Attention Consistency on Visual Corruptions for Single-Source Domain Generalization

Ilke Cugu[1], Massimiliano Mancini[1], Yanbei Chen[1], Zeynep Akata[1,2]
[1]University of Tübingen, [2]MPI for Intelligent Systems
{ilke.cugu, massimiliano.mancini, yanbei.chen, zeynep.akata}@uni-tuebingen.de

Abstract

*Generalizing visual recognition models trained on a single distribution to unseen input distributions (*i.e. *domains) requires making them robust to superfluous correlations in the training set. In this work, we achieve this goal by altering the training images to simulate new domains and imposing consistent visual attention across the different views of the same sample. We discover that the first objective can be simply and effectively met through visual corruptions. Specifically, we alter the content of the training images using the nineteen corruptions of the ImageNet-C benchmark and three additional transformations based on Fourier transform. Since these corruptions preserve object locations, we propose an attention consistency loss to ensure that class activation maps across original and corrupted versions of the same training sample are aligned. We name our model* ***A****ttention* ***C****onsistency on* ***V****isual* ***C****orruptions (ACVC). We show that ACVC consistently achieves the state of the art on three single-source domain generalization benchmarks, PACS, COCO, and the large-scale DomainNet* [1].

1. Introduction

Visual recognition models aim to categorize the semantic content of an image. While existing deep learning methods have achieved impressive results on standard object recognition benchmarks [11, 15], their performance degrades when the test data distribution differs from the training one [34]. This problem, called *domain-shift* [5] is ubiquitous for systems operating in real environments. In fact, since we cannot collect data for every possible change in the input distribution (*e.g.* illumination, background, weather, etc.), we need to develop models that can generalize to *unseen* domains (*i.e.* input distribution) not represented in the training set.

Towards this goal, in this paper, we address the problem of *single-source domain generalization* (single DG), where only a single (source) domain is available for training, and multiple unseen domains are present at test time. This problem is challenging since, contrary to standard domain generalization, we cannot rely on multiple training domains to disentangle domain-specific and domain-invariant information [2, 16, 22, 35]. This led previous approaches to simulate multiple domains via data augmentation and adversarial perturbations [26, 33, 41], using them within standard classification objectives [32, 33, 41], or meta-learning procedures [26].

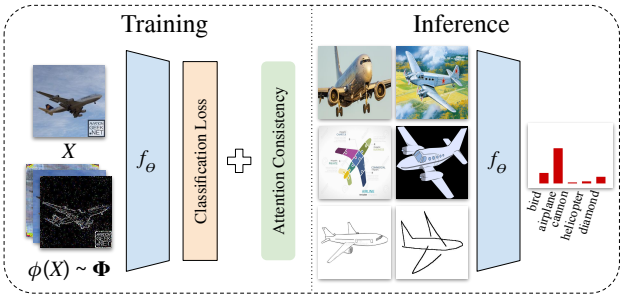

Figure 1. Our approach (1) samples a transformation from a pool of visual corruptions (i.e. $\phi(X) \sim \Phi$) to simulate distinct domains for training, and (2) enforces visual attention consistency between the original and corrupted sample. Once trained, our model is capable of generalizing well to unseen domains.

Training with multiple synthetic domains allows the model to better disentangle domain- and semantic-specific information, eliminating spurious correlations between the model's predictions and the input images. Here we start from the same principle, *i.e.* augmenting data to simulate different training domains. However, we take a step further and we argue that a robust single DG model should provide the same explanation across augmented views of the same training sample. In particular, we compute the model's Class Activation Maps [42] for both the original and augmented samples, imposing consistency among the two (Figure 1). This forces the model to look at the same spatial locations, no matter how different the augmented sample looks like. We found this approach to provide a stronger learning signal

[1]The codes are available at https://github.com/ExplainableML/ACVC

in comparison to alignment on model predictions [13].

Another crucial element of our framework is the data augmentation technique. It should heavily alter the input while not modifying the spatial location of the semantic content. To achieve this, we propose to use visual corruptions. Our idea is that corrupting the images not only creates different input domains, but also produces abundant task-irrelevant visual variations, which together help to prevent the model from memorizing spurious patterns in the training set. We make use of five families of visual corruptions (shown in Figure 2), *i.e*. Weather, Blur, Noise, Digital, and Fourier. The first four groups contain transformations taken from the ImageNet-C [12] benchmark. The last group contains three transformations corrupting the image using the post-Fourier transform components (Figure 2, bottom left) by removing low frequencies, modifying amplitudes, and scaling phases.

To summarize, our contributions are as follows. (1) We analyze the use of visual corruptions as augmentation technique for single DG, using 19 transformations drawn from ImageNet-C and 3 Fourier-based ones. (2) We propose a new consistency loss based on class activation maps, forcing the model to look at the same regions for both the clean and corrupted images (Figure 2, green box)). We name our model **A**ttention **C**onsistency on **V**isual **C**orruptions (ACVC). (3) We propose a new single DG benchmark using three different datasets: PACS [16], COCO [20] and DomainNet [23], that measure generalization performance of models from natural images to other domains; (4) We show that ACVC achieves the state-of-the-art on the proposed single-source DG benchmarks, outperforming information-bottleneck based adversarial (*e.g*. ME-ADA [41]) and advanced data augmentation techniques (*e.g*. MixUp [40], CutMix [39], CutOut [8], RandAugment [7], and AugMix [13]).

2. Related Work

Domain generalization (DG) is the task of learning a model that generalizes to data distributions unseen during training [10,16]. While this problem is usually addressed in the multi-source setting, here we focus on the scenario where only a single domain is available during training [26], *i.e*. single DG. This is challenging since we cannot rely on the presence of multiple training domains to *e.g*. disentangle domain-specific and domain-invariant information [2, 3, 16, 28], or align feature distributions of different domains while preserving their semantics [1, 17, 22, 43]. Typical approaches for single DG simulate the presence of new domains with data augmentation either through adversarial strategies [9, 18, 25, 26, 33, 41] or direct input transformation [32]. For instance, [33] performs adversarial data augmentation under a worst case formulation, assuming samples of unseen domains to be close to the training distribution. [26] relaxes the worst-case formulation of [33] through Wasserstein Auto-Encoders [31], using the augmented domains to perform meta learning. [41] uses information bottleneck (IB) principle [30] to generate adversarial samples far from the source domain. [32] defines new data augmentation rules through an evolutionary strategy, with the fitness measure being the model error.

