2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

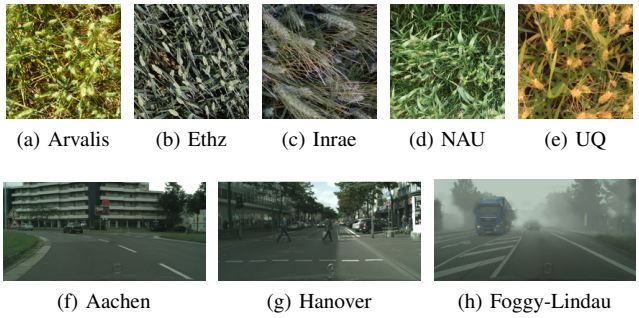
Domain Generalised Fully Convolutional One Stage Detection

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

Abstract— **Real-time vision in robotics plays an important role in localising and recognising objects. Recently, deep learning approaches have been widely used in robotic vision. However, most of these approaches have assumed that training and test sets come from similar data distributions, which is not valid in many real world applications. This study proposes an approach to address domain generalisation (i.e. out-of-distribution generalisation, OODG) where the goal is to train a model via one or more source domains, that will generalise well to unknown target domains using single stage detectors. All existing approaches which deal with OODG either use slow two stage detectors or operate under the covariate shift assumption which may not be useful for real-time robotics. This is the first paper to address domain generalisation in the context of single stage anchor free object detector FCOS 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 Fully Convolutional One Stage (DGFCOS) detection 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 and is able to run in real-time for robotics.**

I. INTRODUCTION

Real-time object detection is a critical task for mobile robots, and benchmark performances have increased recently using deep learning approaches [1]–[7]. While offline systems based on separated training and test sets can learn to make good detections for machine vision benchmarks, an ongoing problem in robotics is that data encountered in the real world rarely shares the same statistics with data previously recorded for use in model training. Factors such as viewpoint, background, weather, and image quality all create variations in object appearance. While it is sometimes possible to collect many different training sets to cover some settings of some factors, it remains difficult to learn a single model which can generalise over all possible factors and settings including to the previously unseen ones which are actually encountered after training, in the real world, by the real-time robot. For example, the autonomous farming and autonomous driving examples in Fig. 1 include factors due to robots moving to previously unseen countries, crop types, and weather conditions.



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

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

Fig. 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.

This problem has become known as *domain shift* and has begun to be measured via degradation of model performance at deployment [8], [9]. The task of compensating for domain shift is known as *domain generalisation*. Domain generalisation aims to consider whatever factors and their values are present in multiple, different sets of training data – such as collected previously by the robot under different geographic or time conditions – and *unlearn* the effects of these factors while learning to make object detections independently of them. This creates more favourable conditions for later real-time detections in new environments, which have different factor levels, in which detection is invariant to their changes.

Domain Generalisation is distinct from other approaches to domain shift, including periodic retraining of models to update them to gradual, continuous shifts of factors over time, and from Domain Adaptation methods [10]–[19] which learn a range of model parameters suited to a known range of domains, then search this space for the best fit on arrival in a new domain.

In our recent work [20], we proposed a Domain Generalisation (DG) modification for improving the offline detection performance of Faster R-CNN on unseen target data, by extending the training of the existing detector with additional entropy based regularisation terms. We used standard Faster-RCNN [4] as the base detector, and new loss functions were created by combining elements from a previous domain-adaptation method used previously for detection [10] and a generalisation method used previously for classification [21]. In the present paper, we adopt our previously proposed DG

*The code to replicate the results in this work can be found at https://github.com/karthikiitm87/domain-generalisation.

[1]Karthik Seemakurthy is with Lincoln Institute of Agri-Food Technology, University of Lincoln, UK. kseemakurthy@lincoln.ac.uk

[2]Petra Bosilj and Charles Fox are with School of Computer Science, University of Lincoln. pbosilj@lincoln.ac.uk, chfox@lincoln.ac.uk

[3]Erchan Aptoula is with Faculty of Engineering and Natural Sciences (VPALab), Sabanci University, Türkiye.

framework [20] to FCOS [22], which is a faster, real-time, single-stage, anchor free detector, and propose a Domain Generalised FCOS (DGFCOS) which has the ability to suppress domain specific features while enhancing the domain invariant features.

