InterAug: A Tuning-Free Augmentation Policy for Data-Efficient and Robust Object Detection

Anonymous ICCV submission

Paper ID 25

Abstract

The recent progress in developing pre-trained models, trained on large-scale datasets, has highlighted the need for robust protocols to effectively adapt them to domain-specific data, especially when there is a limited amount of available data. Data augmentations can play a critical role in enabling data-efficient fine-tuning of pre-trained object detection models. Choosing the right augmentation policy for a given dataset is challenging and relies on knowledge about task-relevant invariances. In this work, we focus on an understudied aspect of this problem – can bounding box annotations be used to design more effective augmentation policies?. Through `InterAug`*, we make a critical finding that, we can leverage the annotations to infer the effective context for each object in a scene, as opposed to manipulating the entire scene or only within the pre-specified bounding boxes. Using a rigorous empirical study with multiple benchmarks and architectures, we demonstrate the efficacy of* `InterAug` *in improving robustness, handling data scarcity and being resilient to high background context diversity. Finally,* `InterAug` *can be used with any off-the-shelf policy, does not require any modification to the model architecture, and significantly outperforms existing protocols.*

1. Introduction

Augmentation design has emerged as a crucial approach to enable robust and data-efficient training of deep models in a variety of computer vision tasks. While a large class of image manipulation strategies can be utilized for synthesizing augmentations [19], *e.g.*, horizontal/vertical flips, changes in brightness or mixup [25], the key focus has been on designing effective augmentation policies. Examples policies include Cutmix [20] that adds a randomly cropped portion of one image onto another, Augmix [11] that utilizes a composition of multiple pre-specified augmentations and more recently, TrivialAug [17] that randomly selects both the type and severity from pre-specified sets of augmentations and severity levels. Despite their widespread adoption, AutoAugment [26, 5] techniques that automatically learn dataset-specific augmentation policies are known to produce superior performance. However, their computational complexity, reliance on large datasets and lack of transferability (from one dataset to another) make them a less preferred choice in practical, data-constrained applications.

In this paper, we explore the problem of designing dataset-agnostic augmentation policies for data-efficient training of object detectors. A common aspect in all existing off-the-shelf policies is that they do not exploit the bounding box (bbox) annotations typically available in object detection datasets. In general, bbox annotations are different from pixel-level labels used in classical instance segmentation tasks, in that they do not accurately represent the object boundaries and often contain some amount of background pixels. Consequently, by enabling invariance to the local context captured by bbox annotations, one must be able to enrich the object detectors and even potentially improve their robustness under real-world distribution shifts. A straightforward approach towards that is to naïvely extend any augmentation policy (e.g., TrivialAug) by manipulating the regions only within the bounding boxes. However, we find that this approach leads to consistently poorer performance when compared to a standard implementation of that policy. This observation can be (at least partly) attributed to the inconsistent nature of bbox labels, *i.e.*, the amount of context captured for each bbox can vary based on factors such as the proximity between objects, the number of objects present, and most importantly the annotator's judgement. As a result, restricting augmentations only within the bounding boxes can lead to inconsistent decision rules even for the same object.

In order to circumvent this, we present `InterAug`, a simple modification applicable to any pre-existing augmentation policy. This involves expanding the bounding box of each object to determine its "effective context" (EC), and subsequently applying the chosen image manipulation within the estimated context. Subsequently,t. Through the

