Domain Generalised Faster R-CNN

Karthik Seemakurthy[1], **Charles Fox**[2], **Erchan Aptoula**[3], **Petra Bosilj**[1, 2]

[1]Lincoln Institute of Agri-Food Technology, University of Lincoln, United Kingdom.
[2]School of Computer Science, University of Lincoln, United Kingdom.
[3]Faculty of Engineering and Natural Sciences (VPALab), Sabanci University, Türkiye.
kseemakurthy@lincoln.ac.uk, chfox@lincoln.ac.uk, erchan.aptoula@sabanciuniv.edu, pbosilj@lincoln.ac.uk

Abstract

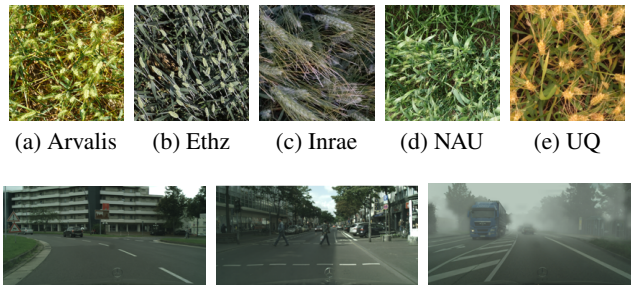
Domain generalisation (i.e. out-of-distribution generalisation) is an open problem in machine learning, where the goal is to train a model via one or more source domains, that will generalise well to unknown target domains. While the topic is attracting increasing interest, it has not been studied in detail in the context of object detection. The established approaches all operate under the covariate shift assumption, where the conditional distributions are assumed to be approximately equal across source domains. This is the first paper to address domain generalisation in the context of object detection, with a rigorous mathematical analysis of domain shift, without the covariate shift assumption. We focus on improving the generalisation ability of object detection by proposing new regularisation terms to address the domain shift that arises due to both classification and bounding box regression. Also, we include an additional consistency regularisation term to align the local and global level predictions. The proposed approach is implemented as a Domain Generalised Faster R-CNN and evaluated using four object detection datasets which provide domain metadata (GWHD, Cityscapes, BDD100K, Sim10K) where it exhibits a consistent performance improvement over the baselines. All the codes for replicating the results in this paper can be found at https://github.com/karthikiitm87/domain-generalisation.git

Introduction

Object detection is the task of identifying and localising all instances of a certain object in an image. Benchmark performances have increased significantly using deep learning approaches (Jocher et al. 2020; Tan, Pang, and Le 2020; Liu et al. 2016; Ren et al. 2015; Lin et al. 2017; Dai et al. 2016; Lin et al. 2014). Factors such as viewpoint, background, weather, and image quality increase the variations in object appearance (autonomous farming and driving examples in Fig. 1). The resulting distribution discrepancy between training and testing data is called *domain shift* and degrades model performance at deployment (Recht et al. 2019; Hendrycks and Dietterich 2019).

Although increasing the amount and diversity of training data can can alleviate the impact of domain shift in theory, image annotation remains an expensive and time consuming task, and there can be no guarantee that all deployment scenarios will be adequately represented. *Domain generalisation* (DG) aims to learn a single, unchanging set of parameters which are able to perform well on previously unseen domains. This is achieved by training jointly on several domains and ensuring that the domain-specific information is suppressed. Unlike *domain adaptation* (DA)(Chen et al. 2018; Cai et al. 2021; Saito et al. 2019; Xu et al. 2020a; Li et al. 2020; Hsu et al. 2020b; Li, Liu, and Yuan 2022; Hsu et al. 2020a; Rezaeianaran et al. 2021; Wang et al. 2021), DG is useful when the system needs to be deployed in an unknown setting, as it addresses domain shift without requiring any additional training data from the target domain, which often cannot be obtained in advance even without annotations. In DA, parameters of a trained model are adapted by exploiting unlabelled data from the target domain(s), however, in practice target domain data is often sparse and unavailable. Despite this, recent surveys of domain shift (Zhou et al. 2021; Koh et al. 2021) show that DA is still the more common approach for addressing domain shift in object detection, with very little work on DG for object detection (Liu, Song, and Ding 2020a; Lin et al. 2021).

Furthermore, this is the first paper to systematically analyse domain shift for object detection without assuming the covariate shift case. While the assumption that the conditional distributions are approximately equal was disputed and addressed in classification (Zhao et al. 2020), techniques



(a) Arvalis (b) Ethz (c) Inrae (d) NAU (e) UQ

(f) Aachen (g) Hanover (h) Foggy-Lindau

Figure 1: Samples from various training and testing domains in the Global Wheat Head Detection (GWHD) (a)-(e) and Cityscapes (f)-(h) datasets used in our experiments.

which improve classification do not directly translate to improvements on more complex tasks such as object detection. We apply this analysis to Faster R-CNN (Ren et al. 2015), a representative two-stage detector, to propose a Domain Generalised Faster R-CNN (DGFR-CNN). This is achieved using a novel loss to assists in aligning both the class and bounding box conditionals (concept shift), which we show to be a necessary criteria to bridge the domain shift across multiple source domains.

Experimental validation has been conducted on four datasets: Sim10K (Johnson-Roberson et al. 2017), Cityscapes (Cordts et al. 2016), BDD100K (Yu et al. 2020), and GWHD (David et al. 2021). GWHD is the only object detection dataset among a number of established and newly proposed DG benchmarks (Fang, Xu, and Rockmore 2013; Koh et al. 2021; Zhou et al. 2021; Saenko et al. 2010; Torralba and Efros 2011), concerning single-object detection across multiple target domains. Both multi-object and single-object detection scenarios are evaluated using the autonomous driving datasets. In addition to evaluating the the DGFR-CNN under a variety of domain shifts (Fig. 1), we also evaluate the reduced version of the proposed architecture under the DA setting (where the class conditional alignment is not performed as this information is unavailable during the DA re-training step).

Related Works

The proposed system builds upon work in object detection, domain adaptation and domain generalisation. **Object detection.** Classical approaches to object detection relied on handcrafted features and formulated object detection as a sliding window classification problem (Dalal and Triggs 2005; Felzenszwalb et al. 2009; Viola and Jones 2001). The empirical success of convolutional neural networks (CNNs) – which automatically extract high performing features – (LeCun et al. 1998; Krizhevsky, Sutskever, and Hinton 2012) has resulted in their widespread adoption. Most CNN-based approaches to object detection can be categorised as either single-stage or two-stage. Single-stage approaches perform localisation and classification simultaneously (Redmon et al. 2016) and have only recently reached the performance of two-stage approaches (Jocher et al. 2020). Two-stage approaches developed from the Region-based CNN (R-CNN) family, and are characterised by a network trained to classify region proposals selected from the image. Region proposals were initially selected by a static approach and processed independently (Girshick et al. 2014). R-CNN was then extended to use a shared feature map for the region proposals from the same image (Girshick 2015). Faster R-CNN (Ren et al. 2015) additionally introduced a Region Proposal Network (RPN), producing object proposals as well as the feature maps used for region classification, thus resulting in an end-to-end trainable system. However, all of these architectures and a number of follow-up works (Liu et al. 2016; Dai et al. 2016; Lin et al. 2017) operate under the assumption that the testing data originates from the same domain and distribution as the training data and their performance consequently degrades on out-of-domain (OOD) data.

