AdvST: Revisiting Data Augmentations for Single Domain Generalization

Guangtao Zheng[1], **Mengdi Huai**[2], **Aidong Zhang**[1]

[1]Department of Computer Science, University of Virginia
[2]Department of Computer Science, Iowa State University
gz5hp@virginia, mdhuai@iastate.edu, aidong@virginia.edu

Abstract

Single domain generalization (SDG) aims to train a robust model against unknown target domain shifts using data from a single source domain. Data augmentation has been proven an effective approach to SDG. However, the utility of standard augmentations, such as translate, or invert, has not been fully exploited in SDG; practically, these augmentations are used as a part of a data preprocessing procedure. Although it is intuitive to use many such augmentations to boost the robustness of a model to out-of-distribution domain shifts, we lack a principled approach to harvest the benefit brought from multiple these augmentations. Here, we conceptualize standard data augmentations with learnable parameters as semantics transformations that can manipulate certain semantics of a sample, such as the geometry or color of an image. Then, we propose Adversarial learning with Semantics Transformations (AdvST) that augments the source domain data with semantics transformations and learns a robust model with the augmented data. We theoretically show that AdvST essentially optimizes a distributionally robust optimization objective defined on a set of semantics distributions induced by the parameters of semantics transformations. We demonstrate that AdvST can produce samples that expand the coverage on target domain data. Compared with the state-of-the-art methods, AdvST, despite being a simple method, is surprisingly competitive and achieves the best average SDG performance on the Digits, PACS, and DomainNet datasets. Our code is available at https://github.com/gtzheng/AdvST.

Introduction

Domain generalization (Balaji, Sankaranarayanan, and Chellappa 2018; Li et al. 2017, 2018a, 2019) aims to learn a model that can generalize well on target (test) domains with unknown distribution shifts using multiple source (training) domains. However, having diverse domains for training is a strong assumption due to various practical considerations, such as data collection budgets or privacy issues. A realistic alternative is *single domain generalization* (SDG) (Zhao et al. 2020; Li et al. 2018a), which only requires data from a single source domain for model training. SDG is challenging for deep image classifiers. Although they have achieved impressive performance on benchmarks, they strongly hinge on the implicit assumption that training and test data follow the same distribution. Their performance can drop significantly when there are shifts between training and test data distributions caused by, for example, changes in object appearance or data collection methods.

Data augmentation is an effective approach to SDG. It augments the source domain data to expand the coverage on the unseen target domain during model training. Methods of data augmentation include using adversarial learning (Volpi et al. 2018; Zhao et al. 2020; Qiao, Zhao, and Peng 2020) or using generative models (Qiao, Zhao, and Peng 2020; Wang et al. 2021; Li et al. 2021) to generate diverse data samples. The utility of standard data augmentations, such as scale, or CutOut (DeVries and Taylor 2017), has not been fully exploited in SDG. In practice, these augmentation methods have been widely used in model training for *in-distribution* generalization. However, their applications in SDG are limited. In most cases, they serve as a part of the data preprocessing procedure in other SDG methods (Volpi et al. 2018; Zhao et al. 2020; Qiao, Zhao, and Peng 2020). Although it is intuitive that applying multiple standard data augmentations to the source domain data can generate diverse samples and hence improve a model's SDG performance, we lack a principled approach to fully realize the benefit brought from multiple standard data augmentations.

Therefore, in this paper, we revisit standard data augmentations for SDG and develop methods that make them a strong competitor in SDG. We consider the composition of several standard data augmentation as a semantics transformation which can manipulate certain kinds of semantics of a sample, such as the brightness and hue of an image. Normally, standard data augmentations have pre-specified and fixed parameters. Here, we make these parameters learnable in a semantics transformation so that we can tune these parameters to produce semantically significant variations and bring new styles that are different from the source domain data. With semantics transformations, we can transform data in the source domain to a fictitious one which has large domain shifts from the source and possibly covers data in target domains, yielding favorable SDG performance.

To learn semantics transformations for SDG, we propose AdvST, an adversarial learning framework that trains a robust model and generates challenging data samples iteratively with mini-max optimization. In the maximization phase, we learn the parameters of semantics transforma-

tions so that the samples transformed by semantics transformations maximize the prediction loss of the model. To avoid learning a trivial solution where the information in the source domain samples is completely lost after semantics transformations, we additionally regularize the distance between the source domain samples and the transformed ones in the deep feature space of the model to keep the core features of the source domain data. In the minimization phase, we train the model with the new samples generated by semantics transformations.

We theoretically show that the learning objective of AdvST connects to that of distributionally robust optimization (DRO) (Blanchet and Murthy 2019; Gao and Kleywegt 2022). DRO trains a robust model using the worst-case distribution that leads to the worst model performance on an uncertainty set—a set of neighboring distributions with a predefined value of distributional shifts from the training data. Increasing the coverage of the uncertainty set on the target domain data can improve the model's SDG performance. AdvST can be considered as a special form of DRO whose uncertainty set consists of semantics-induced data distributions which are generated by applying semantics transformations to samples from the source distribution. We demonstrate that AdvST can produce samples in the uncertainty set that expand the coverage on the target domain data.

Our method, despite being a simple method utilizing standard data augmentations, is surprisingly competitive in SDG. AdvST consistently outperforms existing state-of-the-art methods in terms of the average SDG performance on three benchmark datasets.

Related Work

Domain adaptation and generalization. Domain adaptation methods (French, Mackiewicz, and Fisher 2017; Ganin and Lempitsky 2015; Shu et al. 2018; Li et al. 2018b) have been proposed to solve the problem of generalizing to a target domain where the label information is unknown at training time. These methods mainly aim to align the distributions of source and target domains. However, their setups differ from ours since they require access to samples from the target distribution during training. In contrast, domain generalization methods (Balaji, Sankaranarayanan, and Chellappa 2018; Li et al. 2019; Pandey et al. 2021; Shankar et al. 2018) do not require samples from the target domain during training. However, they use training samples from multiple domains instead of one.