Experimental validation has been conducted on four datasets: Sim10K [23], Cityscapes [24], BDD100K [25], and GWHD [26]. GWHD is the only object detection dataset among a number of established and newly proposed DG benchmarks [27]–[31], 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 DGFCOS 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).

II. Related Works

The proposed system builds upon the following work in object detection, domain shift and real time applications:

Object detection. Classical approaches to object detection relied on handcrafted features and formulated object detection as a sliding window classification problem [32]–[34]. The empirical success of convolutional neural networks (CNNs) which automatically extract high performing features [35], [36] 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 [37] and have only recently reached the performance of two-stage approaches [1]. Two-stage approaches developed from the Region-based CNN (R-CNN) family [4] initially generate possible region proposals and features corresponding to the objects of interest which are further used for classification and localisation in second stage, thus resulting in an end-to-end trainable system. However, all of these architectures and a number of follow-up works [3], [5], [6] 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. Also, the operation of most popular detectors like Faster R-CNN [4] rely on the framework of anchor based object detection which suffers from some drawbacks [22], [38], [39]: 1) Due to fixed anchor sizes and aspect ratios, these detectors tend to exhibit a poor generalisation ability on the unseen test set with objects that have large shape variations. 2) Usage of excessive negative samples aggravates the class imbalance problem during training. 3) Hyper-parameters need to be carefully tuned in anchor based approaches to achieve optimal detection performance. 4) Anchor based detectors have more number of hyper-parameters when compared to anchor free approaches and thus increasing the inference time. Hence the high computational complexity and poor generalisation ability of anchor based detectors makes them less preferred for robotics based real time applications.

Domain shift. Unsupervised Domain Adaptation (UDA) and Domain Generalisation (DG) are two popular approaches to address domain shift problem in object detection. The prior assumes the availability of unlabelled target data while the later is more suited for real time object detection where the access to target data is restricted. According to the recent surveys of domain shift [28], [29], DA is still the more common approach for addressing domain shift in object detection, with very little work on DG for object detection [40]–[43]. Most of these techniques assume the disputed covariate shift case [21], [44]–[46]. In our recent approach [20], we demonstrate an improved modification of Faster R-CNN detector which addresses domain shift by compensating for both concept and covariate shifts together by using additional entropy based regularisation terms. Most of these techniques which attempt to address domain shift for object detection validate the performance using Faster R-CNN as the base detector. However, as mentioned earlier, anchor based approaches might not be a good fit for real time applications.

Recently, among the anchor free single stage detectors, FCOS [22] has been the basis for many deep network architectures used in real time robotic applications in agriculture [39], sports [47], medical [48], remote sensing [49], [50], and grasping [51]. Most of these approaches either focus to address occlusion or to enhance the detection of high speed and tiny objects. There are very few techniques [16] which attempt to improve the generalisation ability of FCOS, and these operate only in DA setting which is not relevant to robotic applications where the target data is not available during training. In this paper, for the first time, we propose Domain Generalised Fully Convolutional One Stage (DGFCOS) detector where we adopt our previously proposed DG framework [20] to improve generalisation ability of FCOS.

III. Mathematical preliminaries

In this section, we initially describe the mathematical preliminaries corresponding to FCOS detector [22] followed by a brief description of our previous DG framework [20]. The following section will then show how to combine them.

A. FCOS detector

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. According to [22], ResNet50 is used as the backbone of FCOS detector which has five levels of feature extraction layers. We assume $\mathbf{I} \in \mathbb{R}^{N_b \times 3 \times H \times W}$ to be the batch of N_b input images (each in 3 channel RGB, $W \times H$ pixels) fed to $F^{(\theta)}$ and $\mathbf{F}_i \in \mathbb{R}^{N_{f_i} \times H_{f_i} \times W_{f_i}}$ be the corresponding output at i^{th} level of ResNet50. The top three levels of features $(\mathbf{F}_3, \mathbf{F}_4, \mathbf{F}_5)$ from $F^{(\theta)}$ are fed as input to a feature pyramid network (FPN) to extract features at multiple different scales. Let $\mathbf{P}_j$ denote the output of FPN at j^{th} level where the index $j \in [3, 7]$. The relation between the features at different levels of $F^{(\theta)}$ and FPN can be mathematically