Domain adaptation. (Chen et al. 2018) proposed an approach to improve the performance of Faster R-CNN detector on unlabelled target data, by aligning the image- and instance-level domain classifiers using a consistency regulariser. However, their assumption that the instance level classifier is consistent across domains does not hold when domain shift is significant (Zhao et al. 2020; Li et al. 2018c; Schölkopf et al. 2012; Janzing and Schölkopf 2010). Subsequent attempts of DA for object detection (Li et al. 2020; Sindagi et al. 2020) include using adversarial learning to diminish domain discrepancy (Chen et al. 2018; He and Zhang 2019; Saito et al. 2019; Xu et al. 2020a; Zhu et al. 2019; Wu et al. 2021), reconstruction-based methods using image style transfer as an auxiliary task (Hsu et al. 2020b; Rodriguez and Mikolajczyk 2019; Gong et al. 2019b), feature disentanglement approaches whose aim is to suppress domain specific features and enhance domain invariant features (Lin et al. 2021; Wu et al. 2021) and self-training methods using pseudolabels to retrain a model (Khodabandeh et al. 2019; RoyChowdhury et al. 2019; Cai et al. 2021). Furthermore, the aforementioned approaches consider a single source and target domain, with multi-source DA for object detection having been considered only recently (Yao et al. 2021).

The present study is most closely related to adversarial approaches which align feature representations of source and target domains (Chen et al. 2018) which assume covariate shift alone, although in our case we align features from multiple source domains by accounting for both covariate and concept shifts.

Domain generalisation. DG was first studied by Blanchard, Lee, and Scott (2011) in the context of medical imaging, while the terminology was introduced later by Muandet, Balduzzi, and Schölkopf (2013). Earlier studies have explored fixed shallow features (Fang, Xu, and Rockmore 2013; Ghifary et al. 2015; Khosla et al. 2012; Muandet, Balduzzi, and Schölkopf 2013; Xu et al. 2014), while more recent investigations design architectures to address domain shift (Li et al. 2017) or learning algorithms to optimise standard architectures (Li et al. 2018a; Shankar et al. 2018; Li et al. 2019). Domain randomisation (Tobin et al. 2017; Yue et al. 2019) is a complementary approach to DG, which relies on synthetically generated variations of the input data to obtain more generalisable features. Domain randomisation was applied to car detection (Khirodkar, Yoo, and Kitani 2019), however, their approach requires 3D models of the detected objects, camera parameters and scene geometry. A single-shot YOLO detector has also been extended with DG components (Liu, Song, and Ding 2020b,a), however they rely on Invariant Risk Minimisation (IRM) (Arjovsky et al. 2019) which is likely to underperform when there is a significant domain shift (Rosenfeld, Ravikumar, and Risteski 2020). (Lin et al. 2021) proposed a feature disentanglement approach at both image and instance levels to extract the domain invariant features across multiple source domains.

The existing DG approaches suffer from the same drawback identified in DA, namely the disputed assumption that the conditional class distributions do not vary across domains (Zhao et al. 2020; Li et al. 2018c; Hu et al. 2020; Li et al. 2018b). According to recent surveys, the majority of DG

work has been conducted in the classification setting (Koh et al. 2021; Zhou et al. 2021). Conversely, the proposed work addresses domain shift in object detection, relying on entropy based regularisation (Zhao et al. 2020) to achieve class conditional invariance.

Preliminaries

In this section we describe the details of the mathematical framework used to describe the DG task for object detection in the following section.

Let $\mathbf{I} \times \mathbf{C} \times \mathbf{B} \times \mathbf{D}$ be the sample space under observation. Here $I \in \mathbf{I}$ denotes the images of interest. $B^I \in \mathbb{R}^4 = \mathbf{B}$ is the bounding box predictor consisting of a vector of image coordinates (x, y, w, h) corresponding to the objects detected in image I. $C^I \in \{0, \dots K\} = \mathbf{C}$ denotes the integer-valued class label assigned to each of the detected bounding boxes B^I, where K is the total number of object classes, and $D^I \in \mathbf{D}$ represents the domain of image I. Let $P_D\left(I, C^I, B^I\right)$ denote the joint distributions defined on the sample space $\mathbf{I} \times \mathbf{C} \times \mathbf{B}$ given domain D. The goal of this work is to train the DGFR-CNN detector that can learn domain-invariant features from N source domains which can be generalised to an unseen target domain (or domains) without compromising the main detection task. This is achieved via proposed regularisation terms along with the main detection loss.

Let (θ, ϕ, β) be the parameters for a backbone feature extractor $F^{(\theta)}$, a classifier $T^{(\phi)}$, and a bounding box predictor $R^{(\beta)}$, respectively. Let $Q^{T^{(\phi)},R^{(\beta)}}(F^{(\theta)}(I), C^I, B^I)$ be the model joint distribution obtained when using all of these parameters together. In this work, we aim to optimise θ to transform the input images into feature vectors $F^{(\theta)}(I)$ such that all the domain specific joint distributions $P_D(F^{(\theta)}(I), C^I, B^I)$ converge to the single best (maximising the fit over ϕ and β) joint distribution $Q^{T^{(\phi)},R^{(\beta)}}(F^{(\theta)}(I), C^I, B^I)$. This will enable $F^{(\theta)}$, $T^{(\phi)}$ and $R^{(\beta)}$ to be optimised for domain invariant object detection. According to Bayes' theorem, in order to map the domain specific joint distributions $P_D(F^{(\theta)}(I), C^I, B^I)$ to a common $Q^{T^{(\phi)},R^{(\beta)}}(F^{(\theta)}(I), C^I, B^I)$, we need to map the domain specific conditionals $P_D(C^I, B^I|F^{(\theta)}(I))$ to a common $Q^{T^{(\phi)},R^{(\beta)}}(C^I, B^I|F^{(\theta)}(I))$ and the domain specific marginals $P_D(F^{(\theta)})$ need to be mapped onto a common $Q(F^{(\theta)})$.

In fact, many of the reported studies in the state-of-the-art (Chen et al. 2018; Muandet, Balduzzi, and Schölkopf 2013) attribute domain shift to the difference in marginals, while they assume conditionals to be stable across domains. The standard approach adopted by (Chen et al. 2018; Matsuura and Harada 2020) of equalising the marginals across domains is through an explicit domain discriminator $S^{(\psi)}$, which is trained by minimising the following negative discriminator ('domain adversarial' or *'dadv'*) loss (Goodfellow et al. 2014),

$$\begin{aligned}\min_{\theta}\max_{\psi} L_{dadv}(\theta, \psi) &= \sum_{D=1}^{N} \mathbb{E}_{P_D}\left[\log\left(S^{(\psi)}\left(F^{(\theta)}(I)\right)\right)\right] \\ &= \sum_{D=1}^{N}\sum_{j=1}^{M_D} \mathbf{d}_j^D . \log\left(S^{(\psi)}\left(F^{(\theta)}(I_j^D)\right)\right)\end{aligned} \tag{1}$$

where $\mathbf{d}_j$ is the one-hot vector encoding of the domain label of the j-th image sample, and M_D is the number of samples in the D-th domain. The maximisation is conducted with respect to parameters corresponding to the domain discriminator $S^{(\psi)}$, while the minimisation is with respect to the feature extractor $F^{(\theta)}$. This minmax game enables $F^{(\theta)}$ to learn features whose domain cannot be distinguished by any $S^{(\psi)}$. This implies that the optimisation in Eq. (1) will lead to equality in marginals,

$$\begin{aligned}P_i(F^{(\theta)}(I)) &= P_j(F^{(\theta)}(I)), \forall i, j \in \{1, \dots, N\} \\ &= Q(F^{(\theta)}(I))\end{aligned} \tag{2}$$