Differently from these works, we focus on *corruptions* of the input images as transformations. We show that removing information from the data provides better generalization performance than more complex data augmentation schemes. Moreover, we are the first to use visual explanation techniques as consistency loss for DG, enforcing the model to attend to the same regions, regardless the style of the input.

Data augmentation is an effective strategy to improve the generalization of deep neural networks, providing different views of the same input. In computer vision, the most common augmentation strategies are label-preserving transformations such as random flipping, cropping and rotations [4, 15]. Recently, various advanced augmentation techniques have been proposed to further improve representation learning, including CutOut [8], CutMix [39], MixUp [40] and automated augmentation schemes such as AutoAugment [6], RandAugment [7], and AugMix [13]. In these techniques, an input image is often randomly corrupted by mixing with another image (*e.g*. CutMix [39], MixUp [40]) or by random occlusion (*e.g*. CutOut [8]). Such corruptions, however, may destroy the underlying semantics of the input image and even alter its corresponding class label [39, 40]. In automated augmentation, augmentation strategies are either learned w.r.t. the performance on the validation set [6], or randomly selected from a pool [7, 13]. In addition, AugMix [13] uses a Jensen-Shannon divergence loss on model's predictions for the original and augmented images.

In this work, we propose to use a diverse set of visual corruptions randomly selected per image during training. Since our transformations alter neither the semantic of the image nor the location of the objects, we formulate a visual attention consistency loss to encourage the model to look at the same regions for both the original and corrupted versions of a given image.

3. Attention Consistency on Visual Corruptions

We aim to solve the problem of single domain generalization (single DG) where a model is trained on data from a single domain (source) but is expected to generalize to domains unseen during training (target). Formally, we are given a training set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where $x \in \mathcal{X}$ is an image in the space $\mathcal{X}$ and y is its corresponding class label $y \in Y = \{1, \dots, C\}$, with C being the number of classes. We are interested in learning the parameters θ of a function $f_\theta : \mathcal{X} \longrightarrow \mathcal{Y}$ mapping images to probability vectors over the class labels, with $\mathcal{Y}$ being a probability simplex defined over Y. Note that, at test time, we receive images X_t from a new dataset $\mathcal{D}_t$, with a different joint distribution, *i.e*. $p^{\mathcal{D}}_{\mathtt{xy}} \neq p^{\mathcal{D}_t}_{\mathtt{xy}}$, with $\mathtt{x}$ and $\mathtt{y}$ being random variables in $\mathcal{X}$

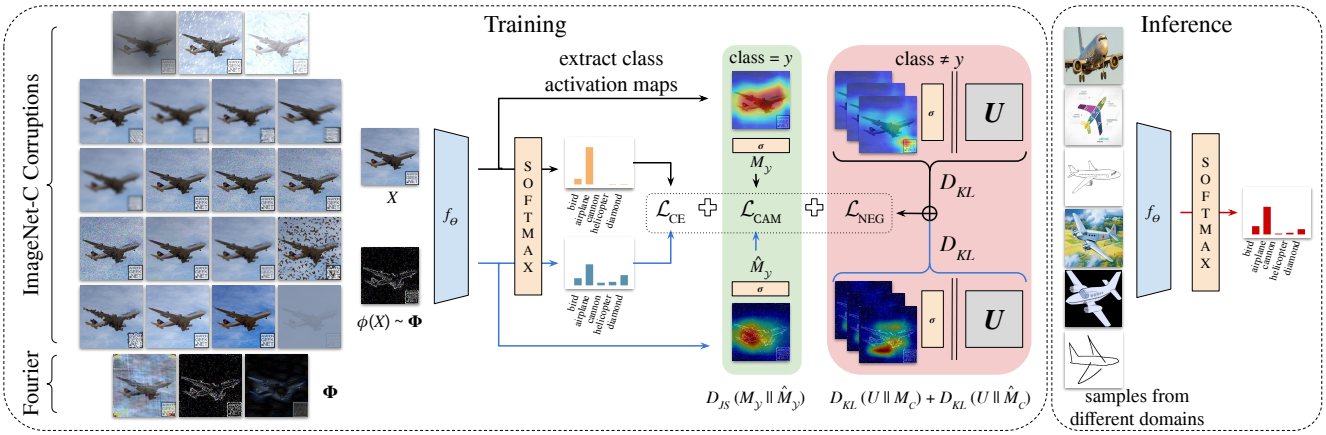

Figure 2. Our ACVC approach (1) randomly samples a corruption ϕ from the set of twenty-two augmentations Φ that consist of ImageNet-C and our Fourier-based corruptions, (2) enforces visual attention consistency between a given model's class activation maps (CAM) for the original M_y and corrupted version $\hat{M}_y$ of a given image X, (3) regularize the CAMs via Negative CAM loss [29] that minimizes the difference between uniform distribution U and top-k negative classes' CAMs $M_{c \in C_k}$.

and Y respectively.

We train our model θ based on two simple principles: 1) simulating the presence of multiple domains via a set of data augmentations, 2) enforcing that the output of the model is consistent across original and simulated domains. Formally, we define our overall learning objective function as:

$$\mathcal{L} = \sum_{(X,y)\in\mathcal{D}} \mathcal{L}_{\text{CE}}(X, \phi(X), y) + \lambda \mathcal{L}_{\text{CON}}(X, \phi(X), y), \quad (1)$$

where ϕ is a label-preserving augmentation function, $\mathcal{L}_{\text{CON}}$ is a consistency term between X and its augmented version $\phi(X)$ given the semantic label y, and λ is a hyperparameter balancing the two loss terms. $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss:

$$\mathcal{L}_{\text{CE}}(X, \hat{X}, y) = -\log f_\theta^y(X) - \log f_\theta^y(\hat{X}), \quad (2)$$

where $f_\theta^y(X)$ is the probability of class y for the input X given by the function f_θ.