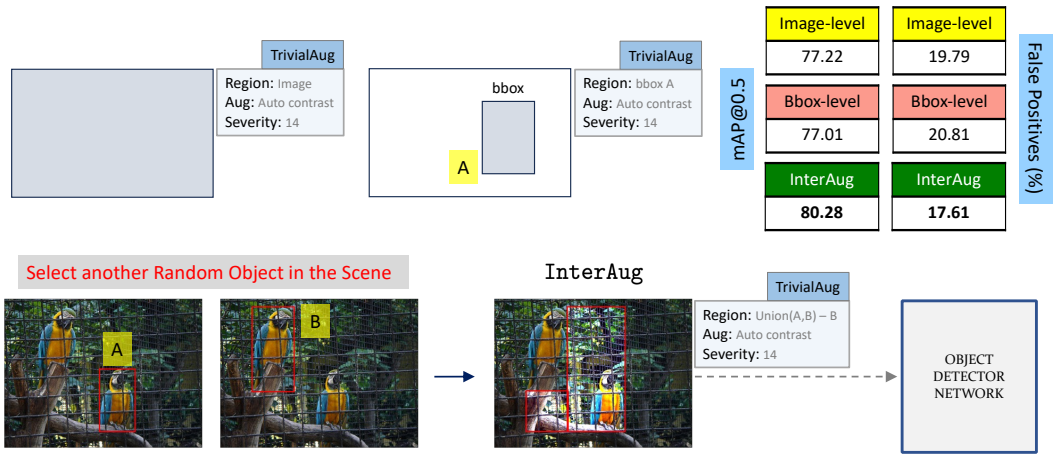

Figure 1: **Proposed Work.** Naïvely extending existing augmentation policies (e.g., TrivialAug) to incorporate bounding box information leads to poorer detection performance (results showed for Pascal VOC, when only 10% data is used for training). Hence, we introduce `InterAug` , which infers the *effective context* to expand the given bbox annotation and restricts image manipulation only within this context. `InterAug` is applicable to any architecture, augmentation policy and leads to improved and more robust object detectors.

consistent use of expanded local context and the systematic elimination of undesirable leakage from other objects, this simple approach enables targeted image manipulation while being cognizant of other co-occurring objects within the scene.

Findings. In our study, we rigorously evaluated the performance of `InterAug` using a suite of commonly adopted benchmarks and model architectures (F-RCNN, RetinaNet, DETR). Motivated by its simplicity and efficacy, we used TrivialAug, a state-of-the-art tuning-free augmentation policy, to implement all our variants (image-level, bbox-level, and `InterAug`). We make the following findings:

- **Robustness under real-world shifts (Section 4.1).** Following the recent DetectBench [22], we considered three sets of splits from the Berkeley Deep Drive dataset, namely weather, scene and time, in order to evaluate the impact of augmentation policies on detector robustness. Across all architectures, we observe consistent gains ($\approx 7.8\%$ average in mAP@0.5) over the bbox-level policy as well as the *de facto* standard of image-level augmentations ($\approx 3.9\%$ average). Further, studying metrics from the recent TIDE framework [2], a toolbox for fine-grained error analysis reveals the importance of considering the effective semantic context;

- **Performance in data-constrained settings (Section 4.2).** Our experiments with the standard Pascal VOC benchmark reveal that, at low training sizes ($10\% - 20\%$), there is no apparent performance gap between bbox- and image-level augmentation policies. Interestingly, via selective context manipulation, `InterAug` provides particularly impressive gains (2.6% in F-RCNN and 3.1% in RetinaNet) in such data-constrained settings;

- **Impact of high context diversity (Section 4.3).** Since `InterAug` relies on exploiting the local context, we evaluated its behavior on the synthetic fruits benchmark, which synthetically places common fruits in unrelated scenes. Surprisingly, even in this challenging case, `InterAug` outperforms the naïve bbox-level policy by large margins ($3.5\% - 4\%$).

Overall, `InterAug` provides an efficient augmentation policy for object detector training, that is effective with any dataset, model architecture or training sample size.

2. Proposed Approach

In conventional object recognition models, only object labels are available and hence image-level manipulations are appropriate for implementing augmentation policies. However, when detecting multiple objects in a scene, the augmentations must be designed to promote invariance to changes in the local context, and bounding box annotations can be useful. To test this hypothesis, we first naïvely extend TrivialAug [17] by restricting the (randomly) chosen image manipulation only within the bounding boxes. We refer to this as bbox-level augmentation policy, as opposed to the conventional image-level policy. We find that, in practical data-constrained settings, a bbox-level policy underperforms (measured using mAP@0.5 and False Positives (%) in Figure 1) in comparison to the image-level policy. This

