $p(\mathcal{T})$. On the other hand, we prefer the transportation is domain-knowledge-free with uncertainty guarantee for broad and safe domain generalization. Towards this goal, we introduce ϕ_p to create feature augmentation $\mathbf{h}^+$ with large domain transportation through increasing the uncertainty with respect to θ.

Adversarial Domain Augmentation. To encourage large domain transportation, we cast the problem in a worst-case scenario [52] and propose to learn the auxiliary mapping ϕ_p via *adversarial domain augmentation*:

$$\underset{\phi_p}{\text{maximize}} \underbrace{\mathcal{L}(\theta; \mathcal{S}^+)}_{\text{Main task}} - \beta \underbrace{\left\|\mathbf{z} - \mathbf{z}^+\right\|_2^2}_{\text{Constraint}}. \tag{1}$$

Here, $\mathcal{L}$ denotes empirical loss such as cross-entropy loss for classification. The second term is the worst-case constraint, bounding the largest domain discrepancy between $\mathcal{S}$ and $\mathcal{S}^+$ in embedding space. $\mathbf{z}$ denotes the FC-layer output right before the activation layer, which is distinguished from $\mathbf{h}$ that denotes the Conv-layer outputs.

One merit of the proposed uncertainty-guided augmentation is that we can effectively relax the constraint to encourage large domain transportation in a curriculum learning scheme, which is significantly more efficient than [45] that has to train an extra WAE-GAN [57] to achieve this goal. We introduce the detailed form of $\mathbf{h}^+$ as follows.

Variational feature perturbation. To achieve adversarial domain augmentation, we apply uncertainty-guided perturbations to latent features instead of directly augmenting raw data, yielding domain-knowledge-free augmentation. We propose to learn layer-wise feature perturbations $\mathbf{e}$ that transport latent features $\mathbf{h} \rightarrow \mathbf{h}^+$ for efficient domain augmentation $\mathcal{S} \rightarrow \mathcal{S}^+$. Instead of a direct generation $\mathbf{e} = f_{\phi_p}(\mathbf{x}, \mathbf{h})$ widely used in previous work [59, 45], we assume $\mathbf{e}$ follows a multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})$, which can be used to easily access the uncertainty. More specifically, the Gaussian parameters are learnable via variational inference $(\boldsymbol{\mu}, \boldsymbol{\sigma}) = f_{\phi_p}(\mathcal{S}, \theta)$, such that:

$$\mathbf{h}^+ \leftarrow \mathbf{h} + \text{Softplus}(\mathbf{e}), \text{ where } \mathbf{e} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}), \quad (2)$$

where Softplus$(\cdot)$ is applied to stabilize the training. ϕ_p can create a series of feature augmentations $\{\mathbf{h}_1^+, \mathbf{h}_2^+, \cdots\}$ in different training iterations. In Sec. 4.5, we empirically show that $\{\mathbf{h}_1^+, \mathbf{h}_2^+, \cdots\}$ gradually enlarge the transportation through increasing the uncertainty of augmentations in a curriculum learning scheme and enable the model to learn from "easy" to "hard" domains.

3.2. Uncertainty-Guided Label Mixup

Feature perturbations not only augment the input but also yield label uncertainty. To explicitly model the label uncertainty, we leverage the input uncertainty, encoded in $(\boldsymbol{\mu}, \boldsymbol{\sigma})$, to infer the label uncertainty encoded in (a, b, τ) through ϕ_m as shown in Fig. 1. We leverage the label uncertainty to propose learnable label mixup, yielding consistent augmentation in both input and output spaces and further reinforcing generalization over unseen domains.

Random Mixup. We start by introducing random *mixup* [66] for robust learning. The key idea is to regularize the training to favor simple linear behavior in-between examples. More specifically, *mixup* performs training on convex interpolations of pairs of examples $(\mathbf{x}_i, \mathbf{x}_j)$ and their labels $(\mathbf{y}_i, \mathbf{y}_j)$:

$$\mathbf{x}^+ = \lambda \mathbf{x}_i + (1-\lambda)\mathbf{x}_j, \quad \mathbf{y}^+ = \lambda \mathbf{y}_i + (1-\lambda)\mathbf{y}_j,$$

where $\lambda \sim \text{Beta}(\alpha, \alpha)$ and the *mixup* hyper-parameter $\alpha \in (0, +\infty)$ controls the interpolation strength.