The form of ϕ and $\mathcal{L}_{\text{CON}}$ influence the performance of the framework. In this work, we randomly sample ϕ from a larger set Φ composed of visual corruptions, *i.e.* transformations that alter the content of the image while not modifying the location of the object of interest. These corruptions provide large visual variations while being simple and efficient w.r.t. other state-of-the-art alternatives. Since the locations of the objects are preserved, we can implement $\mathcal{L}_{\text{CON}}$ by 1) extracting the spatial regions that most contributed to the prediction and 2) enforcing the model to focus on the same regions, independent of the specific corruption of the input. As we will show experimentally, this supervision is more effective than enforcing consistency on model's predictions.

3.1. Visual Corruptions

Here we describe our set of transformations Φ, merging the ImageNet-C with Fourier transform-based corruptions.

3.1.1 ImageNet-C Visual Corruptions

ImageNet-C [12] is a well-known benchmark to evaluate the robustness of visual models under corruptions [13, 21, 41]. It contains 19 corruptions in total, with 5 severity levels. We argue that corruptions can be used as an augmentation technique to train robust vision models. The corruptions in ImageNet-C are grouped into four categories, *i.e.* Weather, Blur, Noise and Digital (see Figure 2 for examples).

Weather simulates meteorological hurdles such as *fog*, *snow*, *frost* and *spatter* whereas **Blur** smooths the intensities of the image pixels using different functions, such as *gaussian*, *glass*, *motion*, *defocus* and *zoom*. **Noise** perturbates the pixel values randomly, using different functions, *i.e. shot*, *impulse*, *Gaussian* and *speckle* while **Digital** gathers diverse set of corruptions caused by either modifying the image resolution (*i.e. JPEG compression*, *pixelation*, *elastic*) or pixel intensity (*i.e. saturation*, *brightness*, and *contrast*).

3.1.2 Fourier-based Visual Corruptions

Early studies showed how the phase component of Fourier transform of images retains most of the semantic in a scene whereas amplitude focuses on textures [24]. Recent works successfully used this property in domain adaptation [37, 38] and multi-source domain generalization [36]. We thus incorporate three frequency-based corruption methods to our pool of transformations. In the following we use $\mathcal{F}(X)$ to denote the Fourier transform of an image X, with $\mathcal{F}^A(X)$ its amplitude and with $\mathcal{F}^P(X)$ its phase.

Phase Scaling. Given a random scalar $\alpha \in (0, 1]$, this corruption uses α to scale the phase component, computing:

$$\phi_{\text{P-scaling}}(X) = \mathcal{F}^{-1}([\mathcal{F}^A(X), \alpha\mathcal{F}^P(X)]), \quad (3)$$

where $\mathcal{F}^{-1}$ is the inverse Fourier transform. By scaling the phase, we are adding more visual artifacts that will occlude elements of the scene as $\alpha \to 0$ (Fig. 2, first Fourier sample).

Constant Amplitude. This corruption replaces $\mathcal{F}^A$ with a constant $\beta \in (0, 1]$, computing the corrupted image as:

$$\phi_{\text{constant-A}}(X) = \mathcal{F}^{-1}([\beta, \mathcal{F}^P(X)]), \tag{4}$$

Since phase information is preserved, the resulting images are recognizable, but lose most color and texture information (Figure 2, second Fourier sample).

High pass filter. This transformation corrupts the input image with a high pass filter via frequency windows. It filters out low frequency components by adjusting its diameter d on the centered Fourier spectrum. Formally:

$$\phi_{\text{high-pass}}(X) = \mathcal{F}^{-1}(H^d(\mathcal{F}(X)) \circ \mathcal{F}(X))), \tag{5}$$

where $H^d(F)$ a filtering mask where each spatial coordinate (u, v) has value:

$$H^d_{u,v}(F) = \begin{cases} 1, & \text{if} \quad F_{u,v} \geq d \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

This leads to a corrupted image where edges are highlighted and shapes are preserved (Figure 2, third Fourier sample).

3.2. Attention Consistency

Visual corruptions provide powerful augmentations for single DG. However, we argue that a good single DG model should also look at the same image regions, no matter of their particular style. This will allow the model to find consistent visual cues across different versions of the same input, re-using these cues in unseen target domain. In this section, we describe how to use CAMs of original and corrupted images to define a consistency loss term for single DG.

CAM consistency. CAMs [42] provide visual explanations to a given model's predictions by visualizing the spatial regions that most contributed to the output in a given feature map. Let us split f_θ in three components: $g : \mathcal{X} \to \mathcal{Z}$ mapping an image into the feature space $\mathcal{Z} \subset \mathbb{R}^{n \times s}$, an average pooling operation P and a linear classifier $W \in \mathbb{R}^{n \times C}$ followed by softmax. In $\mathcal{Z}$, n, denotes number of channels, and s the spatial locations. Following the formulation of [29], given an input X and we define its set of CAMs as:

$$M = \sigma(W^\intercal g(X)), \tag{7}$$

where $M \in \mathbb{R}^{C \times s}$ and $M_c \in \mathbb{R}^s$ denotes the CAM for class c, corresponding to the c-th row of M. In Eq. (7), σ is a softmax operation with temperature T over the locations:

$$\sigma(x)^c_i = \frac{exp(x^c_i/T)}{\sum_{j=1}^{s} exp(x^c_j/T))}. \tag{8}$$

Algorithm 1 Single DG with ACVC

Require: Training set $\mathcal{D}$, parameters θ, set of corruptions Φ, prediction function f.

1: **for all** $(X, y) \in \mathcal{D}$ **do**
2: Randomly sample a corruption operation ϕ from Φ
3: Apply the transformation to the input: $\hat{X} = \phi(X)$
4: Compute predictions $f_\theta(X)$, $f_\theta(\hat{X})$
5: Compute CAMs M, $\hat{M}$ using Eq.(7)
6: Compute the loss: $\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda(\mathcal{L}_{\text{CAM}} + \mathcal{L}_{\text{NEG}})$
7: Compute the gradient of θ w.r.t. $\mathcal{L}$
8: Update θ
9: **end for**

Given a label y we compute our visual attention consistency loss using Jensen-Shannon divergence as:

$$\mathcal{L}_{\text{CAM}}(M, \hat{M}, y) = D_{JS}(M_y || \hat{M}_y), \tag{9}$$

where $\hat{M}_y$ is the CAM of the corrupted image $\hat{X} = \phi(X)$ for the class y. While Eq. (9) can be replaced by other objectives, such as MSE, we found the Jensen-Shannon divergence (JSD) to work better in practice. Moreover, this formulation allows to define more flexible objectives through the temperature T of the softmax, since $T < 1$ leaves only the extreme points of attention whereas $T > 1$ smooths the CAM over the image.