However, as pointed out by recent studies (Schölkopf et al. 2012; Janzing and Schölkopf 2010; Li et al. 2018c; Zhao et al. 2020), the stability of conditionals across domains cannot be guaranteed. Any method with the goal of achieving domain invariance needs to compensate for the variation in conditionals $P_D(C^I, B^I|F^{\theta}(I))$. In other words, the domain discriminator $S^{(\psi)}$ aids in achieving the invariance on the sample space $\mathbf{I} \times \mathbf{D}$ but not on $\mathbf{I} \times \mathbf{C} \times \mathbf{B} \times \mathbf{D}$. Moreover, the techniques proposed in recent studies (Zhao et al. 2020; Li et al. 2018c) are intended for classification and cannot be directly used to achieve generalised object detection.

In the next section, we describe the details of the proposed mathematical framework, to map the domain-specific conditionals $P_D(C^I, B^I|F^{(\theta)}(I))$ to a common $Q^{T^{(\phi)},R^{(\beta)}}(C^I, B^I|F^{(\theta)}(I))$. Our approach in conjunction with Eq. (1) leads to domain generalised object detection.

Proposed Method

The overview of the proposed DGFR-CNN is given in Fig. 2, where the Faster R-CNN is trained in conjunction with two additional modules related to class-conditional invariance and bounding box invariance. These modules aid to optimise the feature extractor so that the input images map onto a feature space where the detection is consistent across multiple domains. In this section, we elaborate on these new modules.

In addition to equalising marginals across domains (Eq. 1), we also aim to transform the domain-specific conditional distribution in every D-th domain to a common distribution $Q^{T^{(\phi)},R^{(\beta)}}(C^I, B^I|F^{\theta}(I))$ which is the main contribution of this paper. This can be done by minimising the KL divergence

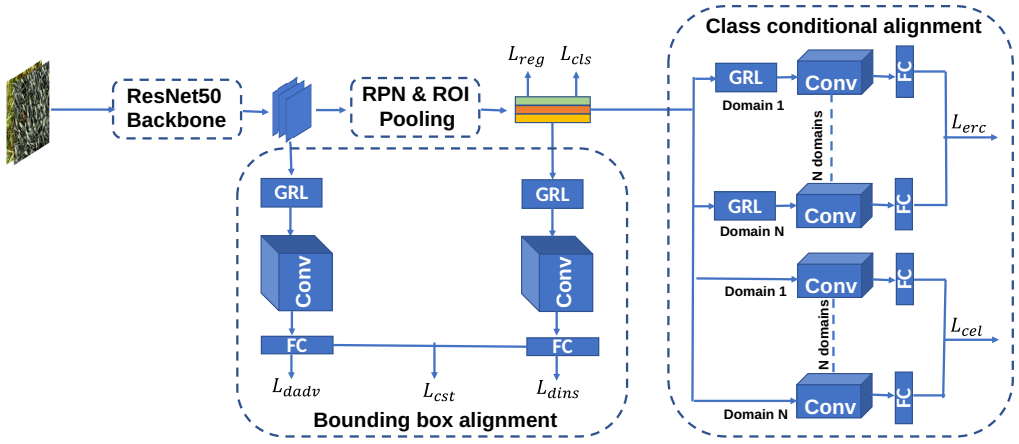


Figure 2: Overview of the proposed DGFR-CNN approach.

between all of the P_D and $Q^{T^{(\phi)},R^{(\beta)}}$:

$$\begin{aligned}\min_{\theta,\phi,\beta}\sum_{D=1}^{N} KL[P_D(C,B|F^{(\theta)}(I))||Q^{T^{(\phi)},R^{(\beta)}}(C,B|F^{(\theta)}(I))]\\ = \min_{\theta,\phi,\beta}\sum_{D=1}^{N} KL[P_D(C|B,F^{(\theta)}(I)||Q^{T^{(\phi)}}(C|B,F^{(\theta)}(I))]\\ + \sum_{D=1}^{N} KL[P_D(B|F^{(\theta)}(I))||Q^{R^{(\beta)}}(B|F^{(\theta)}(I))]\end{aligned} \tag{3}$$

where $Q^{T^{(\phi)}}(C^I|B^I,F^{(\theta)}(I))$ and $Q^{R^{(\beta)}}(B^I|F^{(\theta)}(I))$ denote respectively the distributions associated with the instance level classifier and the bounding box predictor. The minimisation of the first term in Eq. (3) trains the system to transform the input images into a feature space where the domain-specific instance level classifier will be consistent across the domains. The minimisation of the second term in Eq. (3) trains the system to transform input images into a feature space where the bounding box predictor is invariant to any individual domain. This implies that the optimisation in Eq. (3) along with Eq. (1) will result in a domain generalised trained model.

The first term in Eq. (3) can be further modified as:

$$\begin{aligned}\sum_{D=1}^{N}\mathbb{E}_{P_D}\left[\log\frac{P_D(C|B,F^{(\theta)}(I))}{Q^{T^{(\phi)}}(C|B,F^{(\theta)}(I))}\right]\\ = \sum_{D=1}^{N}\mathbb{E}_{P_D}\left[\log P_D(C|B,F^{(\theta)}(I))\right]\\ - \sum_{D=1}^{N}\mathbb{E}_{P_D}\left[\log Q^{T^{(\phi)}}(C|B,F^{(\theta)}(I))\right]\end{aligned}$$

The second term in Eq. (4) denotes the classification loss L_{cls}, while the first term is the sum of N negative conditional entropies $-H_D(C^I|B^I,F^{(\theta)}(I))$ for each domain. Minimising the negative conditional entropy is equivalent to maximising the conditional entropy $H_D(C^I|B^I,F^{(\theta)}(I))$. This implies that maximising the conditional entropy $H_D(C^I|B^I,F^{(\theta)}(I))$ for a specific domain D can increase the uncertainty in assigning the correct class label for the objects in the input image I. In order to maximise the conditional entropy $H_D(C^I|B^I,F^{(\theta)}(I))$ in Eq. (4), we extend the following theorem from (Zhao et al. 2020) for the object detection task:

Theorem 1: Assuming all the object classes are equally likely, maximising $H_D(C^I|B^I,F^{(\theta)}(I))$ is equivalent to minimising the Jensen-Shannon divergence between the conditional distributions $P_D(B^I,F^{(\theta)}(I)|C=j)_{j=1}^{K}$. The global minimum can be achieved if and only if:

$$\begin{aligned}P_D(B^I,F^{(\theta)}(I)|C^I=i) = P_D(B^I,F^{(\theta)}(I)|C^I=j),\\ \forall i,j\in\{1,\ldots,K\}.\end{aligned} \tag{4}$$

Even though this assumption can fail under a class imbalance scenario, balance can still be enforced through batch based biased sampling (proof is given in supplementary material).