Learnable Label Mixup. We improve *mixup* by casting it in a learnable framework specially tailored for single source generalization. First, instead of mixing up pairs of examples, we mix up $\mathcal{S}$ and $\mathcal{S}^+$ to achieve in-between domain interpolations. Second, we leverage the uncertainty encoded in $(\boldsymbol{\mu}, \boldsymbol{\sigma})$ to predict learnable parameters (a, b), which controls the direction and strength of domain interpolations:

$$\mathbf{h}^+ = \lambda \mathbf{h} + (1-\lambda)\mathbf{h}^+, \quad \mathbf{y}^+ = \lambda \mathbf{y} + (1-\lambda)\tilde{\mathbf{y}}, \quad (3)$$

where $\lambda \sim \text{Beta}(a, b)$ and $\tilde{\mathbf{y}}$ denotes a *label-smoothing* [54] version of $\mathbf{y}$. More specifically, we perform *label smoothing* by a chance of τ, such that we assign $\rho \in (0, 1)$ to the true category and equally distribute $\frac{1-\rho}{c-1}$ to the others, where c counts categories. The Beta distribution (a, b) and the lottery τ are jointly inferred by $(a, b, \tau) = f_{\phi_m}(\boldsymbol{\mu}, \boldsymbol{\sigma})$ to integrate the uncertainty of domain augmentation.

3.3. A Unified Framework

To effectively organize domain augmentation and model training, we propose a Bayesian meta-learning framework to *maximize a posterior* of $p(\mathcal{S}^+)$ by jointly optimizing the backbone θ and the auxiliary $\psi = \{\phi_p, \phi_m\}$. Specifically, we *meta-train* the backbone θ on the source $\mathcal{S}$ and *meta-test* its generalization capability over $p(\mathcal{S}^+)$, where $\mathcal{S}^+$ is generated by performing data augmentation in both input (Sec. 3.1) and output (Sec. 3.2) spaces through the auxiliary ψ. Finally, we *meta-update* $\{\theta, \psi\}$ using gradient:

$$\nabla_{\theta,\psi} \mathbb{E}_{p(\mathcal{S}^+)}[\mathcal{L}(\theta^*; \mathcal{S}^+)], \text{where } \theta^* \equiv \theta - \alpha \nabla_\theta \mathcal{L}(\theta; \mathcal{S}). \quad (4)$$

Here θ^* is the meta-trained backbone on $\mathcal{S}$ and α is the learning rate. After training, the backbone θ is expected to bound the generalization uncertainty over unseen populations $p(\mathcal{T})$ in a worst-case scenario (Sec. 3.1) while ψ can be used to access the value of uncertainty efficiently.

Bayesian Meta-learning. The goal is to maximize the conditional likelihood of the augmented domain $\mathcal{S}^+$: $\log p\left(\mathbf{y}^+|\mathbf{x}, \mathbf{h}^+; \theta^*\right)$. However, solving it involves the true posterior $p\left(\mathbf{h}^+|\mathbf{x}; \theta^*, \psi\right)$, which is intractable [30]. Thus, we resort to amortized variational inference with a tractable form of approximate posterior $q\left(\mathbf{h}^+|\mathbf{x}; \theta^*, \psi\right)$. The approximated lower bound is as follows:

$$L_{\theta,\psi} = \mathbb{E}_{q(\mathbf{h}^+|\mathbf{x};\theta^*,\psi)}[\log \frac{p\left(\mathbf{y}^+|\mathbf{x}, \mathbf{h}^+; \theta^*\right)}{q\left(\mathbf{h}^+|\mathbf{x}; \theta^*, \psi\right)}]. \quad (5)$$

We leverage Monte-Carlo (MC) sampling to maximize the lower bound $L_{\theta,\psi}$ by:

$$\begin{aligned} \min_{\theta,\psi} \frac{1}{K} \sum_{k=1}^{K} &\left[-\log p\left(\mathbf{y}_k^+|\mathbf{x}, \mathbf{h}_k^+; \theta^*\right)\right] + \\ &\text{KL}\left[q\left(\mathbf{h}^+|\mathbf{x}; \theta^*, \psi\right) \| p\left(\mathbf{h}^+|\mathbf{x}; \theta^*, \psi\right)\right], \end{aligned} \quad (6)$$

where $\mathbf{h}_k^+ \sim q\left(\mathbf{h}^+|\mathbf{x}; \theta^*, \psi\right)$ and K is the number of MC samples. For KL divergence, traditional Gaussian prior $\mathcal{N}(\mathbf{0}, \mathbf{I})$ [24] is not compatible with our setup, since it may constrain the uncertainty of domain augmentations. Instead, we let $q\left(\mathbf{h}^+|\mathbf{x}; \theta^*, \psi\right)$ approximate $p\left(\mathbf{h}^+|\mathbf{x}; \theta^*, \psi\right)$ through adversarial training on ϕ_p in Eq. 1, so that the learned adversarial distribution is more flexible to approximate unseen domains. Thanks to the Bayesian meta-learning framework, the generalization uncertainty on unseen domains is significantly suppressed (Sec. 4.5). More importantly, a few examples of the target domain can quickly adapt θ to be domain-specific, yielding largely improved performance for few-shot domain adaptation (Sec. 4.1).