Negative CAM loss. One problem with CAMs is that models tend to produce false activations, *i.e.* attention maps localized in precise regions even when a class is not present in the input image [29]. Since our consistency loss heavily relies on the quality of the CAMs, we use the negative CAM loss [29] to penalize attention maps for absent classes in the input (Figure 2, red box). The loss is defined as:

$$\mathcal{L}_{\text{NEG}}(M, C_k) = \sum_{c \in C_k} D_{KL}(U || M_c) + D_{KL}(U || \hat{M}_c), \tag{10}$$

where U is the uniform distribution over the spatial locations s, and C_k is the set of top-k negative classes in terms of their confidence scores for the clean image X. From Eqs. (9) and (10), we can define our final objective as:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda(\mathcal{L}_{\text{CAM}} + \mathcal{L}_{\text{NEG}}). \tag{11}$$

We name our final model **A**ttention **C**onsistency on **V**isual **C**orruptions (**ACVC**).

3.3. Algorithm Overview

We summarize the model training in Algorithm 1. As the algorithm shows, we first sample a training image and its label (line 1) from $\mathcal{D}$. We then sample a random corruption from our set Φ (line 2) and we apply the transformation to the input image (line 3). For each sample in the batch

	Photo	Art	Cartoon	Sketch	Avg.	Max.
Baseline	98.52 ± 0.4	55.62 ± 2.2	18.56 ± 2.6	25.81 ± 4.8	33.33 ± 2.4	37.11
MixUp [40]	97.32 ± 0.7	52.82 ± 0.7	16.97 ± 4.4	23.21 ± 4.5	31.00 ± 1.7	32.83
CutOut [8]	98.49 ± 0.6	59.84 ± 1.3	21.56 ± 1.6	28.83 ± 3.3	36.74 ± 1.5	39.24
CutMix [39]	98.20 ± 0.6	59.63 ± 1.8	21.98 ± 3.9	24.94 ± 4.7	35.52 ± 2.3	38.92
ME-ADA [41]	96.49 ± 0.8	55.61 ± 0.9	28.92 ± 1.5	24.63 ± 4.3	36.39 ± 1.8	39.08
RandAugment [7]	$\mathbf{99.22} \pm 0.6$	$\mathbf{67.81} \pm 0.9$	28.94 ± 2.6	36.96 ± 4.7	44.57 ± 2.3	48.79
AugMix [13]	98.44 ± 0.3	63.94 ± 1.6	27.72 ± 1.4	30.86 ± 3.2	40.84 ± 1.4	43.11
VC (Ours)	$\underline{98.75} \pm 0.6$	67.23 ± 0.5	$\underline{30.26} \pm 2.1$	$\underline{43.81} \pm 3.9$	$\underline{47.10} \pm 1.7$	$\underline{49.48}$
ACVC (Ours)	$\mathbf{99.22} \pm 0.4$	$\underline{67.80} \pm 0.9$	$\mathbf{30.31} \pm 2.1$	$\mathbf{46.42} \pm 6.7$	$\mathbf{48.18} \pm 2.8$	$\mathbf{54.67}$

Table 1. Comparing with the state of the art on PACS benchmark on single DG task using ResNet-18. The models are trained on Photo domain, and tested on Art, Cartoon and Sketch domains. We measure classification accuracy. Baseline: ResNet-18 trained with cross-entropy loss w/o any augmentations. Best numbers are bold, second best are underlined. VC = ACVC w/o attention consistency.

we compute its prediction on both original and corrupted samples (line 4) and their relative CAMs (line 5). Finally, we compute our loss using Eq. (11) (line 6), the gradient of the parameters w.r.t. the loss (line 8) and update the parameters (line 9). During training, we apply the corruptions without any additional augmentation techniques. At inference, no corruption is applied on the images of unseen domains.

4. Experiments

Datasets and setup. We evaluate our model on three challenging benchmarks for single DG: PACS [16], COCO [20], and DomainNet [23] in increasing order of difficulty.

PACS is a standard multi-source domain generalization benchmark [16], with 9,991 images belonging to 7 different classes. We use it in the single DG setting due to the extreme domain-shift between its four domains, *i.e*. Photo, Art painting, Cartoon and Sketch. Since in this work we are specifically interested in generalizing from natural images, we consider Photo as the source domain.

For **COCO**, we propose a new benchmark consisting of 10 shared classes between the original MS-COCO [20] and DomainNet [23]. We take MS-COCO as the training set, and test with the six domains in DomainNet: Real, Infograph, Painting, Clipart, Sketch, and Quickdraw. This setting is similar to that of [44], but we only use training images where the target object covers at least 10% of the pixels. Since this constraint may limit the number of images of some classes, we avoid class imbalance by setting 1,000 as the upper bound on the number of samples per class, obtaining 7,783 images in total. As in [44], we test on this benchmark since the available segmentation masks allows us (and potentially future works) to explore how modeling the location of an object can improve the single DG performance.

Finally, we include a large scale investigation using the full **DomainNet** dataset. It has 345 object classes and contains 596,010 images in total. We use the Real domain for training and validation, and the other five for testing. This dataset is extremely challenging due to the high-variability of the domains and the large number of classes.

For all settings, we resize the RGB images to 224×224, and use the official train/val/test splits. We employ an ImageNet [27] pretrained ResNet-18 [11], and use SGD optimizer with a learning rate of 4×10^{-3}, a batch size of 128 and we train for 30 epochs, dropping the learning rate by 0.1 after 24 epochs. These are the hyperparameters proposed by [14] for multi-source domain generalization using PACS, and we keep these hyperparameters constant across our three benchmarks. In addition, for ACVC, we set $k = 3$ empirically, and $\lambda = 0.06$ as in [29]. For our Fourier-based corruptions, we define 5 severity levels for α, β, and d, as in ImageNet-C (see supplementary). During training, we randomly sample the severity of both ImageNet-C and Fourier-based operations uniformly from these 5 levels.