Algorithm 1: `InterAug` with TrivialAug

Input: Image I, bounding boxes $\{\mathrm{B}_1, \mathrm{B}_2, \cdots, \mathrm{B}_n\}$, List of augmentations $\mathcal{A}$ and strengths $\mathcal{M}$
Output: Augmented Image
1. For any object O_i with bounding box annotation B_i, randomly select another bounding box annotation B_j
2. Construct effective context $\mathrm{S}_{(i,j)}$ as described in Section 2.1
3. Sample $\mathrm{aug} \in \mathcal{A}$ and strength $m \in \mathcal{M}$
4. Perform augmentation $\mathrm{aug}\Big(\mathrm{S}_{(i,j)}, m\Big)$

somewhat surprising result motivated us to take a deeper look into the design of an effective augmentation policy with bbox annotations.

We begin by hypothesizing that the inconsistent nature of bounding box labels can be one of the reasons for this behavior. Unlike pixel-level object labels, the context captured in bbox annotations can vary due to factors like object proximity, the number of objects present, and, most importantly, the annotator's judgment. Consequently, by confining augmentations solely within the bounding boxes, inconsistent decision rules may arise even for the same object in different scenes. To address this challenge, we introduce a simple protocol `InterAug` that can be implemented using any off-the-shelf augmentation policy. As illustrated in Figure 1, with no additional modification to the training pipeline, `InterAug` leads to significantly improved detectors ($> 3\%$ gain in mAP@0.5). We next describe `InterAug` and its implementation details.

Setup. We denote a scene as $\mathrm{I} \in \mathbb{R}^{H \times W \times C}$, where H, W, C represent the height, width and number of channels of the image. Without loss of generality, we assume that the image contains n objects $\{\mathrm{O}_1, \mathrm{O}_2, \cdots, \mathrm{O}_n\}$ with corresponding bounding boxes $\{\mathrm{B}_1, \mathrm{B}_2, \cdots, \mathrm{B}_n\}$. Each B_j is expressed using the top-left and bottom-right spatial coordinates $\mathrm{B}_j = \left\{(\mathrm{x}_j^1, \mathrm{y}_j^1), (\mathrm{x}_j^2, \mathrm{y}_j^2)\right\}$. Finally, we denote the object detector as P_Θ parameterized by Θ.

2.1. `InterAug`: Augmentation Policy Design

Our approach's fundamental idea revolves around achieving invariance to variations in an object's local context and addressing the inconsistency in bbox labels. To accomplish this, we emphasize the significance of considering the semantic context (background) while ensuring that information from co-occurring objects in a scene does not influence the process, thereby avoiding any unintended leakage. For a given object O_i with bounding box B_i, we first select another object O_j (with B_j) to infer the effective context (EC). Note that, the choice of O_j is random in every iteration and hence the inferred EC for an object O_i can vary between iterations.

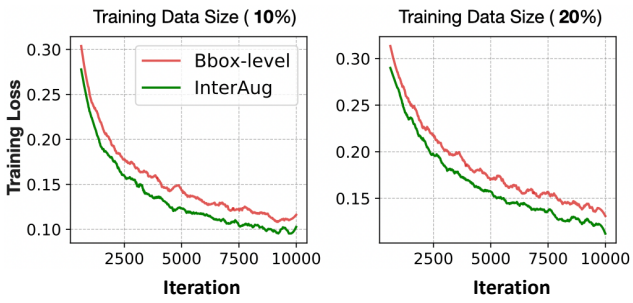

Figure 2: **Convergence.** An illustration of the training convergence observed with naïve bbox-level policy and `InterAug`. Here, we consider two different training settings for Pascal VOC, wherein the training size was fixed at 10% and 20% of the full train data. Interestingly, `InterAug` demonstrates improved convergence characteristics. As we will show in the results, this also reflects in the superior generalization and robustness performance.

More specifically, we first construct the union box $\mathrm{B}^u_{(i,j)}$ as follows:

$$\mathrm{B}^u_{(i,j)} = \Big\{\big(\min(\mathrm{x}_i^1, \mathrm{x}_j^1), \min(\mathrm{y}_i^1, \mathrm{y}_j^1)\big), \big(\max(\mathrm{x}_i^2, \mathrm{x}_j^2), \max(\mathrm{y}_i^2, \mathrm{y}_j^2)\big)\Big\}$$

Now, to identify the effective context for O_i, we compute residual between the union box and the bounding box B_j *i.e.*, $\mathrm{S}_{(i,j)} = \mathrm{B}^u_{(i,j)} - \mathrm{B}_j$. Since the EC's for the same object can focus on different aspects of the background in a scene, we encourage the detectors to avoid shortcut decision rules.