Single domain generalization. SDG requires no access to target distributions and only one single source domain for training. The general idea is to augment the source domain data, and there are three types of methods. Methods of the first type (DeVries and Taylor 2017; Hendrycks et al. 2019; Zoph et al. 2020; Cubuk et al. 2020a; Lian et al. 2021) use traditional data augmentation to improve *in-domain* generalization performance but often fail to generate samples with large domain shifts for out-of-domain generalization. The second type of methods use adversarial data augmentation to augment the source domain data. However, they generate samples either in the pixel space (Volpi et al. 2018; Zhao et al. 2020) or via perturbing latent feature statistics (Zhong et al. 2022; Zhang et al. 2023), which struggle to produce samples with large domain shifts. Our method exploits the domain knowledge in standard data augmentations and uses them as semantics transformations with learnable parameters, generating samples with large domain shifts from the source domain. Adversarial AutoAugment (Zhang et al. 2019) adversarially learns augmentation policies to improve *in-domain* generalization performance. In contrast, our method directly generates worst-case samples to improve *out-of-domain* generalization performance. The third type (Qiao, Zhao, and Peng 2020; Wang et al. 2021; Li et al. 2021) uses generative models to produce diverse training samples. However, since generative models are learned from the source domain, the styles of the generated samples are still related to those in the source domain. In contrast, our method uses semantics transformations to manipulate the semantics that is *independent* of the source domain, allowing us to inject *external* styles to the generated samples.

Semantics transformations. Semantics transformations (Mohapatra et al. 2020) can manipulate certain kinds of the semantics of an image, such as changing hue and saturation (Hosseini and Poovendran 2018) or color and texture (Bhattad et al. 2020). Semantics transformations are used to produce "unrestricted" perturbations (Bhattad et al. 2020) in adversarial samples, which are traditionally generated by finding imperceptible perturbations under a norm ball constraint (Bhattad et al. 2020). However, these methods cannot be directly adopted in our problem setting since they focus on performing adversarial attacks, while our goal of using semantics transformations is to improve a model's SDG performance. Semantics transformations have also been used to improve few-shot generalization (Zheng et al. 2023) via meta-learning (Finn, Abbeel, and Levine 2017; Zheng and Zhang 2022). A parallel work (Gokhale et al. 2022) uses a pre-defined set of linguistic transformations, such as negation and paraphrasing, to augment text data for improved vision-language inference performance. However, these transformations do not have learnable parameters and cannot be fine-tuned into different ones.

AdvST: Adversarial Learning With Semantics Transformations

Semantics Transformation

We define a semantics transformation as a composition of several standard data augmentation functions that manipulate certain kinds of semantics of a sample. For example, we can perturb both the hue and brightness of an image x with $\tau(x;\omega) = o_h(o_b(x;\omega_b);\omega_h)$, where o_h is the function that changes the hue of x, o_b changes the brightness of x, and $\omega = \omega_b \cup \omega_h$ denotes the set of parameters for τ. We construct a set of M semantics transformations $\mathcal{T} = \{\tau_i(\cdot;\omega_i), i = 1,\dots,M\}$ by randomly composing $L(1 \le L \le L_{\max})$ unique standard data augmentation functions (see Appendix).

Intuitively, a semantics transformation with a large L can produce more diverse samples. However, depending on the target domain data, the semantics transformations that

produce more diverse samples are not necessarily better than those producing less diverse ones. Since we have no knowledge about target domains in SDG, we first uniformly choose the length for semantics transformations and then uniformly choose a semantics transformation with the selected length. Thus, we derive the distribution over M semantics transformations as $G(\tau^L) = \frac{1}{M_L L_{\max}}$, where τ^L denotes a semantics transformation with L standard augmentations, $L_{\max}$ is the maximum number of standard augmentations in τ^L, and M_L is the total number of τ^L and satisfies $M = \sum_{L=1}^{L_{\max}} M_L$.

Learning Objective of AdvST

SDG aims to train a model that is robust to unseen domain shifts with the training samples from a single source domain. The robustness of the trained model to unseen domain shifts depends on how much the training data covers target domains. Therefore, with semantics transformations, we aim to generate new data samples that have large domain shifts from the source domain, increasing the chance of covering data samples from unseen target domains.

A key property of the samples from target domains is that they often yield a high average prediction loss because of their large domain shifts from the source. This motivates AdvST, an adversarial learning framework that learns semantics transformations to generate challenging samples with significant semantics variations for model training.

Given a model f_θ with parameters θ, a set of source domain samples $\mathcal{D}_S = \{(x_n, y_n)\}_{n=1}^N$ with N pairs of training sample x_n and its label y_n, and a distribution G over a set of M semantics transformations $\{\tau_i(\cdot;\omega_i)\}_{i=1}^M$, we express the learning objective of AdvST as:

$$\theta^* = \min_{\theta\in\Theta} \max_{\psi\in\Psi} \mathop{\mathbb{E}}_{\tau\sim G} \mathop{\mathbb{E}}_{\xi\sim\mathcal{D}_S} \big[\ell(\theta;\xi') - \lambda d_\theta(\xi',\xi)\big], \quad (1)$$

where $\xi = (x, y)$ denotes a tuple of a sample x and its label y, $\xi' = (\tau(x;\omega), y)$ is the tuple of the same label y and a new sample obtained by applying the semantics transformation τ to x, $\ell(\theta;\xi)$ is the prediction loss for $\xi = (x, y)$, Θ denotes the set of all possible values of θ, $\psi = \cup_{i=1}^M \omega_i$ denotes the union of the parameters of M semantics transformations, Ψ is the set of all possible values of ψ, λ is a nonnegative regularization parameter, and d_θ is the squared Euclidean distance function between ξ and ξ' in the deep feature space of the model f_θ, i.e., $d_\theta(\xi,\xi') = \|v - v'\|_2^2$ with the embeddings v and v' of x and x', respectively.