Baselines and metrics. We establish the single DG performance comparison using (1) a deep neural network trained using cross-entropy loss but without any data augmentation (Baseline), (2) advanced augmentation techniques, *i.e*. MixUp [40], CutOut [8], CutMix [39], (3) methods that randomly select augmentations from a large pool of transformations (where most corruption operations are omitted), *i.e*. RandAugment [7], AugMix [13], and (4) the state-of-the-art adversarial data augmentation technique, *i.e*. ME-ADA [41]. These methods do not have any reported results on our benchmarks, hence, we run our own experiments using the authors' implementations and suggested configurations if applicable. We provide mean accuracy and standard deviation measurements for multiple runs and the maximum achievable average domain generalization performance per method. The code will be released upon acceptance.

4.1. Comparison on PACS

Table 1 shows our evaluation on PACS, where there exists a large distribution shift between the source and target do-

	COCO	Real	Painting	Infograph	Clipart	Sketch	Quickdraw	Avg.	Max.
Baseline	80.44 ± 0.7	84.15 ± 0.7	78.55 ± 0.5	31.56 ± 2.1	62.90 ± 3.4	44.93 ± 1.7	12.56 ± 2.4	52.44 ± 1.0	54.46
MixUp [40]	80.79 ± 0.7	78.61 ± 1.0	73.70 ± 1.0	23.96 ± 1.1	50.39 ± 2.7	38.82 ± 1.7	13.59 ± 1.1	46.51 ± 0.8	48.21
CutOut [8]	$\underline{80.93} \pm 0.4$	84.17 ± 0.4	79.58 ± 0.8	32.45 ± 1.7	61.62 ± 2.6	39.73 ± 3.2	10.42 ± 0.6	51.33 ± 1.1	53.35
CutMix [39]	80.13 ± 0.6	84.03 ± 0.9	78.72 ± 0.7	31.73 ± 1.4	64.08 ± 2.8	43.35 ± 2.1	12.22 ± 0.9	52.35 ± 0.7	53.88
ME-ADA [41]	78.35 ± 0.9	82.28 ± 1.0	77.69 ± 0.5	28.58 ± 1.4	63.88 ± 1.8	45.29 ± 1.7	12.32 ± 1.1	51.67 ± 0.8	52.71
RandAugment [7]	80.51 ± 0.6	85.55 ± 0.6	$\underline{81.67} \pm 0.4$	$\underline{33.87} \pm 0.9$	67.96 ± 2.8	52.58 ± 1.4	14.57 ± 1.3	56.03 ± 0.5	56.75
AugMix [13]	80.50 ± 0.6	$\underline{85.60} \pm 0.6$	80.19 ± 0.7	33.50 ± 1.6	71.37 ± 0.8	51.96 ± 1.3	17.96 ± 2.3	56.76 ± 0.7	57.79
VC (Ours)	80.58 ± 0.5	$\mathbf{85.86} \pm 0.7$	80.94 ± 0.6	32.50 ± 1.0	$\underline{71.93} \pm 1.4$	$\underline{61.98} \pm 1.8$	$\underline{18.42} \pm 1.4$	$\underline{58.61} \pm 0.7$	$\underline{59.63}$
ACVC (Ours)	$\mathbf{81.80} \pm 0.6$	85.27 ± 0.5	$\mathbf{82.37} \pm 0.6$	$\mathbf{35.40} \pm 0.6$	$\mathbf{73.04} \pm 0.8$	$\mathbf{62.72} \pm 1.0$	$\mathbf{21.25} \pm 0.9$	$\mathbf{60.01} \pm 0.3$	$\mathbf{60.23}$

Table 2. Comparing with the state of the art on COCO benchmark on single DG task using ResNet-18. The models are trained on COCO dataset, and tested on DomainNet dataset. We measure classification accuracy. Baseline: ResNet-18 trained with cross-entropy loss only w/o any augmentations. Bold figures are the highest numbers, underlined are the second highest. VC = ACVC w/o attention consistency.

mains (*e.g*. Photo to Sketch). Despite the large domain gap, our proposed methods surpass all competitors on this benchmark. Visual corruptions alone (VC) obtain superior performance on PACS, with an average accuracy of $47.10\% \pm 1.7$ across the different unseen domains in comparison to RandAugment with $44.57\% \pm 2.3$ showing its effectiveness for single DG. On average, due to our visual attention consistency loss, ACVC improves VC results by 1.08%, and on the best case, ACVC can go as high as 54.67% average single DG performance. Moreover, the table shows how data augmentation methods that apply a single type of transformation do not provide enough input variations for training, with ME-ADA and CutOut achieving 36.39% and 36.74% respectively in average. However, applying multiple transformations per image may also hurt the performance, *e.g*. RandAugment outperforms AugMix by 3.73% despite using a single transformation and no contrastive loss term.

We see that, as the domain shift increases (Art $\rightarrow$ Sketch), standard deviations of all methods also increase, which is the reason behind the large gap between the average and maximum performance measurements. We believe this problem to be caused by the training set size, since PACS contains only $1{,}499$ Photo images.

4.2. Comparison on COCO

Table 2 shows our results on the COCO benchmark. VC alone again outperforms the competitors with an average single DG accuracy of $58.61\% \pm 0.7$ where the best method in literature, AugMix achieves $56.76\% \pm 0.7$. Combined with visual attention consistency, *i.e*. ACVC, the average performance reaches to $60.01\% \pm 0.3$, improving the VC accuracy by 1.4%. Contrary to PACS, we see that with enough training data ($7{,}783$ images from COCO), the standard deviation of ACVC's performance is relatively small.