The equality in conditionals $P_D(B,F_\theta(I)|C=j)$ for all the classes implies that the instance level features extracted are independent of the class labels. Inspired by Theorem 1 and the minmax game approach proposed in (Goodfellow et al. 2014; Gong et al. 2019a; Zhao et al. 2020), we introduce N classifiers, $\{T'^{(\phi'_D)}_D\}_{D=1\ldots N}$, each parameterised by a ϕ'_D, and propose the following loss function,

$$\begin{aligned}\min_{\theta}\max_{\{\phi'_D\}_{D=1}^{N}} L_{erc}(\theta,\phi'_i)\\ = \sum_{D=1}^{N}\mathbb{E}_{P_D}\left[\log Q_D^{T'^{(\phi'_D)}_D}(C|B,F^{(\theta)}(I))\right],\end{aligned} \tag{5}$$

where $Q_D^{T'^{(\phi'_D)}_D}(C|B,F^{(\theta)}(I))$ is the instance-level class conditional probability induced by classifier T'_D corresponding to the D-th domain.

To optimise the second term in Eq. (3), we adopt a strategy previously used by (Chen et al. 2018) for DA. Minimising

the KL divergence between the terms $P_D(B|F^{(\theta)}(I))$ and $Q^{R^{(\beta)}}(B|F^{(\theta)}(I))$ is equivalent to building a bounding box predictor independent of the domain label D. Rewriting the term $P_D(B|F^{(\theta)}(I))$ as $P(B|D, F^{(\theta)}(I))$, and using Bayes' theorem, gives:

$$\begin{aligned} &P(D|B, F^{(\theta)}(I))P(B|F^{(\theta)}(I)) \\ &= P(B|F^{(\theta)}(I), D)P(D|F^{(\theta)}(I)), \end{aligned} \tag{6}$$

where $P(D|B, F^{(\theta)}(I))$ represents the instance level domain label predictor, $P(B|F^{(\theta)}(I), D)$ is the domain specific bounding box predictor, and $P(D|F^{(\theta)}(I))$ is the image level domain label predictor. From Eq. (1), we can observe that if there is a consistency between the image and instance level domain label predictor then the bounding box predictor will be invariant to domains, i.e. $P(B|D, F^{(\theta)}(I)) = P(B|F^{(\theta)}(I))$.

The input $F^{(\theta)}(I[B])$ to the instance-level domain classifier will be the subset of image I features, computed at locations within the bounding box B by $F^{(\theta)}$. The loss function that is employed at the instance level domain classifier is:

$$L_{dins} = \sum_{|B|} \mathbb{E}\left[\log(P(D|B, F^{(\theta)}(I)))\right], \tag{7}$$

where $|B|$ denotes the total number of detected bounding boxes in the image I. As shown in Eq. (6), in order to achieve invariant bounding box prediction across domains, we need a consistency regularisation (Chen et al. 2018):

$$L_{cst} = \|\frac{1}{|B|}\sum_{i=1}^{|B|}(\mathbf{p}_i^{ins} - \mathbf{p}^{img})\|_2, \tag{8}$$

where $\mathbf{p}_i^{ins}$ and $\mathbf{p}^{img}$ denote respectively the probability scores corresponding to the instance and domain level classifier outputs. Combining the loss functions defined in Eqs. 1, 5, 7, 8 results in the following:

$$\begin{aligned} &\min_{(\theta,\phi)} \max_{(\psi_{img},\{\phi'_D\},\psi_{ins})} L(\theta, \phi, \psi_{img}, \psi_{ins}, \{\phi'_D\}, \beta) \\ &= L_{cls}(\theta, \phi) + L_{reg}(\theta, \beta) \\ &+\alpha_1 L_{dadv}(\theta, \psi_{img}) + \alpha_2 L_{dins}(\theta, \psi_{ins}) \\ &+\alpha_3 L_{cst}(\theta, \beta) + \alpha_4 L_{erc}(\theta, \{\phi'_D\}), \end{aligned} \tag{9}$$

where α_1, α_2, α_3, α_4 represent the regularisation constants.

It is important to note that the features learned via the maximisation of the domain specific classification loss L_{erc} can have a negative impact on the classifier $T^{(\phi)}$, and on the bounding box predictor $R^{(\beta)}$, and thus can result in instability during the mini-max optimisation of Eq. (9). To overcome this drawback, following (Zhao et al. 2020), we introduce N additional domain-specific classifiers, $\{T_D^{\dagger(\phi_D^\dagger)}\}_{D=1}^N$, with a new cross-entropy loss (L_{cel}):

$$\begin{aligned} &\min_{\theta,\{\phi_D\}_{D=1}^N} L_{cel}(\theta, \{\phi_D^\dagger\}_{D=1}^N) \\ &= -\sum_{D=1}^{N} \mathbb{E}_{P_D}[\log Q_D^{T_D^{\dagger(\phi_D^\dagger)}}(C|B, \bar{F}(I))] \\ &\quad -\sum_{D=1}^{N}\sum_{j=1,j\neq D} \mathbb{E}_{P_j}[\log Q_D^{\bar{T}_D}(C|B, F^{(\theta)}(I))], \end{aligned} \tag{10}$$

where $\bar{F}$ and $\bar{T}_D$ indicate that the parameters are fixed during training. Since these additional N classifiers are domain-specific, there is a high likelihood that they can prevent F_θ from learning domain invariant features, which is against the goal of our work. But at the same time, the inclusion of $\{T_D^\dagger\}$ can help in overcoming the instability introduced by T'_D. Hence, an effective strategy for training these domain specific classifiers is of crucial importance. We initially freeze θ and train each of the classifiers $\{T_D^\dagger\}$ by using data from the D-th domain. This step will aid $\{T_D^\dagger\}$ to learn only domain invariant features. In the next step, we fix all the N parameters $\phi_D^\dagger$ and fine tune θ so that a sample I_D from the D-th domain is classified accurately by all $\{T_D^\dagger\}_{D\neq d}$.

The final loss function that we use to train the system is:

$$\begin{aligned} &\min_{(\theta,\phi,\beta,\{\phi_D^\dagger\})} \max_{(\psi_{img},\{\phi'_D\},\psi_{ins})} \\ &L(\theta, \beta, \phi, \psi_{img}, \psi_{ins}, \{\phi'_D\}, \{\phi_D^\dagger\}) \\ &= L_{cls}(\theta, \phi) + L_{reg}(\theta, \beta) + \alpha_1 L_{dadv}(\theta, \psi_{img}) + \\ &\alpha_2 L_{dins}(\theta, \psi_{ins}) + \alpha_3 L_{cst}(\theta, \beta) + \alpha_4 L_{erc}(\theta, \{\phi'_D\}) \\ &+\alpha_5 L_{cel}(\theta, \{\phi_D^\dagger\}), \end{aligned} \tag{11}$$

where α_5 is the regularisation constant associated with the additional N domain specific classifiers $\{T_D^\dagger\}$. The complete training procedure is described in Algorithm 1, where we train the main detector in conjunction with additional regularisation terms to achieve domain invariant bounding box prediction as well as class-conditional invariance.