Implementation. Algorithm 1 summarizes the proposed augmentation policy. We begin by noting that, image-level and bounding box-level (or shortly bbox-level) policies are special cases of our approach, wherein the former considers the entire image to be the effective context and the latter uses only the bounding-box annotations. The effective context $\mathrm{S}_{(i,j)}$ identified by `InterAug` will be piece-wise rectangular and hence we first split it into its constituent rectangular regions and then apply the pre-specified augmentation within each of those regions. Please refer to Figure 1 for an illustration. While `InterAug` can be implemented with any off-the-shelf policy, we opt for TrivialAug [17], a tuning-free augmentation policy, that involves randomly selecting from a pre-specified set of image transformations $\mathcal{A}$ and list of augmentation strengths $\mathcal{M}$. In all our experiments, we fixed A ={vertical/ horizontal flips, crop, solarize, emboss, enhance color, sharpness, contrast, posterize, blur, add noise, add clouds}, and we randomly pick the corresponding intensity ranges specified in M ={[0.5,1.0],[1.0,1.5], [0.2,1.0], [0.5,2.0], [0.5,3.0], [0.5, 2.0], [0.5,1.5],[1,4],[0,15],[1,2], [0.5,1]}. To improve

Evaluation	Models	Datasets	Section
Robustness under real-world shifts	Faster-RCNN, RetinaNet,DETR	BDD-Weather, BDD-Scene, BDD-Time	sec 4.1
Performance of `InterAug` in data-constrained settings	Faster-RCNN, RetinaNet	Pascal VOC	sec 4.2
Impact of high context diversity	Ftster-RCNN, RetinaNet	Synthetic Fruits	sec 4.3

Table 1: List of experiments considered in our empirical study.

the training process, `InterAug` also considers the EC to be the entire union region $\mathrm{B}^{u}_{(i,j)}$ or the residual region $\mathrm{B}^{u}_{(i,j)} - \mathrm{B}_i - \mathrm{B}_j$ where both bounding boxes are subtracted from the union. More specifically, our implementation uses all the three ways of modeling the effective context (one of them randomly chosen in every minibatch during training) and perform synthetic augmentations within this context.

Convergence Analysis. In Figure 2, we present an illustration of the training convergence observed using the naïve bbox-level policy and our proposed method. For this result, we conducted experiments with two distinct training settings on the Pascal VOC benchmark, where the training size was set to 10% and 20% of the full train data. Interestingly, `InterAug` exhibits a consistently better convergence compared to the naïve augmentation policy. As we will demonstrate in the results (Section 3), this improvement translates into superior generalization and robustness performance.

3. Experiments

Setup. We conduct a number of experiments to assess the performance of `InterAug` in different scenarios, including real-world distribution shifts, data-constrained settings, and its behavior on scenes that exhibit large context diversities. These evaluations are carried out using widely recognized object detection benchmarks, namely Berkeley Deep Drive (BDD), Pascal VOC, and the challenging synthetic fruits datasets. The details of these experiments, including the model architectures employed and the datasets utilized, can be found in Table1. We will now provide a description of the dataset setup for each of these experiments.

(i) To evaluate the robustness of `InterAug` against real-world distribution shifts, we utilize DetectBench [22], a recently introduced benchmark specifically designed to assess the out-of-distribution (OOD) robustness of object detectors. DetectBench constructs three distinct BDD-OOD benchmarks: BDD-Weather, BDD-Scene, and BDD-Time, by leveraging the attribute annotations available in the large-scale autonomous driving dataset, Berkeley Deep Drive (BDD). For instance, the BDD-Weather benchmark aims to assess the OOD performance of object detection models under varying weather conditions. The training set consists of $52,699$ images labeled with weather attributes corresponding to "clear" and "overcast", while the model evaluation is performed on a more challenging set of $17,888$ images containing novel weather attributes "foggy", "cloudy", "rainy" and "snowy". Similarly, the BDD-Scene and BDD-Time benchmarks have non-overlapping attributes related to "scene" and "time of day" respectively, with training and test sizes of $69,506$ and $9,943$ for BDD-Scene, and $47,791$ and $31,900$ for BDD-Time. All three benchmarks are comprised of 10 object categories.