According to the results, corruptions help generalizing to distant domains such as Sketch and Quickdraw. Note that the common subset of COCO and DomainNet datasets includes classes such as *bus*, *car* and *truck* which often have Quickdraw examples that are easily confused one with another. Therefore, any improvement on this domain tends to be limited for this particular benchmark. Nevertheless, ACVC can provide 3.29% improvement over AugMix (17.96%) on Quickdraw. Another interesting observation is that, in addition to the single DG performance rankings, the ranking between different methods change even between COCO dataset and Real domain of DomainNet. For instance, MixUp accuracy decreases $80.79\% \rightarrow 78.61\%$ where ME-ADA accuracy increases $78.35\% \rightarrow 82.28$. Note that, COCO dataset is designed to have multiple target classes in a given scene, whereas Real domain of DomainNet contains mostly centralized images w.r.t. the object of interest, thus performance may increase when testing on the latter.

4.3. Comparison on DomainNet

Table 3 shows our results on large-cale DomainNet benchmark. This is the most challenging setting, due to the large domain-shift among domains (*e.g*. Real to Infograph, Real to Quickdraw), and the large number of classes (345). Even in this benchmark, visual corruptions alone (VC) improve single DG performance, achieving 26.68% accuracy with the best competitor (AugMix) achieving 26.48%.

When visual attention consistency is used (ACVC), the avg. single DG accuracy reaches 26.89%. For Quickdraw images, methods with additional supervision signal tends to perform better, *i.e*. AugMix and ACVCachieving 6.26%, and 6.57 accuracy, respectively. Table 3 validates once again the importance of simulating different visual variations by collecting a set of transformations: the gap between the best single (and adversarial) augmentation technique (ME-ADA) and VC is more than 2% on average.

Finally, this benchmark reveals that even though all methods perform relatively well on the source domain ($[74.27\%, 76.96\%]$), we still do not have robust vision models since their performance significantly drops as the domain shift increases, *e.g*. as in Infograph and Quickdraw cases. The former shows a model's ability to filter out texts, charts and other irrelevant sources of information to focus on the object, and the performance of all methods drops to the

	Real	Painting	Infograph	Clipart	Sketch	Quickdraw	Avg.	Max.
Baseline	76.04 ± 0.8	38.05 ± 0.8	13.31 ± 0.4	37.89 ± 1.2	26.26 ± 1.3	3.36 ± 0.2	23.78 ± 0.8	24.34
MixUp [40]	76.11 ± 0.2	38.60 ± 0.1	$\mathbf{13.94} \pm 0.2$	38.02 ± 0.8	26.01 ± 0.7	3.71 ± 0.3	24.05 ± 0.4	24.45
CutOut [8]	$\mathbf{76.96} \pm 0.8$	38.34 ± 0.7	13.69 ± 0.4	38.44 ± 1.3	26.24 ± 0.8	3.65 ± 0.4	24.07 ± 0.7	24.69
CutMix [39]	75.79 ± 0.7	38.28 ± 1.1	13.45 ± 0.5	38.65 ± 1.8	26.85 ± 1.5	3.60 ± 0.4	24.17 ± 1.1	24.96
ME-ADA [41]	74.27 ± 0.1	37.95 ± 0.1	13.12 ± 0.0	40.31 ± 0.1	26.79 ± 0.1	4.53 ± 0.2	24.54 ± 0.0	24.60
RandAugment [7]	$\underline{76.70} \pm 0.4$	41.30 ± 0.8	13.57 ± 0.3	41.11 ± 1.1	30.40 ± 1.0	5.31 ± 0.5	26.34 ± 0.7	$\underline{26.85}$
AugMix [13]	76.27 ± 0.1	40.79 ± 0.3	$\underline{13.89} \pm 0.1$	41.67 ± 0.3	29.80 ± 0.2	$\underline{6.26} \pm 0.0$	26.48 ± 0.2	26.61
VC (Ours)	75.91 ± 0.3	$\mathbf{41.38} \pm 0.3$	13.58 ± 0.3	$\underline{41.80} \pm 0.7$	$\underline{30.58} \pm 0.5$	6.06 ± 0.4	$\underline{26.68} \pm 0.2$	$\mathbf{26.91}$
ACVC (Ours)	76.16 ± 0.5	$\underline{41.32} \pm 0.6$	12.89 ± 0.6	$\mathbf{42.79} \pm 0.3$	$\mathbf{30.86} \pm 0.5$	$\mathbf{6.57} \pm 0.5$	$\mathbf{26.89} \pm 0.0$	$\mathbf{26.91}$

Table 3. Comparing with the state of the art on large-scale DomainNet benchmark on single DG task using ResNet-18. The models are trained on Real domain, and tested on Painting, Infograph, Clipart, Sketch and Quickdraw domains. We measure classification accuracy. Baseline: ResNet-18 trained with cross-entropy loss only w/o any augmentations. Bold figures are the highest numbers, underlined are the second highest. VC does not contain attention consistency. ACVC is our full model.

	PACS	COCO
Baseline	33.33 ± 2.4	52.44 ± 1.0
Weather	40.36 ± 2.3	$\underline{55.69} \pm 0.4$
Blur	36.83 ± 1.3	53.39 ± 0.1
Noise	35.53 ± 1.9	53.21 ± 0.7
Digital	39.79 ± 3.4	55.12 ± 0.7
Fourier	34.15 ± 1.5	54.18 ± 0.4
ImageNet-C	$\underline{42.12} \pm 2.5$	55.52 ± 0.9
VC	$\mathbf{47.10} \pm 1.7$	$\mathbf{58.61} \pm 0.7$

Table 4. Ablation study of different visual corruptions on PACS, and COCO. ImageNet-C contains Weather, Blur, Noise and Digital corruptions. VC contains all five, including Fourier category.

range $[12.89\%, 13.94\%]$. The latter contains mostly primitive drawings to represent an object without color, texture or background, and the range of classification accuracy becomes $[3.36\%, 6.57\%]$.