The objective in (1) aims to train a robust model with the challenging samples generated from the samples in the source domain while maintaining the core features of the original data. The novel part of (1) is that instead of generating images in the pixel space, we adversarially learn the parameters of semantics transformations, exploiting the domain knowledge in standard data augmentations to generate diverse images.

Learning Algorithm

We adopt an iterative optimization algorithm (Volpi et al. 2018; Zhao et al. 2020) to solve (1). Specifically, the algorithm consists of a minimization and a maximization optimization procedures.

Algorithm 1: Adversarial learning with semantics transformations (AdvST)

Input: Source dataset $\mathcal{D}_S$, extended training set $\mathcal{D}$ with K domains, distribution over M semantics transformations G, initial model weights θ_0, number of training epochs E, batch size B, number of batches per epoch N_B, and number of updates in the maximization procedure $T_{\max}$
Output: learned weights θ

$\theta \leftarrow \theta_0$, $\mathcal{D}$.add($\mathcal{D}_S$)
for $e = 1, \ldots, E$ **do**
  //Minimization procedure
  for $b = 1, \cdots, N_B$ **do**
    Get a batch of B samples $\mathcal{B}$ from $\mathcal{D}$
    Update θ with Eq. (4)
  end for
  //Maximization procedure
  Initialize an empty $\mathcal{D}_e$
  for $(x_n, y_n) \in \mathcal{D}_S$ **do**
    Sample τ from G and initialize its parameters ω_n^0
    for $t = 1, \cdots, T_{max}$ **do**
      Generate a sample $x_n^t = \tau(x_n;\omega_n^{t-1})$
      Update ω_n^t with Eq. (3)
    end for
    Append $(\tau(x_n, \omega_n^{T_{\max}}), y_n)$ to $\mathcal{D}_e$
  end for
  $\mathcal{D}$.add($\mathcal{D}_e$)
end for
return θ

Maximization procedure. We generate worst-case samples via optimized semantics transformations. Specifically, we first sample a semantics transformation τ from G. Then, we sample an example x_n from $\mathcal{D}_S$. We solve the inner maximization problem in (1) by applying $T_{\max}$ steps of stochastic gradient ascent to the parameters of the sampled semantics transformation τ. To facilitate generating diverse samples, we add a maximum entropy regularizer (Zhao et al. 2020) during the optimization. In the tth ($1 \le t \le T_{\max}$) iteration, we have the following steps:

$$x_n^t = \tau(x_n;\omega_n^{t-1}) \quad (2)$$

$$\omega_n^t = \omega_n^{t-1} + \beta\nabla_{\omega_n^{t-1}}\Big(\ell(\theta;x_n^t,y_n) - \lambda d_\theta((x_n^t,y_n),(x_n,y_n)) + \epsilon l_{ent}(\theta;x_n^t,y_n)\Big), \quad (3)$$

where ω_n^t denotes the learnable parameters of τ for the n-th data sample at iteration t, $l_{ent}(\theta;x_n^t,y_n)$ is an entropy regularization term (see Appendix) to further promote learning diverse samples, ϵ is a nonnegative regularization parameter, and β denotes the learning rate in this procedure. We repeat the above steps until all samples in $\mathcal{D}_S$ have been processed. The synthetic data points $\{(\tau(x_n;\omega_n^{T_{\max}}), y_n)\}_{n=1}^N$ are treated as a new domain of data. We add these generated samples to the extended training set denoted as $\mathcal{D}$, which is initialized as $\mathcal{D}_S$.

Minimization procedure. We use samples generated from the maximization step to train a robust model θ against unseen distribution shifts. To avoid model forgetting, at each

iteration, we sample a batch of B samples $\mathcal{B}$ from the extended training set $\mathcal{D}$ to also use previously generated samples. We add a regularizer $\ell_{reg}(\theta;\mathcal{B})$ consisting of contrastive and entropy loss terms (see Appendix) to facilitate learning robust representations. At each iteration, we update the model parameters θ using mini-batch stochastic gradient descent as follows

$$\theta \leftarrow \theta - \alpha\nabla_\theta\Big(\frac{1}{B}\sum_{(x,y)\in\mathcal{B}}\ell(\theta;x,y)+\ell_{reg}(\theta;\mathcal{B})\Big), \quad (4)$$

where "$\leftarrow$" denotes value assignment, and α denotes the learning rate. The complete algorithm is shown in Algorithm 1. We further analyze the space and time complexities of the algorithm with practical considerations in the following.

Space complexity. In the iterative optimization, we keep adding the generated samples to the extended training set $\mathcal{D}$. The size of $\mathcal{D}$ increases with the iteration number, which is not scalable when the initial training set $\mathcal{D}_S$ or the iteration number is large. *Therefore, we implement $\mathcal{D}$ as a domain pool that only stores the generated samples from the most recent K runs of the maximization procedures.* In practice, depending on the size of $\mathcal{D}_S$, we set K in the range of 2 to 5 to ensure that we have sufficient samples for training without incurring the scalability issue.

Time complexity. The time complexity of each iteration of the optimization is $N_BC_\mu + T_{\max}C_GN_B$, where C_μ denotes the complexity of updating the model, C_G denotes the complexity of updating the parameters of semantics transformations, and N_B denotes the number of training batches. Generally, we have $C_G \approx C_\mu$ because C_G and C_μ both include back-propagating the gradients throughout the whole model, and the number of parameters in semantics transformations is negligible compared to the number of model parameters. Therefore, the time complexity is $O(ET_{\max}N_B)$, where E is the total iterations (epochs). In practice, to reduce the impact of $T_{\max}$, we could perform the maximization on different batches in parallel or do early stopping when the difference in loss between consecutive maximization steps is lower than a given threshold.