modelled by using the following equations,

$$\begin{aligned} &\mathbf{P}_5 = f_5(\mathbf{F}_5), \quad \mathbf{P}_4 = f_4(\mathbf{F}_4) \oplus p_4(\mathbf{P}_5^u) \\ &\mathbf{P}_3 = f_3(\mathbf{F}_3) \oplus p_3(\mathbf{P}_4^u), \quad \mathbf{P}_6 = \mathbf{P}_5^d, \quad \mathbf{P}_7 = \mathbf{P}_6^d \end{aligned} \tag{1}$$

where $\oplus$ denotes the element wise addition, $\mathbf{P}_j^u$ and $\mathbf{P}_j^d$ denote the corresponding stride two upsampled and downsampled versions of $\mathbf{P}$, respectively. According to [5], f_j and p_j are modelled as 1×1 and 3×3 convolution layers respectively. In FCOS, features extracted at each of the five layers in FPN are fed as input to heads which essentially computes a classification label and regress the bounding box offsets including an estimate for centeredness measure at every location (x, y) of the FPN feature map. The following is the expression for the loss function which is used to train $T^{(\phi)}$ and $R^{(\beta)}$ [22],

$$\begin{aligned} L(\{\mathbf{p}_{(x,y)}\}, \{\mathbf{t}_{(x,y)}\}) = \frac{1}{N_{pos}} \sum_{(x,y)} L_{cls}(\mathbf{p}_{x,y}, c^*_{x,y}) \\ + \frac{\lambda}{N_{pos}} \sum_{(x,y)} \mathbb{I}_{c^*_{x,y}>0} L_{reg}(\mathbf{t}_{x,y}, \mathbf{t}^*_{x,y}), \end{aligned} \tag{2}$$

where $\mathbf{p}_{(x,y)}$ denotes the predicted classification probabilities at the location (x, y) of the FPN feature map, $c^*_{(x,y)}$ indicates the ground classification label associated with the regressed bounding box at location (x, y), while $\mathbf{t}_{x,y}$ and $\mathbf{t}^*_{x,y}$ denotes the predicted and ground truth offsets of the bounding box. L_{cls} is the focal loss [52] and L_{reg} is the IoU loss [53]. However, training the FCOS detector using Eq. 2 alone can overfit to the training data. In order to minimise the effect of overfitting, we are going to adopt our previously proposed DG framework [20] to the FCOS detector and improve its generalisation ability to unseen target dataset. In the next subsection, we briefly describe the details of our DG framework followed by the details of proposed DGFCOS detector.

B. Domain Generalisation

We assume $\mathbf{I} \in \mathbb{R}^{N_b \times 3 \times H \times W}$ be the batch of input images fed to $F^{(\theta)}$. 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. 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. By Bayes' theorem, 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 $Q(F^{(\theta)})$. The standard technique followed by many approaches in UDA [10], [54] to transform all the domain specific marginals onto a common $Q(F^{(\theta)})$ is by introducing a domain discriminator $S^{(\psi)}$ which is trained by minimising the negative domain discriminator loss [55],

However, as pointed out by recent studies [21], [44], [56], [57], the stability of conditionals across domains cannot be guaranteed with $S^{(\psi)}$. Any method with the goal of achieving domain invariance needs to compensate for the variation in conditionals $P_D(C^I, B^I | F^{\theta}(I))$. By using Bayes' theorem,

$$P_D(C^I, B^I | F^{\theta}(I)) = P_D(C^I | B^I, F^{\theta}(I)) P_D(B^I | F^{\theta}(I)) \tag{3}$$

where $P_D(C^I | B^I, F^{\theta}(I))$ indicates the instance level domain specific classifier and $P_D(B^I | F^{\theta}(I))$ corresponds to domain specific bounding box regressor.

In order to achieve the stability of conditional distributions $P_D(C^I, B^I | F^{\theta}(I))$, across the source domains, we adopt the strategy used in our previously proposed DG framework. In [20], we have proved that the stability of domain specific class conditionals can be achieved by training the main detector in conjunction with N domain specific classifiers $\{\phi'_D\}_{D=1}^N$ while the domain agnostic bounding box regression can be realised through a consistency regularisation term between the instance and image level domain discriminators.