4.4. Ablation Study

Corruptions. Here we study the effect of (1) each visual corruption category; (2) ImageNet-C corruptions; and (3) our VC. As Table 4 shows, different corruptions work differently across domains. For instance, Noise and Fourier families perform well on COCO, but they offer limited improvement upon Baseline for PACS. For Blur, we see the opposite case: it improves PACS performance, but performs similar to Baseline on COCO benchmark. On the other hand, Weather and Digital categories perform close to full ImageNet-C category across all domains. Each category brings improvement over Baseline performance, however, randomly sampling transformations from all five (VC) consistently yields better performance than any individual family. We can also see that our additional Fourier-based visual corruptions bring a significant improvement over the original ImageNet-C

	PACS	COCO
VC	47.10 ± 1.7	58.61 ± 0.7
+ $\mathcal{L}_{\text{JSD}}$	$\underline{47.39} \pm 2.6$	57.83 ± 0.6
+ $\mathcal{L}_{\text{MSE}}$	42.91 ± 1.8	58.02 ± 0.4
+ $\mathcal{L}_{\text{NEG}}$	43.00 ± 0.9	$\underline{59.68} \pm 0.7$
+ $\mathcal{L}_{\text{CAM}}$	46.55 ± 2.7	$\underline{59.67} \pm 0.6$
+ segm. masks + $\mathcal{L}_{\text{NEG}}$	N/A	58.65 ± 0.5
+ $\mathcal{L}_{\text{CAM}}$ + $\mathcal{L}_{\text{NEG}}$ (ACVC)	$\mathbf{48.18} \pm 2.8$	$\mathbf{60.01} \pm 0.3$

Table 5. Ablation study of the different loss terms reported on PACS and COCO benchmarks.

family of transformations. In detail, VC, on average, improves ImageNet-C results by 4.98% on PACS, and 2.81% on COCO. This suggests that randomly combining multiple visual corruptions is the best choice when there is no prior knowledge on the target domains, merging the benefits of all families while diminishing the negative effects that single corruption categories may have in particular benchmarks.

Consistency loss. Here we study the effects of different consistency loss terms on the single DG performance when applied on top of VC. Table 5 shows that not all consistency losses bring the same improvements over VC. For instance, JSD loss on model predictions for the original and augmented versions (as in [13]), slightly improves PACS results ($+0.9\%$), but degrades the performance on COCO (-0.8%). When we apply our attention consistency loss $\mathcal{L}_{\text{CAM}}$, performance is significantly better ($+3.64\%$ on PACS, $+1.65\%$ on COCO) than simple MSE loss between CAMs. Nevertheless, $\mathcal{L}_{\text{CAM}}$ alone is still not robust, as it degrades the performance of VC by 0.55% on PACS, but improves it by 1.08% on COCO. This is also the case for improving the CAMs using only $\mathcal{L}_{\text{NEG}}$ without any consistency loss (*i.e.* -4.1% on PACS but $+1.07\%$ on COCO w.r.t. VC). However, when we combine both terms, we achieve consistent improvement in both benchmarks, *i.e.* $+1.08\%$ on PACS and $+1.4\%$ on COCO w.r.t. VC. Notably, the benefits of using

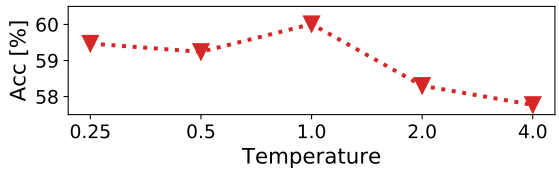

Figure 3. ACVC results on COCO benchmark for different T.

an additional loss terms over VC do not generalize across PACS and COCO, except for ACVC.

Finally, we analyze the effect of replacing our consistency loss on CAMs by imposing as fixed target in Eq.(9) the normalized segmentation mask of the image provided by COCO dataset. Results shows that $\mathcal{L}_{\text{CAM}}$ does not benefit from having a static reference point to optimize towards. On the contrary, ACVC is able to achieve, on average, 1.36% higher single DG performance than using segmentation masks. We ascribe this behaviour to the nature of the softmax that spreads the intensity of the focus over the whole extent of the object and penalizes peaked values of the attention maps, even when they fall inside the object. This can also be seen on Figure 3, where the avg. single DG performance is relatively better for $T \leq 1$, which shows how imposing consistency on peaks of the attention maps is more beneficial for single DG than smoothing the attention over larger spatial regions.

4.5. Qualitative Results

In this section, we show CAMs for four different approaches, (1) the baseline model, (2) RandAugment and our VC as powerful pure data augmentation techniques, and (3) our final ACVC method. In Figure 4, we see that ACVC can recognize and focus on the relevant objects in unseen domains. In detail, the top two rows show paintings where ACVC is able to focus on the correct objects even in frames within a crowded scene. The last two rows show images from the challenging Infograph domain which contains charts, texts and symbols in addition to the target objects. Nevertheless, ACVC can still recognize the bus in both images.

5. Conclusion

In this work, we addressed the problem of single source domain-generalization (single DG) where the goal is to classify images of arbitrary unseen distributions, given a single domain at training time. Similar to previous works, we address the problem by synthesizing multiple training domains. However, unlike previous approaches, we propose to generate new domains by applying randomly sampled visual corruptions on the training data. Specifically, we consider a set of transformations that corrupt the original content in twenty-two different ways belonging to five categories of transformations (*i.e*. Weather, Blur, Noise, Digital, and Fourier). Since these transformations keep the object locations intact, we propose a visual attention consistency loss

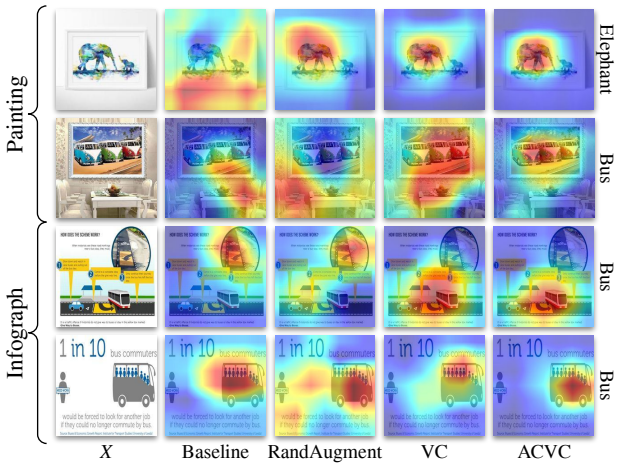

Figure 4. Class activation maps of (1) the baseline model, (2) two different sets of data augmentation techniques, *i.e*. RandAugment and the proposed VC models, (3) attention consistency guided VC, *i.e*. ACVC. Our ACVC approach obtains more fine-grained attention maps on unseen domains.

between the model's class activation maps for the original and corrupted versions of an input image. This loss ensures that the model focuses on the same image regions, disregarding the particular style of the input. Experiments show that our method, ACVC, consistently outperforms the state of the art in PACS, COCO and DomainNet benchmarks.