Theoretical Analysis: Connection to DRO

DRO Formulation

The learning objective of SDG can be expressed via DRO (Gao and Kleywegt 2022) since it does not rely on the notion of a known target distribution. Specifically, DRO chooses a set of probability distributions $\mathcal{U}$ called uncertainty set, and then finds a decision θ from Θ that provides the best hedge against $\mathcal{U}$ by solving the following mini-max problem:

$$\min_{\theta\in\Theta}\max_{Q\in\mathcal{U}}\mathbb{E}_{\xi\sim Q}[\ell(\theta;\xi)], \quad (5)$$

where $\ell(\theta;\xi)$ is the prediction loss with the data-label pair $\xi=(x,y)$, Θ denotes the set of all possible model parameters, and $\mathcal{U}$ contains distributions that are at most δ-distance away from the source distribution P. The uncertainty set, $\mathcal{U}=\{Q|D(P,Q)<\delta\}$, depends on a distance metric $D(\cdot,\cdot)$ and a predefined threshold $\delta>0$. The objective in (5) finds an optimized model under the worst-case distribution Q^* found in $\mathcal{U}$ that maximizes the prediction loss.

Semantics-Induced Distribution

Given a set of M semantics transformations, a semantics-induced distribution $Q_\psi(\xi')$ is defined as follows

$$Q_\psi(\xi')=\sum_{\tau_i}G(\tau_i)\int_\xi p(\xi'|\tau_i,\xi,\omega_i)dP, \quad (6)$$

where $\xi'=(x',y')$, $\xi=(x,y)$ is a sample from the source distribution P, $\psi=\cup_{i=1}^M\omega_i$ denotes the parameters of M semantics transformations, and $p(\xi'|\tau_i,\xi,\omega_i)$ is the probability of obtaining ξ' from ξ and the ith semantics transformation τ_i with parameters ω_i. We require that transformed samples are still assigned with their original labels. Therefore, we have $p(\xi'|\tau_i,\xi,\omega_i)=0$ if $y'\neq y$. Moreover, if τ_i is a deterministic transformation, then $p(\xi'|\tau_i,\xi,\omega_i)=1$ when $\tau_i(x;\omega_i)=x'$ and $y'=y$ and $p(\xi'|\tau_i,\xi,\omega_i)=0$ otherwise. If τ_i is a stochastic transformation, then $p(\xi'|\tau_i,\xi,\omega_i)$ follows the distribution of $\tau_i(x;\omega_i)$. A sample ξ' from Q_ψ can be obtained by first sampling ξ from P with $y=y'$ and τ_i from G, and then obtaining $x'=\tau_i(x;\omega_i)$. Given G and P, Q_ψ fully depends on ψ. We denote the set of all semantics-induced distributions as $\mathcal{Q}_\Psi=\{Q_\psi|\psi\in\Psi\}$, where Ψ is the set of all possible parameters ψ.

Uncertainty Set of AdvST

The uncertainty set of AdvST consists of semantics-induced distributions Q_ψ around the source distribution P to simulate unseen target distributions. These distributions should not deviate too much from the source to avoid hedging against noisy distributions that are not learnable. Hence, we need a proper distance metric $D(\cdot,\cdot)$ to control the distribution shifts. Since semantics transformations create new data samples, we use Wasserstein distances (Definition 1) as the metric D to allow a data distribution Q_ψ to have a different support from that of P.

Definition 1 *(Wasserstein distances (Chen and Paschalidis 2021; Rahimian and Mehrotra 2019; Kuhn et al. 2019)) Let Ξ be a measurable space. Given a transportation cost function $c:\Xi\times\Xi\to[0,\infty)$, which is nonnegative, lower semi-continuous, and satisfies $c(\xi,\xi)=0$, for probability measures Q and P on Ξ, the Wasserstein distance between Q and P is*

$$W_c(Q,P)=\inf_{J\in\prod(Q,P)}\mathbb{E}_{(\xi,\xi')\sim J}[c(\xi,\xi')], \quad (7)$$

where $\prod(Q,P)$ denotes all joint distributions with marginal distributions being P and Q.

We define the transportation cost function c in the deep feature space (Zhao et al. 2020; Volpi et al. 2018) to include distributions whose samples have large style variations since these samples may still be close to the samples from the source distribution in the deep feature space. To exclude noisy distributions whose data samples change their original labels in the source domain after transformations, we design the cost of moving a source distribution sample to such a sample as infinity. Specifically, the cost function of moving $\xi=(x,y)\sim P$ to $\xi'=(x',y')\sim Q_\psi$ given the model θ is defined as follows

$$c_\theta((x,y),(x',y')):=||v-v'||_2^2+\infty\cdot 1\{y\neq y'\}, \quad (8)$$

where v and v' are the model-dependent embeddings for x and x', respectively. Therefore, the uncertainty set that we consider in AdvST is

$$\mathcal{U}_\Psi = \{Q | Q \in \mathcal{Q}_\Psi, W_c(Q, P) < \delta\}, \tag{9}$$

where δ ($\delta > 0$) denotes the predefined distance threshold between the source P and the semantics-induced distributions $\mathcal{Q}_\Psi$.

DRO Learning Objective for AdvST

Directly solving Eq. (5) with $\mathcal{U} = \mathcal{U}_\Psi$ is intractable since it requires searching over the infinite dimension space of distribution functions. We consider the following Lagrangian relaxation with the penalty parameter λ:

$$\min_{\theta \in \Theta} \max_{Q \in \mathcal{Q}_\Psi} \{\mathbb{E}_{(x,y)\sim Q}[\ell(\theta; x, y)] - \lambda W_c(Q, P)\}. \tag{10}$$

However, Eq. (10) is still hard to compute. For the inner maximization term of Eq. (10), Proposition 1 provides a tractable form which only requires the source distribution P and the distribution over semantics transformations G.