Experimental Validation

Datasets. We demonstrate the generalisation ability of our approach on the following four popular multi-source object detection datasets.

GWHD (David et al. 2021): This dataset comprises of a total of 6000 images corresponding to wheat heads (resolution: 1024 $\times$ 1024 pixels) acquired across 47 different sessions; with each being restricted to a single domain/farm. The training set has 18 domains with a total of 2943 images while the validation set contains samples captured across 8 different sessions with 1424 images and the test set has data from 21 different sessions with a total of 1434 images. Here we assume a unique domain label for each of the sessions. A few of the domains are shown in Fig. 1 illustrating the high level of domain shift across acquisition locations.

Algorithm 1: Training strategy for domain generalised object detection.

Input: $\{X_D\}_{D=1}^{N}$, N domain training datasets
Input: α_1, α_2, α_3, α_4, α_5
Output: $F, B, T, D, D_{ins}, \{T_D\}_{D=1}^{N}, \{T'_D\}_{D=1}^{N}$
while *iter* $\leq$ *MAX_EPOCHS* **do**
  while *batch* $\leq$ *MAX_BATCHES* **do**
    Sample random batch of images from training data
    Update θ, β, ϕ in Eq. (11)
    Using fixed θ, update ψ_{img} and ψ_{ins} in Eq. (11)
    Sample random batch of images from training data
    Update θ, β, ϕ in Eq. (11)
    Using fixed θ, update $\{\phi_D^{\dagger}\}$ in Eq. (11) (minimises first term of Eq. 10)
    Sample random batch of images from training data
    Update θ, β, ϕ in Eq. (11)
    Update θ, $\{\phi'_D\}$ in Eq. (11)
    Sample random batch of images from training data
    Update θ, β, ϕ in Eq. (11)
    Using fixed $\{\phi_D^{\dagger}\}$, update θ in Eq. (11) (minimises second term of Eq. 10)
  end
end

Cityscapes (C), *Foggy Cityscapes (FC)*, *Rainy Cityscapes (RC)*: Cityscapes (Cordts et al. 2016) deals with the semantic understanding of urban street scenes. It has a total of 2975 training (from 18 cities) and 500 validation images (from 3 cities). The fog in *Foggy Cityscapes* (Sakaridis, Dai, and Van Gool 2018) images is synthetically created using a standard fog image formation model (Middleton 1957) with an airlight coefficient of 0.02. Similarly, 36 different patterns of rain were introduced in 295 images of *Cityscapes* to create the *Rainy Cityscapes* dataset (Hu et al. 2019).

BDD100k (B) (Yu et al. 2020): This dataset has 100K diverse video clips where each clip is of 40 seconds. The annotations are collected on six different scene types, six different weather conditions, three distinct times of the day. Unlike (Sakaridis, Dai, and Van Gool 2018; Hu et al. 2019), the fog and rain in the images of *BDD100k* is real. The train, validation, and test splits has 70K, 10K, 20K images, respectively. In our experiments, we use only train and validation splits of this dataset due to the lack of test set annotations.

Sim10k (S) (Johnson-Roberson et al. 2017): This data is generated by capturing the snapshots of Grand Threft Auto V (GTA-V) game. There is no official train and validation splits available for this dataset and hence we randomly split it into 8K images as training set and the rest as validation split. Four different weather types will appear in this dataset.

Experiments. We evaluate the generalisation abilities of the proposed DGFR-CNN through the following experiments:

(i) *DG (multi-class, single target domain):* We evaluate the generalisation ability of DGFR-CNN when the nature of domain shift during training and testing is different. During training, the domain shift comes from the acquisition location in *Cityscapes*, while shift during testing on *Foggy/Rainy Cityscapes* is manifested in the form of weather conditions.

(ii) *DG (single-class, multiple target domains):* The *GWHD* dataset allows evaluating the generalisation on a scenario with multiple target domains, where the shift in both the training and target domains comes from the acquisition location.

(iii) *DG (single-class, single target domain):* We use all the autonomous driving datasets together to evaluate Faster R-CNN when the source domains are related but do not follow a uniform standard.

(iv) *DA (single source and target domains):* To analyse the performance of DGFR-CNN in the DA setting, a simplified model is used which uses only the bounding box alignment module. The class conditional alignment is removed, as only the ulabelled target images are available during training.

Network Architecture. We use a Faster R-CNN detector with a ResNet50 backbone initialised with pre-trained ImageNet weights. The output of the backbone network is fed as input to domain adversarial network ($S^{(\psi)}$) while the output of ROI Pooling layer is fed as an input to instance level domain classifier ($S_{ins}^{(\psi)}$), $2N$ domain specific classifiers (T'_D and $T_D^{\dagger}$). All terms in the loss function described in Eq. (11) correspond to either a domain or object classifier. We use cross-entropy to train each of these classifier modules.

Training details. From empirical observations, the regularisation constants were set to $\alpha_1 = 1$, $\alpha_2 = 0.1$, $\alpha_3 = 1$, $\alpha_4 = 0.001$, and $\alpha_5 = 0.05$. We used early stopping with a patience of 10 epochs. AdamW (weight decay = 0.0005, learning rate = 0.001, batchsize=2) has been used as optimiser while training with GWHD and Stochastic Gradient Descent (SGD) (weight decay = 0.0005, momentum=0.9, learning rate=2×10^{-3}, batchsize=2) has been used for other datasets (Cityscapes, BDD100K, Sim10K). The experiments were implemented using the PyTorch deep learning framework and Torchvision Faster R-CNN library on a NVIDIA RTX 3090 GPU with 24GB of GPU memory.

Metrics. Following (David et al. 2021), we use *weighted average domain accuracy* (WADA) to report the performance of our approach on the OOD test set of *GWHD*. For the rest of the datasets, we use *mean average precision* (mAP).

Quantitative Analysis

In experiment (i), we present the performance of proposed approach on *Cityscapes* dataset (Table 1) where the training is done using city information as domain label and tested on *Foggy/Rainy Cityscapes*. None of the existing DG approaches for object detection use citylabel as domain label and hence we limit our comparisons to the methods from which our approach was inspired from. It can be seen that the proposed architecture performs the best in majority of object categories for *Foggy Cityscapes* while a better overall performance for *Rainy Cityscapes*. This signifies the need for compensating both the covariate and concept shifts in a multi-source domain scenario.

	Person	Rider	Car	Truck	Bus	Mcycle	Train	Bicycle	mAP
Cityscapes Foggy									
Faster R-CNN	26.9	38.2	35.6	18.3	32.4	**25.8**	**9.6**	28.6	26.9
BBA	**50.34**	43.49	75.93	20.62	28.00	22.37	4.55	42.50	35.97
CCA	45.27	**48.90**	72.44	23.75	35.11	24.73	6.82	43.41	37.55
DGFR-CNN (ours)	47.17	47.84	**77.43**	**24.79**	**36.00**	19.62	9.09	**48.72**	**38.90**
Cityscapes Rainy									
Faster R-CNN	**42.69**	67.62	75.43	19.64	50.19	13.00	**4.66**	50.54	40.47
BBA	41.45	**72.89**	69.88	15.08	55.00	**22.55**	3.17	48.28	41.03
CCA	42.25	69.61	**78.53**	25.83	**60.71**	20.00	3.74	52.16	44.22
DGFR-CNN (ours)	42.66	71.64	76.83	**29.50**	60.02	20.3	2.98	**53.43**	**44.66**

Table 1: Results for the *Foggy* and *Rainy Cityscapes* datasets. BBA: Bounding Box Alignment. CCA: Class Conditional Alignment. The best results per class and overall are highlighted in bold.