Uncertainty Estimation. At testing time, given a novel domain $\mathcal{T}$, we propose a *normalized domain uncertainty score*, $|\frac{\boldsymbol{\sigma}(\mathcal{T}) - \boldsymbol{\sigma}(\mathcal{S})}{\boldsymbol{\sigma}(\mathcal{S})}|$, to estimate its uncertainty with respect to learned θ. Considering ψ is usually much smaller than

θ, this score can be calculated efficiently by one-pass data forwarding through ψ. In Sec. 4.1, we empirically prove that our estimation is consistent with conventional Bayesian methods [3], while the time consumption is significantly reduced by an order of magnitude.

4. Experiments

To best validate the performance, we conduct a series of experiments to compare our approach with existing methods that can be roughly grouped in four categories: **1) Adversarial training:** PAR [60], Self-super [21], and PGD [38]. **2) Data augmentation:** Mixup [66], JiGen [4], Cutout [8], and AutoAug [6]. **3) Domain adaptation:** DIRT-T [51], SE [12], SBADA [47], FADA [39], and CCSA [40]. **4) Domain generalization:** ERM [25], GUD [59], and M-ADA [45]. The experimental results prove that our method achieves superior performance on a wide scope of tasks, including *image classification* [20], *semantic segmentation* [46], *text classification* [5], and *speech recognition* [62]. Please refer to supplementary for more details about experiment setup.

4.1. Image Classification

Datasets. We validate our method on the following two benchmark datasets for image classification. (1) *Digits* is used for digit classification and consists of five sub-datasets: MNIST [28], MNIST-M [13], SVHN [43], SYN [13], and USPS [7]. Each sub-dataset can be viewed as a different domain. Each image in these datasets contains one single digit with different styles and backgrounds. (2) *CIFAR-10-C* [20] is a robustness benchmark consisting of 19 corruptions types with five levels of severity applied to the test set of CIFAR-10 [26]. The corruptions consist of four main categories: noise, blur, weather, and digital. Each corruption has five-level severities and "5" indicates the most corrupted one.

Setup. *Digits*: following the setup in [59], we use 10,000 samples in the training set of MNIST for training, and evaluate models on the other four sub-datasets. We use a ConvNet [27] with architecture *conv-pool-conv-pool-fc-fc-softmax* as the backbone. All images are resized to 32×32, and the channels of MNIST and USPS are duplicated to make them as RGB images. *CIFAR-10-C*: we train models on CIFAR-10 and evaluate them on CIFAR-10-C. Following the setting of [22], we evaluate the model on 15 corruptions. We train models on AllConvNet (AllConv) [48] and Wide Residual Network (WRN) [65] with 40 layers and width of 2.

Results. 1) Classification accuracy. Tab. 1 shows the classification results of *Digits* and *CIFAR-10-C*. On the experiment of *Digits*, GUD [59], M-ADA [45], and our method outperform all baselines of the second block. And our method outperforms M-ADA [45] on *SYN* and the average accuracy by 8.1% and 1.8%, respectively. On the experiment of *CIFAR-10-C*, our method consistently outperforms all baselines on two different backbones, suggesting its strong generalization on various image corruptions. **2) Uncertainty estimation.** We compare the proposed *domain uncertainty score* (Sec.3.3) with a more time-consuming one based on Bayesian models [3]. The former computes the uncertainty through one-pass forwarding, while the latter computes the variance of the output through repeated sampling of 30 times. Fig. 2 show the results of uncertainty estimation on *Digits* and *CIFAR-10-C*. As seen, our estimation shows consistent results with Bayesian uncertainty estimation on both *Digits* and *CIFAR-10-C*, suggesting its high efficiency. **3) Few-shot domain adaptation.** Although our method is designed for single domain generalization, we also show that our method can be easily applied for few-shot domain adaptation [39] due to the meta-learning training scheme. Following the setup in [45], the model is first pre-trained on the source domain $\mathcal{S}$ and then fine-tuned on the target domain $\mathcal{T}$. We conduct three few-shot domain adaptation tasks: *USPS(U)→MNIST(M)*, *MNIST(M)→SVHN(S)*, and *SVHN(S)→MNIST(M)*. Results of the three tasks are shown in Tab. 2. Our method achieves the best performance on the average of three tasks. The result on the hardest task (*M→S*) is even competitive to that of SBADA [47] which uses all images of the target domain for training. Full results are provided in supplementary.

4.2. Semantic Segmentation

Datasets. *SYTHIA* [46] is a synthetic dataset of urban scenes, used for semantic segmentation in the context of driving scenarios. This dataset consists of photo-realistic frames rendered from virtual cities and comes with precise pixel-level semantic annotations. It is composed of the same traffic situation but under different locations (Highway, New York-like City, and Old European Town are selected) and different weather/illumination/season conditions (Dawn, Fog, Night, Spring, and Winter are selected).