Proposition 1 *Let $\ell : \Theta \times \mathcal{X} \times \mathcal{Y} \to [0, \infty)$ denote the loss function which is upper semi-continuous and integrable. The transportation cost function $c : \Xi \times \Xi \to [0, \infty)$ with $\Xi = \mathcal{X} \times \mathcal{Y}$ is a lower semi-continuous function satisfying $c(\xi, \xi) = 0$ for $\xi \in \Xi$. Let G denote the distribution over M semantics transformations $\{\tau_i | i = 1, \ldots, M\}$. Then, for any given P and $\lambda \geq 0$, it holds that*

$$\begin{aligned} &\sup_{Q \in \mathcal{Q}_\Psi} \{\mathbb{E}_Q[\ell(\theta; x, y)] - \lambda W_c(Q, P)\} \\ &= \mathbb{E}_{\tau_i \sim G} \mathbb{E}_P \big[\sup_{\xi \in \Xi_i} (\ell(\theta; \xi) - \lambda c_\theta(\xi, (x, y)))\big]. \end{aligned} \tag{11}$$

where $\mathcal{Q}_\Psi$ is a set of distributions induced by M semantics transformations parameterized by ψ, $\Xi_i = \{(x', y) | x' = \tau_i(x; \omega_i), \xi \in \Xi_0, \omega_i \subset \psi\}$, and $\Xi_0 \subseteq \Xi$ is the support of P.

The proof of Proposition 1 (see Appendix) includes taking the dual reformulation of Eq. (10) and considering a semantics-induced distribution Q as a mixture of M distributions. We observe that the objective in (1) actually minimizes the empirical version of (11) with P and c_θ being replaced by $\mathcal{D}_S$ and d_θ, respectively.

Experiment

Experimental Settings

Datasets. We use the following three benchmark datasets in the experiments and arrange them in increasing order of difficulty. (1) **Digits** is used for digit classification and contains five datasets: MNIST (LeCun et al. 1998), MNIST-M (Ganin and Lempitsky 2015), SVHN (Netzer et al. 2011), SYN (Ganin and Lempitsky 2015), and USPS (Denker et al. 1989). Each dataset has the same 10 digits ranging from 0 to 9. We use MNIST as the source domain and the other four as the test domains. (2) **PACS** (Li et al. 2017) is a collection of four domains, namely, Art, Cartoon, Photo and Sketch. The four domains share seven common object categories and differ in the styles of their images. We use one domain as the source domain and the other three as the unseen target domains. (3) **DomainNet** (Peng et al. 2019) is a large-scale dataset which has 345 object classes and contains six domains, namely Real, Infograph, Clipart, Painting, Quickdraw, and Sketch. We use Real as the source domain and the remaining five as the test domains. This is the most challenging dataset in our experiments due to the large number of classes and the high variability of domains.

Config.	Semantics	Contrastive	Entropy	Avg.
1				59.3±1.5
2	✓			77.0±0.4
3	✓	✓		77.8±0.2
4	✓		✓	78.8±0.2
5	✓	✓	✓	80.0±0.4

Table 1: Ablation study on the Digits dataset. We report average classification accuracy over the four target domains.

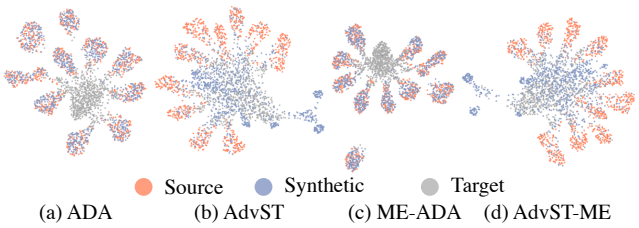
AdvST implementations. We used 12 standard augmentations commonly used in image transformations (see Appendix), such as Rotate and Translate, to construct semantics transformations. Most augmentation functions have specific learnable parameters controlling the magnitude of the transformations. We designed a semantics transformation as a composition of at most $L_{\max} = 3$ standard augmentations since more augmentations bring marginal gains. We used the differentiable library (Riba et al. 2020) to implement these transformations. We denote our method as **AdvST** when $\epsilon = 0$ in Eq. (3) and **AdvST-ME** when $\epsilon > 0$.

Training details. (1) Experiments on **Digits**: we adopted the LeNet (LeCun et al. 1998) as the backbone and used the first $10,000$ images in MNIST to train the model. All images are resized to 32×32 and converted to RGB images. For our method, we set $E = 50$, $T_{\max} = 20$, $\lambda = 100$, $\beta = 0.2$, $B = 32$, and $\alpha = 1 \times 10^{-4}$ which is dropped by 0.1 after 25 epochs. (2) Experiments on **PACS**: we used a ResNet-18 (He et al. 2016) pre-trained on ImageNet as the backbone and fine-tuned it on the source domain. All images are resized to 224×224. We set $E = 50$, $T_{\max} = 50$, $\lambda = 10$, $\beta = 5.0$, $B = 32$, and $\alpha = 0.001$ which decays following a cosine annealing scheduler. (3) Experiments on **DomainNet**: we used the same backbone as for the PACS datasets. We set $E = 200$, $T_{\max} = 50$, $\lambda = 10$, $\beta = 1.0$, $B = 128$, and $\alpha = 0.001$ which decays following a cosine annealing scheduler. We did not specifically tune hyperparameters as our method is robust to them as long as the training converges. Additional training details are in Appendix.

We ran our experiments on Nvidia Quadro RTX 8000 GPUs. We ran our experiments 5 times with different random seeds and reported the average accuracy with standard deviation.

Ablation Studies

We conducted ablation studies on AdvST-ME using the Digits dataset. We evaluated how semantics transformations (Semantics), the contrastive regularizer (Contrastive), and the entropy regularizer (Entropy) affect the average SDG performance. We observe from Table 1 that semantics trans-



(a) ADA (b) AdvST (c) ME-ADA (d) AdvST-ME

Figure 1: Visualization of how samples from the source domain, target domains, and synthetic domains distribute in the embedding space. We compare AdvST and AdvST-ME with their non-semantics counterparts ADA and ME-ADA.

formations can significantly boost the average classification accuracy by 17.7% (Configurations 1 and 2). The contrastive and entropy regularizers can further boost the performance of Configuration 1 by 0.8% (Configuration 3) and 1.8% (Configuration 4), respectively. Our method (Configuration 5) achieves the highest average classification accuracy with all three components.