Broader societal impact. Our method focuses on scenario where generalizing to unseen data distributions is crucial. As a consequence, ACVC can be applied in all scenarios involving robustness to different environmental conditions (*e.g*. illumination, weather) as well as recognition across different visual modalities (*e.g*. photo, cartoon, sketch). The ability to generalize to unseen domains without collecting additional unlabeled (as in domain adaptation [5]) or labeled (as in domain generalization [16]) data from different distributions could bring a positive impact on scenarios with privacy constraints (*e.g*. federated learning [19]), since it reduces the need of collecting data for specializing the recognition model to single users. We want to highlight that the data used for experiments (PACS, COCO, and DomainNet) are all public datasets and do not contain any private information or disclose any identifiable personal information.

Limitations. One limitation of our work is that we explicitly focus on generalizing from natural images, containing rich visual information. In this context, removing information through corruptions is beneficial for single DG performance. However, our approach may not be suitable for source domains where the input already presents limited information, such as sketches. In these cases we may need to replace our pool of corruptions with tailored augmentation techniques.

Acknowledgements This work has been partially funded by the ERC (853489 - DEXIM) and by the DFG (2064/1 – Project number 390727645).

References

[1] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *CVPR*, 2019. 2

[2] Prithvijit Chattopadhyay, Yogesh Balaji, and Judy Hoffman. Learning to balance specificity and invariance for in and out of domain generalization. In *ECCV*. Springer, 2020. 1, 2

[3] Yang Chen, Yu Wang, Yingwei Pan, Ting Yao, Xinmei Tian, and Tao Mei. A style and semantic memory mechanism for domain generalization. In *ICCV*, 2021. 2

[4] Dan Ciregan, Ueli Meier, and Jürgen Schmidhuber. Multi-column deep neural networks for image classification. In *CVPR*, 2012. 2

[5] Gabriela Csurka. A comprehensive survey on domain adaptation for visual applications. *Domain adaptation in computer vision applications*, 2017. 1, 8

[6] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. In *CVPR*, 2019. 2

[7] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPRW*, 2020. 2, 5, 6, 7

[8] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 2, 5, 6, 7

[9] Xinjie Fan, Qifei Wang, Junjie Ke, Feng Yang, Boqing Gong, and Mingyuan Zhou. Adversarially adaptive normalization for single domain generalization. In *CVPR*, 2021. 2

[10] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *ICCV*, 2015. 2

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1, 5

[12] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR*, 2019. 2, 3

[13] 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. 2020. 2, 3, 5, 6, 7

[14] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *ECCV*, 2020. 5

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *NeurIPS*, 2012. 1, 2

[16] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *ICCV*, 2017. 1, 2, 5, 8

[17] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *CVPR*, 2018. 2

[18] Lei Li, Ke Gao, Juan Cao, Ziyao Huang, Yepeng Weng, Xiaoyue Mi, Zhengze Yu, Xiaoya Li, and Boyang Xia. Progressive domain expansion network for single domain generalization. In *CVPR*, 2021. 2

[19] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3), 2020. 8

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 2, 5

[21] Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S. Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *NeurIPSW*, 2019. 3

[22] Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *ICCV*, 2017. 1, 2

[23] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *ICCV*, 2019. 2, 5

[24] Leon N Piotrowski and Fergus W Campbell. A demonstration of the visual importance and flexibility of spatial-frequency amplitude and phase. *Perception*, 1982. 3

[25] Fengchun Qiao and Xi Peng. Uncertainty-guided model generalization to unseen domains. In *CVPR*, 2021. 2

[26] Fengchun Qiao, Long Zhao, and Xi Peng. Learning to learn single domain generalization. In *CVPR*, 2020. 1, 2

[27] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3), 2015. 5

[28] Seonguk Seo, Yumin Suh, Dongwan Kim, Geeho Kim, Jongwoo Han, and Bohyung Han. Learning to optimize domain specific normalization for domain generalization. In *ECCV*, 2020. 2

[29] Guolei Sun, Salman Khan, Wen Li, Hisham Cholakkal, Fahad Shahbaz Khan, and Luc Van Gool. Fixing localization errors to improve image classification. In *ECCV*, 2020. 3, 4, 5

[30] Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In *Allerton Conference on Communication, Control and Computing*, 1999. 2

[31] I Tolstikhin, O Bousquet, S Gelly, and B Schölkopf. Wasserstein auto-encoders. In *ICLR*, 2018. 2

[32] Riccardo Volpi and Vittorio Murino. Addressing model vulnerability to distributional shifts over image transformation sets. In *ICCV*, 2019. 1, 2

[33] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. In *NeurIPS*, 2018. 1, 2

[34] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312, 2018. 1

[35] Shujun Wang, Lequan Yu, Caizi Li, Chi-Wing Fu, and Pheng-Ann Heng. Learning from extrinsic and intrinsic supervisions for domain generalization. In *ECCV*. Springer, 2020. 1

[36] Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *CVPR*, 2021. 3

[37] Yanchao Yang, Dong Lao, Ganesh Sundaramoorthi, and Stefano Soatto. Phase consistent ecological domain adaptation. In *CVPR*, 2020. 3

[38] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *CVPR*, 2020. 3

[39] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *CVPR*, 2019. 2, 5, 6, 7

[40] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 2, 5, 6, 7

[41] Long Zhao, Ting Liu, Xi Peng, and Dimitris Metaxas. Maximum-entropy adversarial data augmentation for improved generalization and robustness. In *NeurIPS*, 2020. 1, 2, 3, 5, 6, 7

[42] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016. 1, 4

[43] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *ICLR*, 2020. 2

[44] Andrea Zunino, Sarah Adel Bargal, Riccardo Volpi, Mehrnoosh Sameki, Jianming Zhang, Stan Sclaroff, Vittorio Murino, and Kate Saenko. Explainable deep classification models for domain generalization. In *CVPRW*, 2021. 5