The final loss function used to train the system is thus:

$$\begin{aligned} &\min_{(\theta,\phi,\beta,\{\phi^\dagger_D\})} \max_{(\psi_{img},\{\phi'_D\},\psi_{ins})} \\ &L(\theta, \beta, \phi, \psi_{img}, \psi_{ins}, \{\phi'_D\}, \{\phi^\dagger_D\}) \\ &= L_{cls}(\theta, \phi) + L_{reg}(\theta, \beta) + \alpha_1 L_{dadv}(\theta, \psi_{img}) + \\ &\alpha_2 L_{dins}(\theta, \psi_{ins}) + \alpha_3 L_{cst}(\theta, \beta) + \alpha_4 L_{erc}(\theta, \{\phi'_D\}) \\ &+ \alpha_5 L_{cel}(\theta, \{\phi^\dagger_D\}), \end{aligned} \tag{4}$$

where L_{dadv} and L_{dins} is the loss used to train an image-level ($S^{(\psi_{img})}$) and instance-level ($S^{(\psi_{ins})}$) domain discriminators, respectively, while L_{cst} is the loss which enforces consistent domain label predictions by both image and instance level domain discriminators. L_{erc} is the sum of negative cross entropy losses corresponding to each of the N additional domain-specific classifiers $\{T'_D\}$ while L_{cel} is the sum of cross entropy losses used for training each of the N domain specific classifiers $\{T^\dagger_D\}$. The losses (L_{dadv}, L_{dins}, L_{cst}) are used to achieve the bounding box alignment while (L_{erc}, L_{cel}) assist in achieving the class-conditional alignment. In order to use our DG framework [20] for FCOS, it remains crucial to appropriately tap the features at both image and instance levels. The next section describes details of how the preliminaries described in this section can be adapted to a FCOS detector.

IV. Proposed Approach

The overview of the new proposed DGFCOS architecture is given in Fig. 2, where the FCOS is trained in conjunction with two additional modules related to class-conditional invariance and bounding box invariance. Similar to [20], these additional modules aim to optimise the feature extractor so that the input images map onto a feature space where the detection is consistent across multiple domains. The main detection loss in Eq. 2 in conjunction with the additional regularisation terms defined in Eq. 4 is used to train the proposed detector. Similar