We further compared the coverage of generated samples on target domain data between our methods, AdvST and AdvST-ME, and their pixel-level counterparts, ADA (Volpi et al. 2018) and ME-ADA (Zhao et al. 2020), which directly generate images in the pixel space. We visualized how the samples generated by ADA, ME-ADA, and our methods distribute in the embedding space in Figure 1. We color the samples from the source domain MNIST orange and the samples from the four target domains gray. We give details for obtaining the figure in Appendix. From Figure 1(a) and (c), we observe that most of the synthetic samples distribute very close to the source domain data and have little coverage on the target domains. In contrast, the synthetic samples in Figure 1(b) and (d) deviate from the source domain and have broad coverage on the target domains.

We provide analyses on the sensitivity of λ and the effect of different semantics transformations in Appendix.

Comparison on Digits

Baselines. We included ADA, ME-ADA, and the following methods for comparison: ERM, which trains a model only using the standard cross-entropy loss; CCSA (Motiian et al. 2017), which aligns samples from different domains to improve generalization; d-SNE (Xu et al. 2019), which minimizes the maximum distance between sample pairs of the same class and maximizes the minimum distance among sample pairs of different categories; JiGen (Carlucci et al. 2019), which is a multi-task learning method that combines the target recognition task and the Jigsaw classification task; M-ADA (Qiao, Zhao, and Peng 2020), which uses generative models and meta-learning (Finn, Abbeel, and Levine 2017; Chen and Zhang 2021, 2022) to improve ADA; AutoAug (Cubuk et al. 2018) and RandAug (Cubuk et al. 2020b), which augment data based on the searched augmentation policies; RSDA (Volpi and Murino 2019), which randomly searches image transformations to train a robust model; and PDEN (Li et al. 2021) and L2D (Wang et al. 2021), which use generative models for data augmentation.

Method	SVHN	MNIST-M	SYN	USPS	Avg.
ERM	27.8	52.7	39.7	76.9	49.3
CCSA	25.9	49.3	37.3	83.7	49.1
d-SNE	26.2	51.0	37.8	93.2	52.1
JiGen	33.8	57.8	43.8	77.2	53.1
ADA	35.5	60.4	45.3	77.3	54.6
ME-ADA	42.6	63.3	50.4	81.0	59.3
M-ADA	42.6	67.9	49.0	78.5	59.5
AutoAug	45.2	60.5	64.5	80.6	62.7
RandAug	54.8	74.0	59.6	77.3	66.4
RSDA	47.7	81.5	62.0	83.1	68.5
L2D	62.9	**87.3**	63.7	84.0	74.5
PDEN	62.2	82.2	69.4	85.3	74.8
AdvST	**67.5±0.7**	79.8±0.7	**78.1±0.9**	94.8±0.4	**80.1±0.5**
AdvST-ME	66.7±1.0	80.0±0.5	77.9±0.7	**95.4±0.4**	80.0±0.4

Table 2: Classification accuracy (%) results on the four target domains SVHN, MNIST-M, SYN, and USPS, with MNIST as the source domain. Best results are in bold font.

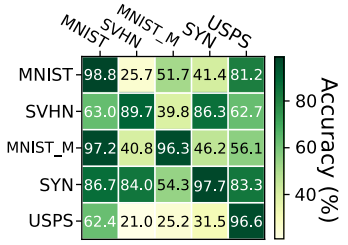


Figure 2: Accuracy heatmap for models trained individually on the five domains from the Digit dataset using ERM.

Results. We observe from Table 2 that our methods, AdvST and AdvST-ME, significantly improve the performance of the pixel-level adversarial data augmentations, ADA and ME-ADA, across the four target domains and achieve a maximum gain of 25.5% in average classification accuracy. Regarding per-domain performance, our methods achieve the best performance on all the target domains except the MNIST-M domain. It is common to observe that a method does not perform the best on all the target domains. For example, PDEN performs better than L2D on SYN but worse than L2D on MNIST-M. We reason that the knowledge that helps a model generalize in one domain does not necessarily work for the other. To demonstrate this, we trained models on one of the five domains and evaluated their generalization performance on each of the remaining domains. From the accuracy heatmap in Figure 2, we see that the learned knowledge for MNIST-M cannot transfer well to SYN and vice versa, which explains the performance tradeoff between MNIST-M and SYN when comparing AdvST with AdvST-ME or AdvST with L2D. Nevertheless, our methods achieve the best average classification accuracy over the four target domains among all the methods.

Comparison on PACS

Baselines. We compared our methods AdvST and AdvST-ME with ADA, ME-ADA, MixUp (Zhang et al. 2018), CutOut (DeVries and Taylor 2017), CutMix (Yun et al.

Target	MixUp	CutOut	ADA	ME-ADA	AugMix	RandAug	ACVC	L2D	AdvST	AdvST-ME
Art	52.8	59.8	58.0	60.7	63.9	67.8	67.8	67.6	**69.2±1.4**	67.0±1.1
Cartoon	17.0	21.6	25.3	28.5	27.7	28.9	30.3	42.6	**55.3±2.0**	53.2±1.1
Sketch	23.2	28.8	30.1	29.6	30.9	37.0	46.4	47.1	**67.7±1.5**	67.2±2.2
Avg.	31.0	36.7	37.8	39.6	40.8	44.6	48.2	52.5	**64.1±0.4**	62.5±0.8

Table 3: Classification accuracy (%) comparison on the PACS dataset. Best results are in bold font.

Target	MixUp	CutOut	CutMix	ADA	ME-ADA	RandAug	AugMix	ACVC	AdvST	AdvST-ME
Painting	38.6	38.3	38.3	38.2	39.0	41.3	40.8	41.3	42.3±0.1	**42.4±0.2**
Infograph	13.9	13.7	13.5	13.8	14.0	13.6	13.9	12.9	14.8±0.2	**14.9±0.1**
Clipart	38.0	38.4	38.7	40.2	41.0	41.1	41.7	**42.8**	41.5±0.4	41.7±0.2
Sketch	26.0	26.2	26.9	24.8	25.3	30.4	29.8	30.9	30.8±0.3	**31.0±0.2**
Quickdraw	3.7	3.7	3.6	4.3	4.3	5.3	6.3	**6.6**	5.9±0.2	6.1±0.2
Avg.	24.0	24.1	24.2	24.3	24.7	26.3	26.5	26.9	27.1±0.2	**27.2±0.1**

Table 4: Classification accuracy (%) comparison on the DomainNet dataset. Best results are in bold font.