Setup. In this experiment, Highway is the source domain, and New York-like City together with Old European Town are unseen domains. Following the protocol in [59, 45], we only use the images from the left front camera and 900 images are randomly sample from each source domain. We use FCN-32s [35] with the backbone of ResNet-50 [19].

Results. We report the mean Intersection Over Union (mIoU) of *SYTHIA* in Tab. 3. As can be observed, our method outperforms previous SOTA in most unseen environments. Results demonstrate that our model can better generalize to the changes of locations, weather, and time. We provide visual comparison in the supplementary.

4.3. Text Classification

Datasets. *Amazon Reviews* [5] contains reviews of products belonging to four categories - books(b), DVD(d), electronics(e) and kitchen appliances(k). The difference in tex-

Domain	Mixup [66]	PAR [60]	Self-super [21]	JiGen [4]	ERM [25]	GUD [59]	M-ADA [45]	**Ours**
SVHN [28]	28.5	30.5	30.0	33.8	27.8	35.5	<u>42.6</u>	**43.3**
MNIST-M [13]	54.0	58.4	58.1	57.8	52.8	60.4	**67.9**	<u>67.4</u>
SYN [13]	41.2	44.1	41.9	43.8	39.9	45.3	<u>49.0</u>	**57.1**
USPS [7]	76.6	76.9	77.1	77.2	76.5	77.3	**78.5**	<u>77.4</u>
Avg.	50.1	52.5	51.8	53.1	49.3	54.6	<u>59.5</u>	**61.3**

Model	Mixup [66]	Cutout [8]	AutoAug [6]	PGD [38]	ERM [25]	GUD [59]	M-ADA [45]	**Ours**
AllConv [48]	75.4	67.1	70.8	71.9	69.2	73.6	<u>75.9</u>	**79.6**
WRN [65]	77.7	73.2	76.1	73.8	73.1	75.3	<u>80.2</u>	**83.4**

Table 1: Image classification accuracy (%) on *Digits* [59] (**top**) and *CIFAR-10-C* [20] (**bottom**). We compare with *robust training* (**Columns 1-4**) and *domain generalization* (**Columns 5-7**). For *Digits*, all models are trained on *MNIST* [28]. For *CIFAR-10-C*, two widely employed backbones are evaluated. Our method outperforms M-ADA [45] (previous SOTA) consistently in all settings.

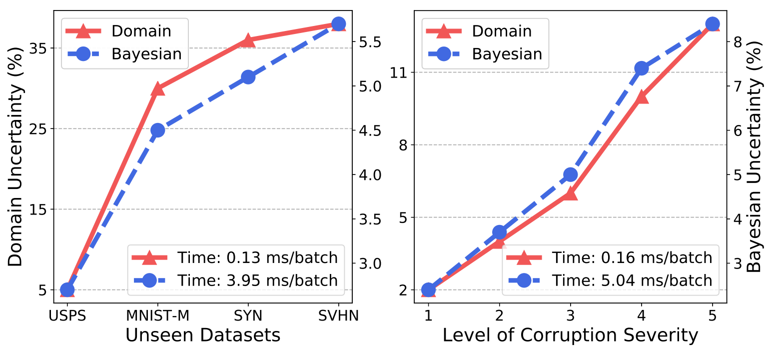

Figure 2: Uncertainty estimation on *Digits* (**left**) and *CIFAR-10-C* (**right**). Our prediction of *domain uncertainty* is consistent with *Bayesian uncertainty*, while our method is an order of magnitude faster since we forward data only once.

Method	$\lvert\mathcal{T}\rvert$	M $\rightarrow$ S	Avg.
DIRT-T [51]	All	54.5	-
SE [12]		14.0	70.4
SBADA [47]		**61.1**	78.3
FADA [39]	7	47.0	75.2
CCSA [40]	10	37.6	76.0
Ours	7	58.1	<u>80.1</u>
	10	<u>59.8</u>	**81.5**

Table 2: Few-shot domain adaptation accuracy (%) on *MNIST(M), USPS(U), and SVHN(S)*. $\lvert\mathcal{T}\rvert$ denotes the number of target samples (per class) used during model training.

tual description of the four product categories manifests as domain shift. Following [13], we use unigrams and bigrams as features resulting in 5000 dimensional representations.

Setup. We train the models on one source domain (books or dvd), and evaluate them on the other three domains. Similar to [13], we use a neural network with two hidden layers (both with 50 neurons) as the backbone.

Results. Tab. 4 shows the results of text classification on *Amazon Reviews* [5]. It appears that our method outperform previous ones on all the three unseen domains when the source domain is "books". We note that there is a little drop in performance on "electronics" when the source domain is "dvd". One possible reason is that "electronics" and "dvd" may share a similar distribution. And our method creates large distribution shift, degrading the performance on "electronics".