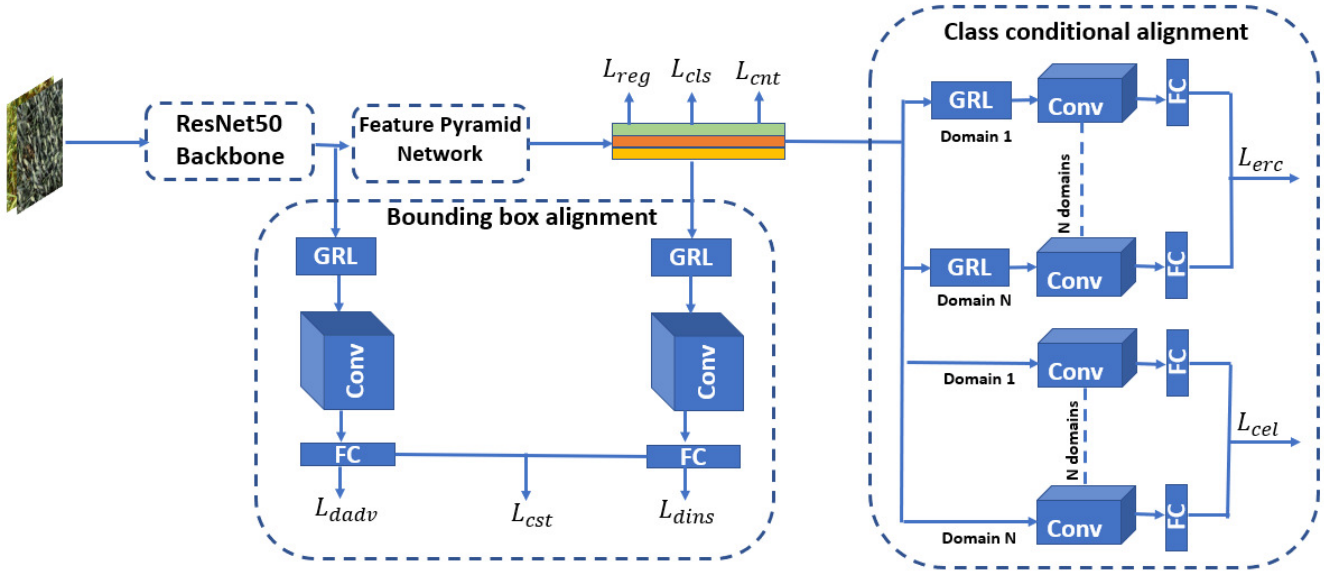


Fig. 2: Proposed New Architecture for DGFCOS. Ouptut of ResNet50 backbone will be used as image level features. Unlike the previous DGFRCNN [20], we now tap instance level features from the output of FPN.

to our previous approach [20] we tap the image level features as the output of the final layer of backbone network ($\mathbf{F}_5$) and feed as an input to the image-level domain discriminator $S^{(\psi_{img})}$. One of the biggest challenge associated with the multi-source domain datasets is the heavy imbalance in the number of images in each of the source domains can induce the domain specific bias from the dominant domains which inturn can lead to poor generalisation ability of FCOS. The negative cross entropy loss used in our previous work to train $S^{(\psi_{img})}$ does not address this domain imbalance issue. However, motivated from the focal loss in [52], which was originally proposed to address class imbalance issue, we train $S^{(\psi_{img})}$ by minimising the following negative focal loss,

$$\begin{aligned} L_{dadv} &= -\frac{1}{N_b}\sum_{i=1}^{N_b} \text{FL}(\mathbf{d}_t^i, \mathbf{d}_p^i) \\ &= -\frac{1}{N_b}\sum_{i=1}^{N_b} ((1-\mathbf{d})^{\gamma})\text{BCE}(\mathbf{d}_t^i, \mathbf{d}_p^i) \end{aligned} \tag{5}$$

where $\mathbf{d}_t^i$ and $\mathbf{d}_p^i$ are the one-hot ground-truth labels and predicted domain probability scores, respectively. FL and BCE denote the focal loss and binary cross entropy. The value for $\mathbf{d}$ can be computed by using the following expression:

$$\mathbf{d} = \mathbf{d}_t^i \mathbf{d}_p^i + (1-\mathbf{d}_t^i)(1-\mathbf{d}_p^i) \tag{6}$$

The output of every layer of feature pyramid network is utilised to model the instance level features. Let $\mathbf{P}_j \in \mathbb{R}^{N_b \times 256 \times W_j \times H_j}$ be the dimension of the features maps extracted from j^{th} layer of FPN, W_j and H_j indicate the width and height of the feature maps. In FCOS [22], an estimate of the object presence is evaluated at every location (x, y) of the feature map $\mathbf{P}_j$. Here the 256D feature vector at every (x, y) is used as an instance level feature. We stack all of the instance level features extracted from multiple levels of FPN as $\mathbf{P}_{ins} \in \mathbb{R}^{N_b \times \sum_{i=3}^{7}(W_i H_i) \times 256}$ which is further fed as an input to instance level domain discriminator and trained using a similar negative focal loss as defined in Eq. 5. Also, these instance level features are used by the domain specific classifiers to achieve the class conditional alignment.

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.

Algorithm 1: Training strategy for DGFCOS

Input: $\{X_D\}_{D=1}^{N}$, 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 ;
    Update θ, β, ϕ in Eq. (4);
    Using fixed θ, update ψ_{img} and ψ_{ins} in Eq. (4);
    Sample random batch of images ;
    Update θ, β, ϕ in Eq. (4);
    Using fixed θ, update $\{\phi_D^{\dagger}\}$ in Eq. (4);
    Sample random batch of images ;
    Update θ, β, ϕ in Eq. (4);
    Update θ, $\{\phi'_D\}$ in Eq. (4);
    Sample random batch of images ;
    Update θ, β, ϕ in Eq. (4);
    Using fixed $\{\phi_D^{\dagger}\}$, update θ in Eq. (4)
  end
end

TABLE I: Quantitative Analysis on Cityscapes (C), Foggy Cityscapes (F), BDD100K (B). The best AP and mAP in each setting is highlighted in bold. Single-best indicates the best performing FCOS model trained on each of the source domains. Source-combined is when FCOS is trained using the dataset obtained by merging all the source domains. Oracle Train on Target indicates the performance evaluation when the detector is trained using the target domain annotations.