	GWHD	(S,C) $\rightarrow$ B	(S, B) $\rightarrow$ C	(B, C) $\rightarrow$ S	L_{cls}	L_{reg}	L_{dadv}	L_{dins}	L_{cst}	L_{erc}	L_{cel}
Faster R-CNN	53.73	50.52	71.43	61.84	+	+	-	-	-	-	-
BBA	54.48	35.76	70.02	54.30	+	+	+	+	+	-	-
CCA	53.65	38.07	71.27	59.50	+	+	-	-	-	+	+
DGFR-CNN	**54.92**	**52.02**	**71.78**	**62.50**	+	+	+	+	+	+	+

Table 2: Quantitative analysis for proposed approach in DG setting. The symbol '+' indicates inclusion of loss component while '-' indicates exclusion of loss component. Generalisation performance of the proposed approach across: *Sim10K* (S), *Cityscapes* (C) and *BDD100K* (B). The left and right sides of $\longrightarrow$ indicate the source and target datasets, respectively. The best results are highlighted in bold.

	Person	Rider	Car	Truck	Bus	Mcycle	Train	Bicycle	mAP
Source-only	26.9	38.2	35.6	18.3	32.4	25.8	9.6	28.6	26.9
GPA (Xu et al. 2020b)	32.9	46.7	54.1	24.7	45.7	41.1	32.4	38.7	39.5
DIDN (Lin et al. 2021)	38.3	44.4	51.8	28.7	**53.3**	34.7	32.4	40.4	40.5
DSS (Wang et al. 2021)	42.9	**51.2**	53.6	**33.6**	49.2	18.9	**36.2**	41.8	40.9
SDA (Rezaeianaran et al. 2021)	38.8	45.9	57.2	29.9	50.2	**51.9**	31.9	40.9	**43.3**
DGFR-CNN (ours)	**53.9**	49.6	**76.9**	23.13	28.67	29.30	9.09	**54.14**	40.6

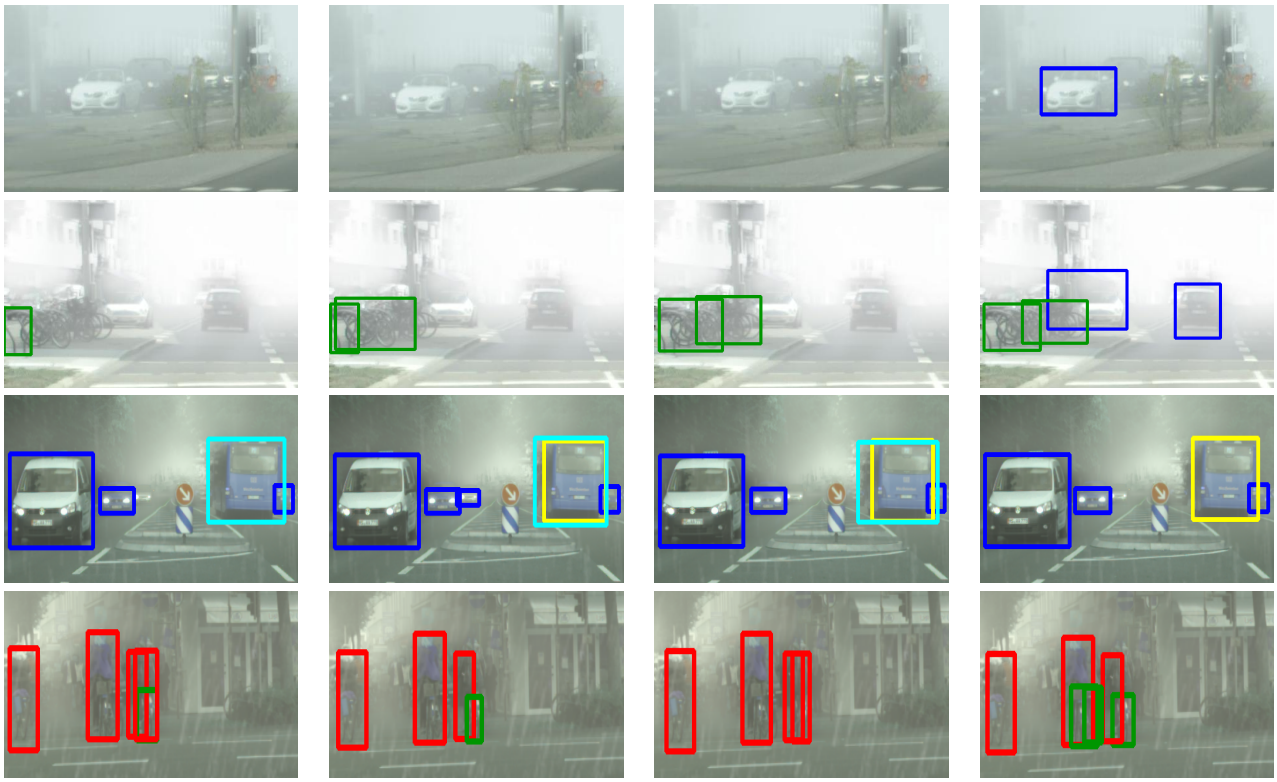
Table 3: Performance of proposed approach in DA setting while adapting from *Cityscapes* to *Foggy Cityscapes*. The best results per class and overall are highlighted in bold.

Ablation Study Table 2 (experiments (ii) and (iii)) shows the quantitative analysis for the official out-of-distribution (OOD) test split of *GWHD*. Also, we evaluate the generalisation ability of our detector to a completely new dataset where we use two among the datasets from *Sim10K*, *Cityscapes*, *BDD100K* as source and the target as the other dataset. It can be seen that our approach outperforms the baseline Faster R-CNN used in WildS benchmark (Koh et al. 2021). The second and third rows report the influence of individual components used in the proposed architecture while the last row indicates the effect of complete architecture. It can be seen that the proposed approach improvises over the baseline as well as when the individual components used alone. This signifies the need for the additional constraints which regularises the main detection loss so as to equalise the conditional distributions of class-labels and bounding box detector across the domains. Also, this highlights the need for addressing both the concept and covariate shifts rather than covariate shift alone.

Table 3 (experiment (iv)) shows the performance of proposed approach while adapting from *Cityscapes* to *Foggy Cityscapes* in DA setting. We compare against a number of representative state-of-the-art DA (Xu et al. 2020b; Wang et al. 2021; Rezaeianaran et al. 2021) and DG (Lin et al. 2021) approaches. The proposed DGFR-CNN improves upon the results of a DG-based approach (Lin et al. 2021) in the DA setting, and is competitive with or improves upon most other DA approaches. While we do not reach the performance of SDA (Rezaeianaran et al. 2021), we note that this is, on one hand, an approach specifically designed for DA, which has recently outperformed competing DA approaches by a large margin. On the other hand, our DGFR-CNN is designed to handle domain shift using different input data available in the DG setting, and was simplified by removing class conditional alignment for the DA experiment.