4.4. Speech Recognition

Datasets. *Google Commands* [62] contains 65000 utterances (one second long) from thousands of people. The goal is to classify them to 30 command words. There are 56196, 7477, and 6835 examples for training, validation, and test. To simulate domain shift in real-world scenario, we apply five common corruptions in both time and frequency domains. This creates five test sets that are "harder" than training sets, namely amplitude change (Amp.), pitch change (Pit.), background noise (Noise), stretch (Stretch), and time shift (Shift).

Setup. We train the models on the clean train set, and evaluate them on the corrupted test sets. We encode each audio into a mel-spectrogram with the size of 1x32x32 and feed them to LeNet [29] as one-channel input.

Results. Tab. 5 shows the results of speech recognition on *Google Commands* [62]. Our method outperforms the other three methods on all the five corrupted test sets, indicating its strong generalization ability in both time and frequency domain. In detail, our method outperforms the second best by 0.8% on "amplitude change", 1.4% on "pitch change", 0.4% on "background noise", 1.2% on "stretch", and 1.1% on "time shift", respectively. We can see that the

Source Domain	Method	New York-like City					Old European Town					
		Dawn	Fog	Night	Spring	Winter	Dawn	Fog	Night	Spring	Winter	Avg.
Highway/Dawn	ERM [25]	27.8	2.7	0.9	6.8	1.7	52.8	31.4	15.9	33.8	13.4	18.7
	GUD [59]	27.1	4.1	1.6	7.2	2.8	52.8	34.4	18.2	33.6	14.7	19.7
	M-ADA [45]	29.1	4.4	**4.8**	**14.1**	5.0	54.3	36.0	23.2	**37.5**	14.9	22.3
	Ours	**29.3**	**7.6**	2.8	12.7	**10.2**	**54.9**	**37.0**	**25.3**	37.2	**17.7**	**23.5**
Highway/Fog	ERM [25]	17.2	34.8	12.4	26.4	11.8	33.7	55.0	26.2	41.7	12.3	27.2
	GUD [59]	18.8	35.6	12.8	26.0	13.1	37.3	56.7	28.1	43.6	**13.6**	28.5
	M-ADA [45]	21.7	32.0	9.7	26.4	13.3	42.8	56.6	**31.8**	42.8	12.9	29.0
	Ours	**23.0**	**36.2**	**13.5**	**27.6**	**14.2**	**43.1**	**57.4**	31.0	**44.6**	13.1	**30.4**

Table 3: Semantic segmentation mIoU (%) on *SYNTHIA* [46]. All models are trained on the single source from *Highway* and evaluated on unseen environments from *New York-like City* and *Old European Town*.

Method	books			dvd		
	d	k	e	b	k	e
ERM [25]	78.7	74.6	63.6	78.5	82.1	**75.2**
GUD [59]	79.1	75.6	64.7	78.1	82.0	74.6
M-ADA [45]	79.4	76.1	65.3	78.8	82.6	74.3
Ours	**80.2**	**76.8**	**67.1**	**80.1**	**83.5**	75.0

Table 4: Text classification accuracy (%) on *Amazon Reviews*. Models are trained on one text domain and evaluated on unseen text domains. Our method outperforms others in all settings except "$dvd \rightarrow electronics$".

Method	Time		Frequency		
	Amp.	Pit.	Noise	Stretch	Shift
ERM [25]	63.8	71.6	73.9	72.9	70.5
GUD [59]	64.1	72.1	74.8	73.1	70.9
M-ADA [45]	64.5	71.9	75.4	73.8	71.4
Ours	**65.3**	**73.5**	**75.8**	**75.0**	**72.5**

Table 5: Speech recognition accuracy (%) on *Google Commands*. Models are trained on clean set and evaluated on five corrupted sets. Results validate our strong generalization on corruptions in both time and frequency domains.

	Digits [59]	CIFAR-10-C [20]
Full Model	**61.3**±0.73	**70.2**±0.62
Random Gaussian	51.0±0.36	64.0±0.18
Determ. perturb.	59.7±0.70	67.0±0.57
Random $\boldsymbol{\mu}$	60.5±0.75	69.1±0.61
Random $\boldsymbol{\sigma}$	60.7±0.65	69.5±0.60

Table 6: Ablation study of feature perturbation.

improvements on "pitch change", "stretch", and "time shift" are more significant than those on "amplitude change" and "background noise".

4.5. Ablation Study

In this section, we perform ablation study to investigate key components of our method. For *Digits* [59], we report the average performance of all unseen domains. For *CIFAR-10-C* [20], we report the average performance of all types of corruptions at the highest level of severity.

Uncertainty assessment. We visualize feature perturbation $|\mathbf{e}| = |\mathbf{h}^+ - \mathbf{h}|$ and the embedding of domains at different training iterations T on MNIST [28]. We use t-SNE [37] to visualize the source and augmented domains without and with uncertainty assessment in the embedding space. Results are shown in Fig. 3. In the model without uncertainty (left), the feature perturbation $\mathbf{e}$ is sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I})$ without learnable parameters. In the model with uncertainty (right), we observe that most perturbations are located in the background area which increases the variation of $\mathcal{S}^+$ while keeping the category unchanged. As a result, models with uncertainty can create large domain transportation in a curriculum learning scheme, yielding safe augmentation and improved accuracy on unseen domains. We visualize the density of $\mathbf{y}^+$ in Fig. 4. As seen, models with uncertainty can significantly augment the label space.