DG Setting	Methods	person	rider	car	truck	bus	train	motor	bike	mAP
F & B to C	Single-best	52.1	**47.9**	71.0	26.3	40.7	**18.2**	**27.4**	37.3	40.1
	Source-combined	60.0	47.2	80.2	**44.0**	62.9	6.0	25.2	41.5	45.9
	DGFCOS (Ours)	**60.7**	47.5	**81.2**	43.2	**63.3**	6.2	26.8	**44.6**	**46.7**
	Oracle-Train on Target	54	52.9	70.5	34.0	59.6	29.6	32.6	46.4	47.5
C & B to F	Single-best	49.4	**44.1**	74.1	16.0	32.0	**9.1**	21.5	**43.8**	36.3
	Source-combined	51.1	41.4	73.9	27.8	41.1	6.9	21.9	38.5	37.8
	DGFCOS (Ours)	**52.2**	41.3	**74.2**	**28.8**	**42.0**	7.1	**22.6**	39.1	**38.4**
	Oracle-Train on Target	56.2	48.4	73.3	33.5	48.5	37.9	30.1	46.4	46.8
F & C to B	Single-best	**38.4**	15.8	**58.3**	14.6	**22.5**	-	13.4	23.2	26.6
	Source-combined	35.9	23.9	54.5	20.7	19.5	-	16.9	26.9	28.3
	DGFCOS (Ours)	36.5	**25.1**	56.0	**21.0**	20.0	-	**17.8**	**27.6**	**29.1**
	Oracle-Train on Target	61.0	42.4	78.9	59.0	57.7	-	37.8	42.8	54.2

V. Results

Datasets. We demonstrate the generalisation ability of our approach on the following four popular multi-source object detection datasets related to precision agriculture and autonomous driving.

GWHD [26]: 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.

Cityscapes (C), *Foggy Cityscapes (FC)*: Cityscapes [24] 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* [58] images is synthetically created using a standard fog image formation model [59] with an airlight coefficient of 0.02.

BDD100k (B) [25]: 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 [58], [60], 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) [23]: This data is generated by capturing the snapshots of *Grand Threft Auto V* (GTA-V) video game. There are 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 DGFCOS through the following experiments:

(i) *DG (multi-class, single target domain):* We evaluate the generalisation ability of DGFCOS when the nature of domain shift is due to acquisition/simulation setup used to capture/generate the data.

(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 FCOS 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 DGFCOS 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 unlabelled target images are available during training.

Training details. From empirical observations, we set the regularisation constants 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). We trained and tested our model using PyTorch and Torchvision's FCOS library on a NVIDIA RTX 3090 GPU with 24GB of GPU memory. Following [26], 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).

A. Quantitative Analysis

In experiment (i), we present the performance of proposed approach on *Cityscapes*, *Foggy Cityscapes* and *BDD100K* datasets (Table I) where two of these three are used as source

TABLE II: Quantitative analysis for proposed approach in DG setting for GWHD dataset. 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. BBA: Bounding Box Alignment. CCA: Class Conditional Alignment. The best results are highlighted in bold.

	GWHD	(S,C) $\rightarrow$ B	(S, B) $\rightarrow$ C	L_{cls}	L_{reg}	L_{dadv}	L_{dins}	L_{cst}	L_{erc}	L_{cel}
FCOS	52.11	50.41	71.5	+	+	-	-	-	-	-
BBA	54.62	52.5	71.9	+	+	+	+	+	-	-
CCA	53.97	51.3	71.7	+	+	+	-	-	+	+
DGFCOS (ours)	**54.90**	**53.51**	**72.5**	+	+	+	+	+	+	+

TABLE III: 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.