Qualitative Results

Cityscapes Fig. 3 presents the qualitative analysis of the proposed approach against the outputs of BBA and CCA alone as well as the baseline Faster R-CNN. The first two



(a) (b) (c) (d)

Figure 3: Qualitative Results for Cityscapes. Outputs of: (a) Faster R-CNN, (b) BBA, (c) CCA, and (d) Proposed approach. (Readers are requested to zoom in to visualise the pictures.). The colour codes used in this figure are indicated in Table 4.

Figure 4: Qualitative Analysis for GWHD. Predictions are indicated in red colour boxes while the ground truth are indicated in blue coloured boxes.

Category	Colour code
Person	
Rider	
Car	
Truck	
Bus	
Motorcycle	
Bicycle	

Table 4: Colour codes for Cityscapes.

rows corresponds to Foggy-Cityscapes while the third and fourth rows shows the detections on Rainy-Cityscapes. In the first row of Fig. 3, it can be seen that the Car on the left of the road has been very well detected by our approach in spite of the dense fog being present in the scene while the state-of-the-art approaches fail to detect. Similarly, our approach robustly detects the Car in the left lane (second row Fig. 3 (d)). In the third row, the bus on the right hand side of the image is correctly detected by our approach where as the baseline or the individual components (BBA, CCA) either incorrectly detects the bus or multiple labels are assigned to that specific object. It has to be noted that the car in the dense fog on the left side of the image is correctly detected by BBA. However, the cost of incorrectly classifying a larger

size object is higher than missing a smaller size object. The bicycles in the last row of Fig. 3 are detected well by using our proposed approach over the existing techniques. We are able to accurately localise the objects in-spite of the dense fog/rain present in the scene. Note that we do not include foggy or rainy images in the training set. This emphasise the need and importance of imposing class-conditional and bounding box invariance across all source domains to enable robust detection on the unseen target data.

GWHD 2021 Fig. 4 shows the result of wheat head detection by using our proposed approach. It can be seen that our detection is able to localise the wheat heads accurately in-spite of the wide visual diversity of the wheat heads and illumination conditions.

Conclusion

We have proposed a Domain Generalised Faster R-CNN, which improves the generalisation ability of Faster R-CNN architecture. To perform consistently across domains, the proposed method does not assume equality between class conditional probabilities, but instead introduces the consistency regularisation term along with the class entropy regulariser to align the feature distribution resulting from different domains. The method has been validated by showing performance improvements when used with Faster R-CNN, on four standard object detection datasets related to autonomous driving and agriculture.

Acknowledgments

This work was supported by Lincoln Agri-Robotics as part of the Expanding Excellence in England (E3) Programme. E. Aptoula was partly supported by the TÜBA GEBIP'21 Award

References

Arjovsky, M.; Bottou, L.; Gulrajani, I.; and Lopez-Paz, D. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Blanchard, G.; Lee, G.; and Scott, C. 2011. Generalizing from several related classification tasks to a new unlabeled sample. *Advances in neural information processing systems*, 24: 2178–2186.

Cai, M.; Luo, M.; Zhong, X.; and Chen, H. 2021. Uncertainty-Aware Model Adaptation for Unsupervised Cross-Domain Object Detection. *arXiv preprint arXiv:2108.12612*.

Chen, Y.; Li, W.; Sakaridis, C.; Dai, D.; and Van Gool, L. 2018. Domain adaptive Faster R-CNN for object detection in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3339–3348.

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The Cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Cision and Pattern Recognition*, 3213–3223.

Dai, J.; Li, Y.; He, K.; and Sun, J. 2016. R-FCN: Object Detection via Region-based Fully Convolutional Networks. In Lee, D.; Sugiyama, M.; Luxburg, U.; Guyon, I.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Dalal, N.; and Triggs, B. 2005. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, 886–893. Ieee.

David, E.; Serouart, M.; Smith, D.; Madec, S.; Velumani, K.; Liu, S.; Wang, X.; Pinto, F.; Shafiee, S.; Tahir, I. S.; et al. 2021. Global Wheat Head Detection 2021: an improved dataset for benchmarking wheat head detection methods. *Plant Phenomics*, 2021.

Fang, C.; Xu, Y.; and Rockmore, D. N. 2013. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *Proceedings of the IEEE International Conference on Computer Vision*, 1657–1664.

Felzenszwalb, P. F.; Girshick, R. B.; McAllester, D.; and Ramanan, D. 2009. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9): 1627–1645.

Ghifary, M.; Kleijn, W. B.; Zhang, M.; and Balduzzi, D. 2015. Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE International Conference on Vision*, 2551–2559.

Girshick, R. 2015. Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, 1440–1448.

Girshick, R.; Donahue, J.; Darrell, T.; and Malik, J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 580–587.

Gong, M.; Xu, Y.; Li, C.; Zhang, K.; and Batmanghelich, K. 2019a. Twin auxiliary classifiers GAN. *Advances in neural information processing systems*, 32: 1328.

Gong, R.; Li, W.; Chen, Y.; and Gool, L. V. 2019b. Dlow: Domain flow for adaptation and generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2477–2486.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*.

He, Z.; and Zhang, L. 2019. Multi-adversarial Faster R-CNN for unrestricted object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6668–6677.

Hendrycks, D.; and Dietterich, T. 2019. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*.

Hsu, C.-C.; Tsai, Y.-H.; Lin, Y.-Y.; and Yang, M.-H. 2020a. Every pixel matters: Center-aware feature alignment for domain adaptive object detector. In *European Conference on Computer Vision*, 733–748. Springer.

Hsu, H.-K.; Yao, C.-H.; Tsai, Y.-H.; Hung, W.-C.; Tseng, H.-Y.; Singh, M.; and Yang, M.-H. 2020b. Progressive domain adaptation for object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 749–757.

Hu, S.; Zhang, K.; Chen, Z.; and Chan, L. 2020. Domain generalization via multidomain discriminant analysis. In *Uncertainty in Artificial Intelligence*, 292–302. PMLR.

Hu, X.; Fu, C.-W.; Zhu, L.; and Heng, P.-A. 2019. Depth-Attentional Features for Single-Image Rain Removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Janzing, D.; and Schölkopf, B. 2010. Causal inference using the algorithmic Markov condition. *IEEE Transactions on Information Theory*, 56(10): 5168–5194.

Jocher, G.; Stoken, A.; Borovec, J.; NanoCode012; ChristopherSTAN; Changyu, L.; Laughing; tkianai; Hogan, A.; lorenzomammana; yxNONG; AlexWang1900; Diaconu, L.; Marc; wanghaoyang0106; ml5ah; Doug; Ingham, F.; Frederik; Guilhen; Hatovix; Poznanski, J.; Fang, J.; Yu, L.; changyu98; Wang, M.; Gupta, N.; Akhtar, O.; PetrDvoracek; and Rai, P. 2020. ultralytics/yolov5: v3.1 - Bug Fixes and Performance Improvements.