Variational feature perturbation. We investigate different designs of feature perturbation: *1) Random Gaussian:* the feature perturbation $\mathbf{e}$ is sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I})$ without learnable parameters. *2) Deterministic perturbation:* we directly add the learned $\boldsymbol{\mu}$ to $\mathbf{h}$ without sampling, yielding $\mathbf{h}^+ \leftarrow \mathbf{h} + \mathrm{Softplus}(\boldsymbol{\mu})$. *3) Random $\boldsymbol{\mu}$:* the feature perturbation $\mathbf{e}$ is sampled from $\mathcal{N}(\mathbf{0}, \boldsymbol{\sigma})$, where $\boldsymbol{\mu} = \mathbf{0}$. *4) Random $\boldsymbol{\sigma}$:* $\mathbf{e}$ is sampled from $\mathcal{N}(\boldsymbol{\mu}, \mathbf{I})$, where $\boldsymbol{\sigma} = \mathbf{I}$. Results on these different choices are shown in Tab. 6. As seen, *Random Gaussian* yields the lowest accuracy on both datasets, indicating the necessity of learnable perturbations. *Deterministic perturbation* is inferior to *Random $\boldsymbol{\mu}$* and *Random $\boldsymbol{\sigma}$*, suggesting that sampling-based perturbation can effectively increase the domain capacity. Finally, either *Random*

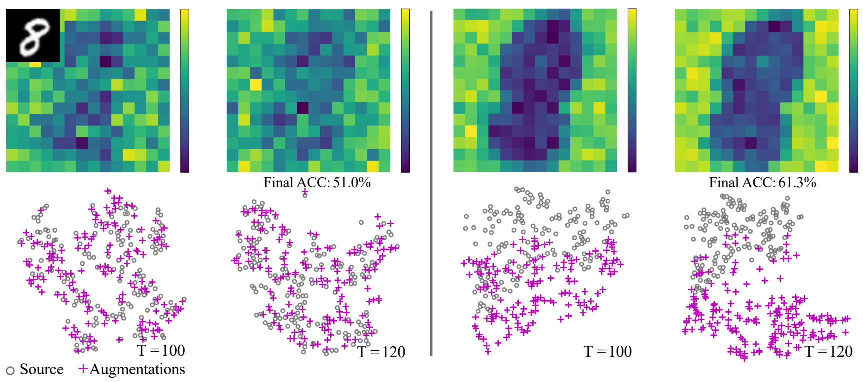

Figure 3: Visualization of feature perturbation $|\mathbf{e}| = |\mathbf{h}^+ - \mathbf{h}|$ (**Top**) and embedding of domains (**Bottom**) at different training iterations T on *MNIST*. **Left:** Models w/o uncertainty; **Right:** Models w/ uncertainty. Most perturbations are located in the background area and models w/ uncertainty can create large domain transportation in a curriculum learning scheme.

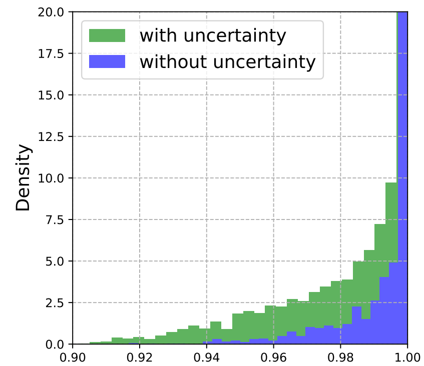

Figure 4: Visualization of label mixup $\mathbf{y}^+$ on *MNIST*. Models w/ uncertainty can encourage more smoothing labels and significantly increase the capacity of label space.

	Digits [59]	CIFAR-10-C [20]
Full Model	**61.3**±0.73	**70.2**±0.62
w/o mixup	60.6±0.76	67.4±0.64
Random mixup	60.9±1.10	69.4±0.58

Table 7: Ablation study of label mixup.

	Digits [59]	CIFAR-10-C [20]
Full Model	**61.3**±0.73	**70.2**±0.62
w/o adv. training	51.8±0.71	60.0±0.55
w/o meta-learning	60.9±1.24	68.7±0.81
w/o minimizing ϕ_p	60.6±0.91	69.6±0.75

Table 8: Ablation study of training strategy.

$\boldsymbol{\mu}$ or *Random* $\boldsymbol{\sigma}$ is slightly worse than the full model. We conclude that both learnable $\boldsymbol{\mu}$ and learnable $\boldsymbol{\sigma}$ contribute to the final performance.