	Person	Rider	Car	Truck	Bus	Train	MCycle	Bicycle	mAP
Source-only	26.9	38.2	35.6	18.3	32.4	9.6	25.8	28.6	26.9
EPM [17]	39.9	38.1	57.3	28.7	**50.7**	37.2	30.2	34.2	39.5
SIGMA [16]	**44.0**	**43.9**	**60.3**	**31.6**	50.4	**51.5**	**31.7**	**40.6**	**44.2**
DGFCOS (ours)	35.8	39.7	52.4	30.6	39.6	23.1	25.1	32.5	34.9
Oracle Train on target	56.2	48.4	73.3	33.5	48.5	37.9	30.1	46.4	46.8

domains and the other as target domain. These three datasets share eight common category of objects. It can be seen that the proposed architecture performs the best in majority of the object categories in all of the three settings. This signifies the need for compensating both the covariate and concept shifts in a multi-source domain generalisation scenario. Due to less number of instances, the 'train' category was not considered during the evaluation for third setting (F & C to B).

B. Ablation studies

Table II (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. The combination when *Sim10K* as target domain is not included as it is not realistic to have generalisation from real to synthetic datasets. It can be seen that our approach outperforms the baseline Faster R-CNN used in WildS benchmark [28]. 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 III (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 [16], [17] approaches which use single stage anchor free detectors. Our DGFCOS 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.

TABLE IV: Comparison of inference times for Faster R-CNN vs FCOS in frames per second (fps).

Dataset (resolution)	Faster R-CNN	FCOS
GWHD (1024 $\times$ 1024)	5	8
Cityscapes (600 $\times$ 1200)	6	10
BDD100K (600 $\times$ 1200)	6	10

Finally, we present comparative analysis in terms of frames per second between Faster R-CNN and FCOS detectors. Here we train the detectors on the official train splits of GWHD, Cityscapes and BDD100K datasets while inference is done on their corresponding official validation splits. From Table IV, it can be seen that FCOS always have a better inference speed over Faster R-CNN. This is primarily due to the higher computational complexity associated with Faster R-CNN and this makes FCOS a preferred choice for real time object detection.

VI. CONCLUSIONS

In this paper, for the first time, we have proposed a Domain Generalised FCOS architecture. Here, we have adopted our previously proposed DG framework for real-time FCOS and proposed a 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 FCOS, on four standard object detection datasets from autonomous driving and agricultural robotics. We showed that FCOS has better inference speed over Faster R-CNN which makes FCOS a preferred choice for real time robotics.

ACKNOWLEDGMENT

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

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

[2] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 781–10 790.

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

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

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

[6] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29. Curran Associates, Inc., 2016.

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

[8] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do imagenet classifiers generalize to imagenet?" in *International Conference on Machine Learning*. PMLR, 2019, pp. 5389–5400.

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

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

[11] M. Cai, M. Luo, X. Zhong, and H. Chen, "Uncertainty-aware model adaptation for unsupervised cross-domain object detection," *arXiv preprint arXiv:2108.12612*, 2021.

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

[13] C.-D. Xu, X.-R. Zhao, X. Jin, and X.-S. Wei, "Exploring categorical regularization for domain adaptive object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 724–11 733.

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

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

[16] W. Li, X. Liu, and Y. Yuan, "Sigma: Semantic-complete graph matching for domain adaptive object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5291–5300.

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

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

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

[20] K. Seemakurthy, C. Fox, E. Aptoula, and P. Bosilj, "Domain generalisation for object detection," *arXiv preprint arXiv:2203.05294*, 2022.

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

[22] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9627–9636.

[23] M. Johnson-Roberson, C. Barto, R. Mehta, S. N. Sridhar, K. Rosaen, and R. Vasudevan, "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)*. IEEE, 2017, pp. 746–753.

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

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

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

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

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

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

[30] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," in *European Conference on Computer Vision*. Springer, 2010, pp. 213–226.

[31] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *CVPR 2011*. IEEE, 2011, pp. 1521–1528.

[32] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1. Ieee, 2005, pp. 886–893.

[33] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2009.

[34] P. Viola and M. Jones, "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*, vol. 1. Ieee, 2001, pp. I–I.