Johnson-Roberson, M.; Barto, C.; Mehta, R.; Sridhar, S. N.; Rosaen, K.; and Vasudevan, R. 2017. Driving in the Matrix: Can virtual worlds replace human-generated annotations for real world tasks? In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 746–753. IEEE.

Khirodkar, R.; Yoo, D.; and Kitani, K. 2019. Domain randomization for scene-specific car detection and pose estimation. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1932–1940. IEEE.

Khodabandeh, M.; Vahdat, A.; Ranjbar, M.; and Macready, W. G. 2019. A robust learning approach to domain adaptive object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 480–490.

Khosla, A.; Zhou, T.; Malisiewicz, T.; Efros, A. A.; and Torralba, A. 2012. Undoing the damage of dataset bias. In *European Conference on Computer Vision*, 158–171. Springer.

Koh, P. W.; Sagawa, S.; Xie, S. M.; Zhang, M.; Balsubramani, A.; Hu, W.; Yasunaga, M.; Phillips, R. L.; Gao, I.; Lee, T.; et al. 2021. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, 5637–5664. PMLR.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25: 1097–1105.

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, H.; Pan, S. J.; Wang, S.; and Kot, A. C. 2018b. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5400–5409.

Li, W.; Li, F.; Luo, Y.; Wang, P.; et al. 2020. Deep domain adaptive object detection: A survey. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, 1808–1813. IEEE.

Li, W.; Liu, X.; and Yuan, Y. 2022. SIGMA: Semantic-complete Graph Matching for Domain Adaptive Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5291–5300.

Li, Y.; Tian, X.; Gong, M.; Liu, Y.; Liu, T.; Zhang, K.; and Tao, D. 2018c. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 624–639.

Lin, C.; Yuan, Z.; Zhao, S.; Sun, P.; Wang, C.; and Cai, J. 2021. Domain-invariant disentangled network for generalizable object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8771–8780.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2117–2125.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, 740–755. Springer.

Liu, H.; Song, P.; and Ding, R. 2020a. Towards domain generalization in underwater object detection. In *2020 IEEE International Conference on Image Processing (ICIP)*, 1971–1975. IEEE.

Liu, H.; Song, P.; and Ding, R. 2020b. WQT and DG-YOLO: Towards domain generalization in underwater object detection. *arXiv preprint arXiv:2004.06333*.

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; and Berg, A. C. 2016. SSD: Single shot multibox detector. In *European Conference on Computer Vision*, 21–37. Springer.

Matsuura, T.; and Harada, T. 2020. Domain generalization using a mixture of multiple latent domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 11749–11756.

Middleton, W. E. K. 1957. Vision through the atmosphere. In *Geophysik II/Geophysics II*, 254–287. Springer.

Muandet, K.; Balduzzi, D.; and Schölkopf, B. 2013. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, 10–18. PMLR.

Recht, B.; Roelofs, R.; Schmidt, L.; and Shankar, V. 2019. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, 5389–5400. PMLR.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 779–788.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28: 91–99.

Rezaeianaran, F.; Shetty, R.; Aljundi, R.; Reino, D. O.; Zhang, S.; and Schiele, B. 2021. Seeking similarities over differences: Similarity-based domain alignment for adaptive object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9204–9213.

Rodriguez, A. L.; and Mikolajczyk, K. 2019. Domain adaptation for object detection via style consistency. In *Proceedings BMVC*.

Rosenfeld, E.; Ravikumar, P.; and Risteski, A. 2020. The risks of invariant risk minimization. *arXiv preprint arXiv:2010.05761*.

RoyChowdhury, A.; Chakrabarty, P.; Singh, A.; Jin, S.; Jiang, H.; Cao, L.; and Learned-Miller, E. 2019. Automatic adaptation of object detectors to new domains using self-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 780–790.

Saenko, K.; Kulis, B.; Fritz, M.; and Darrell, T. 2010. Adapting visual category models to new domains. In *European Conference on Computer Vision*, 213–226. Springer.

Saito, K.; Ushiku, Y.; Harada, T.; and Saenko, K. 2019. Strong-weak distribution alignment for adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6956–6965.

Sakaridis, C.; Dai, D.; and Van Gool, L. 2018. Semantic Foggy Scene Understanding with Synthetic Data. *International Journal of Computer Vision*, 126(9): 973–992.

Schölkopf, B.; Janzing, D.; Peters, J.; Sgouritsa, E.; Zhang, K.; and Mooij, J. 2012. On causal and anticausal learning. *arXiv preprint arXiv:1206.6471*.

Shankar, S.; Piratla, V.; Chakrabarti, S.; Chaudhuri, S.; Jyothi, P.; and Sarawagi, S. 2018. Generalizing across domains via cross-gradient training. In *ICLR*.

Sindagi, V. A.; Oza, P.; Yasarla, R.; and Patel, V. M. 2020. Prior-based domain adaptive object detection for hazy and rainy conditions. In *European Conference on Computer Vision*, 763–780. Springer.

Tan, M.; Pang, R.; and Le, Q. V. 2020. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10781–10790.

Tobin, J.; Fong, R.; Ray, A.; Schneider, J.; Zaremba, W.; and Abbeel, P. 2017. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 23–30. IEEE.

Torralba, A.; and Efros, A. A. 2011. Unbiased look at dataset bias. In *CVPR 2011*, 1521–1528. IEEE.

Viola, P.; and Jones, M. 2001. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, I–I. Ieee.

Wang, Y.; Zhang, R.; Zhang, S.; Li, M.; Xia, Y.; Zhang, X.; and Liu, S. 2021. Domain-specific suppression for adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9603–9612.

Wu, A.; Han, Y.; Zhu, L.; and Yang, Y. 2021. Instance-Invariant Domain Adaptive Object Detection via Progressive Disentanglement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Xu, C.-D.; Zhao, X.-R.; Jin, X.; and Wei, X.-S. 2020a. Exploring categorical regularization for domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11724–11733.

Xu, M.; Wang, H.; Ni, B.; Tian, Q.; and Zhang, W. 2020b. Cross-domain detection via graph-induced prototype alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12355–12364.

Xu, Z.; Li, W.; Niu, L.; and Xu, D. 2014. Exploiting low-rank structure from latent domains for domain generalization. In *European Conference on Computer Vision*, 628–643. Springer.

Yao, X.; Zhao, S.; Xu, P.; and Yang, J. 2021. Multi-Source Domain Adaptation for Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3273–3282.

Yu, F.; Chen, H.; Wang, X.; Xian, W.; Chen, Y.; Liu, F.; Madhavan, V.; and Darrell, T. 2020. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2636–2645.

Yue, X.; Zhang, Y.; Zhao, S.; Sangiovanni-Vincentelli, A.; Keutzer, K.; and Gong, B. 2019. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2100–2110.

Zhao, S.; Gong, M.; Liu, T.; Fu, H.; and Tao, D. 2020. Domain generalization via entropy regularization. *Advances in Neural Information Processing Systems*, 33.

Zhou, K.; Liu, Z.; Qiao, Y.; Xiang, T.; and Loy, C. C. 2021. Domain generalization: A survey. *arXiv preprint arXiv:2103.02503*.

Zhu, X.; Pang, J.; Yang, C.; Shi, J.; and Lin, D. 2019. Adapting object detectors via selective cross-domain alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 687–696.