Learnable label mixup. We implement two variants of label *mixup*: *1) Without mixup:* the model is trained without label augmentation. *2) Random mixup:* the mixup coefficient λ is sampled from a fixed distribution $\mathrm{Beta}(1, 1)$. Results on the two variants are reported in Tab. 7. We notice that *Random mixup* achieves better performance than *without mixup*. The results support our claim that label augmentation can further improve the model performance. The learnable mixup (full model) achieves the best results, suggesting that the proposed learning label mixup can create informative domain interpolations for robust learning.

Training strategy. At last, we compare different training strategies. *1) Without adversarial training:* models are learned without adversarial training (Eq. 1). *2) Without meta-learning:* the source $\mathcal{S}$ and augmentations $\mathcal{S}^+$ are trained together without the meta-learning scheme. *3) Without minimizing ϕ_p:* ϕ_p is not optimized in Eq. 6. Results are reported in Tab. 8. The adversarial training contributes most to the improvements: 9.5% on *Digits* and 10.2% on *CIFAR-10-C*. Meta-learning consistently improve the accuracy and reduce the deviation on both datasets. We notice that the accuracy is slightly dropped without minimization of ϕ_p, possibly due to the excessive accumulation of perturbations.

5. Conclusion

In this work, we introduced uncertainty-guided model generalization to unseen domains to tackle the problem of single source generalization. Our method explicitly model the uncertainty of domain augmentations in both input and label spaces. In input space, the proposed uncertainty-guided feature perturbation resolves the limitation of raw data augmentation, yielding a domain-knowledge-free solution for various modalities. In label space, the proposed uncertainty-guided label mixup further increases the domain capacity. Finally, the proposed Bayesian meta-learning framework can maximize the posterior distribution of domain augmentations, such that the learned model can generalize well on unseen domains. The experimental results prove that our method achieves superior performance on a wide scope of tasks, including *image classification*, *semantic segmentation*, *text classification*, and *speech recognition*.

Acknowledgments

This work is partially supported by National Science Foundation (NSF) CMMI-2039857 D-ISN-1.

References

[1] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. In *Annual Conference on Neural Information Processing Systems*, pages 998–1008, 2018.

[2] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *International Conference on Machine Learning*, pages 41–48, 2009.

[3] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International Conference on Machine Learning*, pages 1613–1622, 2015.

[4] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2229–2238, 2019.

[5] Minmin Chen, Zhixiang Xu, Kilian Q Weinberger, and Fei Sha. Marginalized denoising autoencoders for domain adaptation. In *International Conference on Machine Learning*, pages 1627–1634, 2012.

[6] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 113–123, 2019.

[7] John S Denker, WR Gardner, Hans Peter Graf, Donnie Henderson, Richard E Howard, W Hubbard, Lawrence D Jackel, Henry S Baird, and Isabelle Guyon. Neural network recognizer for hand-written zip code digits. In *Annual Conference on Neural Information Processing Systems*, pages 323–331, 1989.

[8] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[9] Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. In *Annual Conference on Neural Information Processing Systems*, pages 6447–6458, 2019.

[10] 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, 2017.

[11] Chelsea Finn, Kelvin Xu, and Sergey Levine. Probabilistic model-agnostic meta-learning. In *Annual Conference on Neural Information Processing Systems*, pages 9516–9527, 2018.

[12] Geoffrey French, Michal Mackiewicz, and Mark Fisher. Self-ensembling for visual domain adaptation. In *International Conference on Learning Representations*, 2018.

[13] Yaroslav Ganin and Victor Lempitsky. Unsupervised Domain Adaptation by Backpropagation. In *International Conference on Machine Learning*, pages 1180–1189, 2015.

[14] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2551–2559, 2015.

[15] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

[16] Erin Grant, Chelsea Finn, Sergey Levine, Trevor Darrell, and Thomas Griffiths. Recasting gradient-based meta-learning as hierarchical bayes. In *International Conference on Learning Representations*, 2018.

[17] Alex Graves. Practical variational inference for neural networks. In *Annual Conference on Neural Information Processing Systems*, pages 2348–2356, 2011.

[18] Thomas Grubinger, Adriana Birlutiu, Holger Schöner, Thomas Natschläger, and Tom Heskes. Multi-domain transfer component analysis for domain generalization. *Neural Processing Letters (NPL)*, 46(3):845–855, 2017.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[20] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *International Conference on Learning Representations*, 2019.

[21] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. In *Annual Conference on Neural Information Processing Systems*, pages 15637–15648, 2019.

[22] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. In *International Conference on Learning Representations*, 2020.

[23] Geoffrey E Hinton and Drew Van Camp. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the Sixth Annual ACM Conference on Computational Learning Theory*, pages 5–13, 1993.

[24] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

[25] Vladimir Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole d'Eté de Probabilités de Saint-Flour XXXVIII-2008*, volume 2033. Springer Science & Business Media, 2011.