[35] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[36] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

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

[38] X. Zhang, H. Lu, C. Hao, J. Li, B. Cheng, Y. Li, K. Rupnow, J. Xiong, T. Huang, H. Shi, *et al.*, "Skynet: a hardware-efficient method for object detection and tracking on embedded systems," *Proceedings of Machine Learning and Systems*, vol. 2, pp. 216–229, 2020.

[39] M. Liu, W. Jia, Z. Wang, Y. Niu, X. Yang, and C. Ruan, "An accurate detection and segmentation model of obscured green fruits," *Computers and Electronics in Agriculture*, vol. 197, p. 106984, 2022.

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

[41] H. Liu, P. Song, and R. Ding, "Towards domain generalization in underwater object detection," in *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 1971–1975.
[42] ——, "WQT and DG-YOLO: Towards domain generalization in underwater object detection," *arXiv preprint arXiv:2004.06333*, 2020.
[43] C. Lin, Z. Yuan, S. Zhao, P. Sun, C. Wang, and J. Cai, "Domain-invariant disentangled network for generalizable object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8771–8780.
[44] Y. Li, X. Tian, M. Gong, Y. Liu, T. Liu, K. Zhang, and D. Tao, "Deep domain generalization via conditional invariant adversarial networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 624–639.
[45] S. Hu, K. Zhang, Z. Chen, and L. Chan, "Domain generalization via multidomain discriminant analysis," in *Uncertainty in Artificial Intelligence*. PMLR, 2020, pp. 292–302.
[46] H. Li, S. J. Pan, S. Wang, and A. C. Kot, "Domain generalization with adversarial feature learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5400–5409.
[47] B. Tian, D. Zhang, and C. Zhang, "High-speed tiny tennis ball detection based on deep convolutional neural networks," in *2020 IEEE 14th International Conference on Anti-counterfeiting, Security, and Identification (ASID)*. IEEE, 2020, pp. 30–33.
[48] X. Dai, Y. Lei, T. Wang, Z. Tian, J. Zhou, M. McDonald, S. Y. David, B. B. Ghavidel, J. D. Bradley, T. Liu, *et al.*, "Automated ct segmentation for rapid assessment of anatomical variations in head-and-neck radiation therapy," in *Medical Imaging 2022: Image-Guided Procedures, Robotic Interventions, and Modeling*, vol. 12034. SPIE, 2022, pp. 306–311.
[49] M. Zhu, G. Hu, H. Zhou, S. Wang, Z. Feng, and S. Yue, "A ship detection method via redesigned fcos in large-scale sar images," *Remote Sensing*, vol. 14, no. 5, p. 1153, 2022.
[50] M. Zhu, G. Hu, S. Li, H. Zhou, S. Wang, and Z. Feng, "A novel anchor-free method based on fcos+ atss for ship detection in sar images," *Remote Sensing*, vol. 14, no. 9, p. 2034, 2022.
[51] Z. Liu, K. Ding, Q. Xu, Y. Song, X. Yuan, and Y. Li, "Scene images and text information-based object location of robot grasping," *IET Cyber-Systems and Robotics*, 2022.
[52] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.
[53] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang, "Unitbox: An advanced object detection network," in *Proceedings of the 24th ACM international conference on Multimedia*, 2016, pp. 516–520.
[54] T. Matsuura and T. Harada, "Domain generalization using a mixture of multiple latent domains," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 11 749–11 756.
[55] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *arXiv preprint arXiv:1406.2661*, 2014.
[56] B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij, "On causal and anticausal learning," *arXiv preprint arXiv:1206.6471*, 2012.
[57] D. Janzing and B. Schölkopf, "Causal inference using the algorithmic markov condition," *IEEE Transactions on Information Theory*, vol. 56, no. 10, pp. 5168–5194, 2010.
[58] C. Sakaridis, D. Dai, and L. Van Gool, "Semantic foggy scene understanding with synthetic data," *International Journal of Computer Vision*, vol. 126, no. 9, pp. 973–992, Sep 2018.
[59] W. E. K. Middleton, "Vision through the atmosphere," in *Geophysik II/Geophysics II*. Springer, 1957, pp. 254–287.
[60] X. Hu, C.-W. Fu, L. Zhu, and P.-A. Heng, "Depth-attentional features for single-image rain removal," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.