2019), RandAug (Cubuk et al. 2020a), AugMix (Hendrycks et al. 2019), and L2D (Wang et al. 2021). We also included ACVC (Cugu et al. 2022), which applies attention consistency to learning from augmented samples.

Results. We used Photo as the source domain and evaluated models on the Art, Cartoon, and Sketch domains. Generalizing raw images to artificial images is the most challenging SDG setting in the PACS dataset since the domain shift between the source and target domains is substantial. Results in Table 3 show that our methods significantly improve the performance of pixel-level adversarial data augmentations, ADA and ME-ADA, in all three domains. Moreover, our method AdvST performs the best on the three target domains and achieves the best average classification accuracy over the three domains. AdvST-ME performs the second best in this setting, indicating that maximizing output entropy to further encourage generating diverse samples does not help the generalization from a natural domain to an artificial one.

Comparison on DomainNet

Baselines. We compared our methods AdvST and AdvST-ME with ADA, ME-ADA, MixUp (Zhang et al. 2018), CutOut (DeVries and Taylor 2017), CutMix (Yun et al. 2019), RandAug (Cubuk et al. 2020a), and AugMix (Hendrycks et al. 2019).

Results. Table 4 shows our results in the most challenging SDG setting, DomainNet, which has 345 classes and significant domain shifts from the source domain, such as Real to Infograph and Real to Quickdraw. Under this challenging setting, our methods outperforms pixel-level adversarial data augmentations, ADA and ME-ADA, and complex data augmentations, such as RandAug and AugMix.

Learning With Limited Source Data

We further demonstrated the utility of our methods by evaluating the average classification accuracy of our methods on target domains with limited training data. We used the Art dataset from PACS as the source domain and the remaining three datasets in PACS as the target domains. We used partial training data of the Art domain and reported the average classification accuracy over the three target domains in

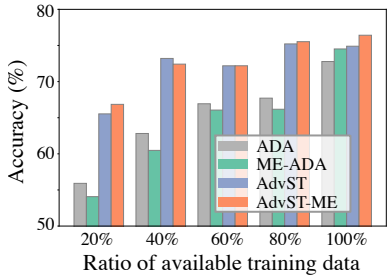


Figure 3: Average classification accuracy under different ratios of available training data.

Figure 3. We observe that under different ratios of available training data, our methods, AdvST and AdvST-ME, consistently outperform ADA and ME-ADA, respectively. The gains are significant when the ratio is small, demonstrating the effectiveness of our method when there is a lack of available training data.

Conclusion

We revisited data augmentation for SDG and focused on leveraging the domain knowledge in standard data augmentations. We conceptualized a composition of several standard data augmentations as a semantics transformation with learnable parameters and proposed AdvST, an adversarial learning framework that aims to train a robust model with diverse samples generated by semantics transformations. We theoretically showed that AdvST optimizes a DRO objective with semantics-induced distributions. Although built on standard data augmentations, AdvST is surprisingly competitive. It achieves the best average domain generalization performance on three benchmark datasets and is effective with limited source data. A promising future improvement is to expand the pool of standard data augmentations and selectively choose augmentations given the partial knowledge of target domain data, such as style descriptions.

Acknowledgments

This work is supported in part by the US National Science Foundation under grants 2217071, 2213700, 2106913, 2008208, 1955151.

References

Balaji, Y.; Sankaranarayanan, S.; and Chellappa, R. 2018. Metareg: Towards domain generalization using meta-regularization. *Advances in Neural Information Processing Systems*, 31: 998–1008.

Bhattad, A.; Chong, M. J.; Liang, K.; Li, B.; and Forsyth, D. A. 2020. Unrestricted Adversarial Examples via Semantic Manipulation. In *International Conference on Learning Representations*.

Blanchet, J.; and Murthy, K. 2019. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2): 565–600.

Carlucci, F. M.; D'Innocente, A.; Bucci, S.; Caputo, B.; and Tommasi, T. 2019. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2229–2238.

Chen, J.; and Zhang, A. 2021. Hetmaml: Task-heterogeneous model-agnostic meta-learning for few-shot learning across modalities. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 191–200.

Chen, J.; and Zhang, A. 2022. Topological transduction for hybrid few-shot learning. In *Proceedings of the ACM Web Conference 2022*, 3134–3142.

Chen, R.; and Paschalidis, I. C. 2021. Distributionally robust learning. *arXiv preprint arXiv:2108.08993*.

Cubuk, E. D.; Zoph, B.; Mane, D.; Vasudevan, V.; and Le, Q. V. 2018. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*.

Cubuk, E. D.; Zoph, B.; Shlens, J.; and Le, Q. V. 2020a. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.

Cubuk, E. D.; Zoph, B.; Shlens, J.; and Le, Q. V. 2020b. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 702–703.

Cugu, I.; Mancini, M.; Chen, Y.; and Akata, Z. 2022. Attention consistency on visual corruptions for single-source domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 4165–4174.

Denker, J. S.; Gardner, W.; Graf, H. P.; Henderson, D.; Howard, R. E.; Hubbard, W.; Jackel, L. D.; Baird, H. S.; and Guyon, I. 1989. Neural network recognizer for handwritten zip code digits. In *Advances in Neural Information Processing Systems*, 323–331. Citeseer.

DeVries, T.; and Taylor, G. W. 2017. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, volume 70, 1126–1135.

French, G.; Mackiewicz, M.; and Fisher, M. 2017. Self-ensembling for visual domain adaptation. *arXiv preprint arXiv:1706.05208*.

Ganin, Y.; and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, 1180–1189. PMLR.

Gao, R.; and Kleywegt, A. 2022. Distributionally robust stochastic optimization with Wasserstein distance. *Mathematics of Operations Research*.