[26] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[27] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation (NC)*, 1(4):541–551, 1989.

[28] 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.

[29] Yann Lecun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, 86(11), 1998.

[30] Hae Beom Lee, Hayeon Lee, Donghyun Na, Saehoon Kim, Minseop Park, Eunho Yang, and Sung Ju Hwang. Learning to balance: Bayesian meta-learning for imbalanced and out-of-distribution tasks. In *International Conference on Learning Representations*, 2020.

[31] Hae Beom Lee, Taewook Nam, Eunho Yang, and Sung Ju Hwang. Meta dropout: Learning to perturb latent features for generalization. In *International Conference on Learning Representations*, 2019.

[32] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Deeper, Broader and Artier Domain Generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5542–5550, 2017.

[33] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Learning to Generalize: Meta-Learning for Domain Generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3490–3497, 2018.

[34] Hong Liu, Mingsheng Long, Jianmin Wang, and Michael Jordan. Transferable adversarial training: A general approach to adapting deep classifiers. In *International Conference on Machine Learning*, pages 4013–4022, 2019.

[35] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[36] Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[37] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research (JMLR)*, 9(Nov):2579–2605, 2008.

[38] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[39] Saeid Motiian, Quinn Jones, Seyed Iranmanesh, and Gianfranco Doretto. Few-shot adversarial domain adaptation. In *Annual Conference on Neural Information Processing Systems*, pages 6670–6680, 2017.

[40] Saeid Motiian, Marco Piccirilli, Donald A. Adjeroh, and Gianfranco Doretto. Unified Deep Supervised Domain Adaptation and Generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5715–5725, 2017.

[41] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain Generalization via Invariant Feature Representation. In *International Conference on Machine Learning*, pages 10–18, 2013.

[42] Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. Image to image translation for domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4500–4509, 2018.

[43] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

[44] Xi Peng, Zhiqiang Tang, Fei Yang, Rogerio S Feris, and Dimitris Metaxas. Jointly Optimize Data Augmentation and Network Training: Adversarial Data Augmentation in Human Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2226–2234, 2018.

[45] 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*, pages 12556–12565, 2020.

[46] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3234–3243, 2016.

[47] Paolo Russo, Fabio M Carlucci, Tatiana Tommasi, and Barbara Caputo. From source to target and back: symmetric bi-directional adaptive gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8099–8108, 2018.

[48] Tim Salimans and Durk P Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *Annual Conference on Neural Information Processing Systems*, pages 901–909, 2016.

[49] Jürgen Schmidhuber. *Evolutionary principles in self-referential learning*. PhD thesis, Technische Universität München, 1987.

[50] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing Across Domains via Cross-Gradient Training. In *International Conference on Learning Representations*, 2018.

[51] Rui Shu, Hung H Bui, Hirokazu Narui, and Stefano Ermon. A dirt-t approach to unsupervised domain adaptation. In *International Conference on Learning Representations*, 2018.

[52] Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

[53] David Stutz, Matthias Hein, and Bernt Schiele. Disentangling adversarial robustness and generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6976–6987, 2019.

[54] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[55] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

[56] Sebastian Thrun and Lorien Pratt. *Learning to learn*. Springer Science & Business Media, 2012.

[57] I Tolstikhin, O Bousquet, S Gelly, and B Schölkopf. Wasserstein auto-encoders. In *International Conference on Learning Representations*, 2018.

[58] Riccardo Volpi and Vittorio Murino. Addressing model vulnerability to distributional shifts over image transformation sets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7980–7989, 2019.

[59] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. In *Annual Conference on Neural Information Processing Systems*, pages 5334–5344, 2018.

[60] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Annual Conference on Neural Information Processing Systems*, pages 10506–10518, 2019.

[61] Shujun Wang, Lequan Yu, Caizi Li, Chi-Wing Fu, and Pheng-Ann Heng. Learning from extrinsic and intrinsic supervisions for domain generalization. In *Proceedings of the European Conference on Computer Vision*, pages 159–176. Springer, 2020.

[62] Pete Warden. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209*, 2018.

[63] Xiang Xu, Xiong Zhou, Ragav Venkatesan, Gurumurthy Swaminathan, and Orchid Majumder. d-sne: Domain adaptation using stochastic neighborhood embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2497–2506, 2019.

[64] Jaesik Yoon, Taesup Kim, Ousmane Dia, Sungwoong Kim, Yoshua Bengio, and Sungjin Ahn. Bayesian model-agnostic meta-learning. In *Annual Conference on Neural Information Processing Systems*, pages 7332–7342, 2018.

[65] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference*, 2016.

[66] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

[67] Long Zhao, Ting Liu, Xi Peng, and Dimitris Metaxas. Maximum-entropy adversarial data augmentation for improved generalization and robustness. In *Annual Conference on Neural Information Processing Systems*, pages 14435–14447, 2020.