Gokhale, T.; Chaudhary, A.; Banerjee, P.; Baral, C.; and Yang, Y. 2022. Semantically distributed robust optimization for vision-and-language inference. In *Findings of the Association for Computational Linguistics: ACL 2022*, 1493–1513.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Hendrycks, D.; Mu, N.; Cubuk, E. D.; Zoph, B.; Gilmer, J.; and Lakshminarayanan, B. 2019. AugMix: A simple data processing method to improve robustness and uncertainty. In *International Conference on Learning Representations*.

Hosseini, H.; and Poovendran, R. 2018. Semantic adversarial examples. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1695–16955.

Kuhn, D.; Esfahani, P. M.; Nguyen, V. A.; and Shafieezadeh-Abadeh, S. 2019. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations Research & Management Science in the Age of Analytics*, 130–166. INFORMS.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324.

Li, D.; Yang, Y.; Song, Y.-Z.; and Hospedales, T. M. 2017. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE International Conference on Computer Vision*, 5542–5550.

Li, D.; Yang, Y.; Song, Y.-Z.; and Hospedales, T. M. 2018a. Learning to generalize: Meta-learning for domain generalization. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Li, D.; Zhang, J.; Yang, Y.; Liu, C.; Song, Y.-Z.; and Hospedales, T. M. 2019. Episodic training for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1446–1455.

Li, L.; Gao, K.; Cao, J.; Huang, Z.; Weng, Y.; Mi, X.; Yu, Z.; Li, X.; and Xia, B. 2021. Progressive domain expansion network for single domain generalization. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 224–233.

Li, Y.; Wang, N.; Shi, J.; Hou, X.; and Liu, J. 2018b. Adaptive batch normalization for practical domain adaptation. *Pattern Recognition*, 80: 109–117.

Lian, Q.; Ye, B.; Xu, R.; Yao, W.; and Zhang, T. 2021. Geometry-aware data augmentation for monocular 3D object detection. *arXiv preprint arXiv:2104.05858*.

Mohapatra, J.; Weng, T.-W.; Chen, P.-Y.; Liu, S.; and Daniel, L. 2020. Towards verifying robustness of neural networks against a family of semantic perturbations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 244–252.

Motiian, S.; Piccirilli, M.; Adjeroh, D. A.; and Doretto, G. 2017. Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE International Conference on Computer Vision*, 5715–5725.

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; and Ng, A. Y. 2011. Reading digits in natural images with unsupervised feature learning.

Pandey, P.; Raman, M.; Varambally, S.; and AP, P. 2021. Domain generalization via inference-time label-preserving target projections. *arXiv preprint arXiv:2103.01134*.

Peng, X.; Bai, Q.; Xia, X.; Huang, Z.; Saenko, K.; and Wang, B. 2019. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1406–1415.

Qiao, F.; Zhao, L.; and Peng, X. 2020. Learning to learn single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12556–12565.

Rahimian, H.; and Mehrotra, S. 2019. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*.

Riba, E.; Mishkin, D.; Ponsa, D.; Rublee, E.; and Bradski, G. 2020. Kornia: An open source differentiable computer vision library for PyTorch. In *Winter Conference on Applications of Computer Vision*.

Shankar, S.; Piratla, V.; Chakrabarti, S.; Chaudhuri, S.; Jyothi, P.; and Sarawagi, S. 2018. Generalizing across domains via cross-gradient training. *arXiv preprint arXiv:1804.10745*.

Shu, R.; Bui, H. H.; Narui, H.; and Ermon, S. 2018. A dirt-t approach to unsupervised domain adaptation. *arXiv preprint arXiv:1802.08735*.

Volpi, R.; and Murino, V. 2019. Addressing model vulnerability to distributional shifts over image transformation sets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7980–7989.

Volpi, R.; Namkoong, H.; Sener, O.; Duchi, J.; Murino, V.; and Savarese, S. 2018. Generalizing to unseen domains via adversarial data augmentation. In *Advances in Neural Information Processing Systems*, 5339–5349.

Wang, Z.; Luo, Y.; Qiu, R.; Huang, Z.; and Baktashmotlagh, M. 2021. Learning to diversify for single domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 834–843.

Xu, X.; Zhou, X.; Venkatesan, R.; Swaminathan, G.; and Majumder, O. 2019. d-sne: Domain adaptation using stochastic neighborhood embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2497–2506.

Yun, S.; Han, D.; Oh, S. J.; Chun, S.; Choe, J.; and Yoo, Y. 2019. CutMix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018. Mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*.

Zhang, X.; Wang, Q.; Zhang, J.; and Zhong, Z. 2019. Adversarial AutoAugment. In *International Conference on Learning Representations*.

Zhang, Y.; Deng, B.; Li, R.; Jia, K.; and Zhang, L. 2023. Adversarial style augmentation for domain generalization. *arXiv preprint arXiv:2301.12643*.

Zhao, L.; Liu, T.; Peng, X.; and Metaxas, D. 2020. Maximum-entropy adversarial data augmentation for improved generalization and robustness. *Advances in Neural Information Processing Systems*.

Zheng, G.; Suo, Q.; Huai, M.; and Zhang, A. 2023. Learning to learn task transformations for improved few-shot classification. In *Proceedings of the 2023 SIAM International Conference on Data Mining*, 784–792.

Zheng, G.; and Zhang, A. 2022. Knowledge-guided semantics adjustment for improved few-shot classification. In *2022 IEEE International Conference on Data Mining*, 1347–1352. IEEE.

Zhong, Z.; Zhao, Y.; Lee, G. H.; and Sebe, N. 2022. Adversarial style augmentation for domain generalized urban-scene segmentation. *Advances in Neural Information Processing Systems*, 35: 338–350.

Zoph, B.; Cubuk, E. D.; Ghiasi, G.; Lin, T.-Y.; Shlens, J.; and Le, Q. V. 2020. Learning data augmentation strategies for object detection. In *European Conference on Computer Vision*, 566–583. Springer.