(ii) To evaluate the performance of `InterAug` under limited training sample size settings, we utilized the standard Pascal VOC object detection benchmark of scenes comprising different combinations of 20 distinct objects. Following standard practice, we first combined Pascal VOC 2007 and Pascal VOC 2012 train-validation sets resulting in a training dataset of $16,550$ images. From this combined dataset, we randomly sub-selected 10% and 20% of data for training the detectors. Training object detectors with such limited data is known to be challenging and data augmentations are expected to help. In each case, we report the performance on the same held-out, full Pascal 2007 test set consisting of 4952 samples.

(iii) Finally, we utilized the synthetic fruits dataset [1], which contains images with fruits artificially inserted into natural scenes. Through this dataset, we investigate the impact of `InterAug` under scenarios characterized by high context diversity. Intuitively, due to the synthetic placement of fruits in unrealistic settings, the background context does not provide any useful signals for improving detection. As a result, this provides an assessment of how `InterAug` handles such high diversity in local context, given that its effective context generation invariably includes background pixels. This benchmark contains 65 different object categories, and to emulate data-scarcity settings, we use only a subset of 1000 images for training and report the performance on the held-out validation set.

Model Architecture. To systematically benchmark the impact of different augmentation strategies on fine-tuning object detectors with extremely limited data, we performed experiments with three popular object detection architectures: (i) Faster-RCNN [18], a two-stage detector based on Resnet-50 along with an FPN [13] backbone; (ii) RetinaNet [14], a single-stage detector based on Resnet50 and FPN; and (iii) DETR [3] a transformer-based object detec-

[1] https://public.roboflow.com/object-detection/synthetic-fruit

tor based on Resnet50 backbone. All there architectures were pre-trained on the MS-COCO [15] benchmark.

Experimental Implementation. We implemented `InterAug` using the imgaug library [12] [2] and incorporated it into the popular Detectron2 object detection framework [23] for Faster-RCNN and RetinaNet, and into HuggingFace [21] library for DETR. Although we present results using all three architectures for the BDD-OOD benchmarks, we report performance only for the Faster-RCNN and RetinaNet architectures due to the limited training sizes in the data-efficiency and high context diversity experiments. In all our experiments, we initialized the networks with weights from a model pre-trained on the COCO benchmark, and performed fine-tuning for 10K iterations with batch size 8 (2 NVIDIA TESLA V100 GPUs).

Baselines. For comparison, we consider the two widely-adopted augmentation policies, namely image-level and bbox-level, evaluated under the same experiment setup. In the former baseline, we randomly sample from the pre-specified augmentation and strength sets, and apply it to the whole image. In the latter approach, we only augment the region within the bounding box of an object (provided in the ground truth annotations). As described earlier, both these baselines can be viewed as special cases of our method, and the performance variation across these choices clearly evidences the need to achieve invariance to the context captured by the bounding box (bbox) annotations and exploring the optimal effective context (EC) for applying image manipulations.

Metrics. In addition to the commonly employed Average Precision score (mAP@0.5 score aggregated from 3 independent trials.), we also consider an additional suite of metrics to perform fine-grained error characterization. To this end, we follow the recent work by Bolya *et al*. [2] and study the following error components [3]: (i) classification error (Cls. Error): instances where the model correctly localized an object but incorrectly classified it; (ii) localization error (Loc. Error): instances where the model correctly identifies the class of an object, but the predicted bounding box is incorrect; (iii) CE Error: instances where the models makes incorrect predictions for both the bounding box and the class label; (iv) background error (Bck. Error): instances where the model incorrectly identifies the background or an area without an object as containing an object; (v) missed: instances where the model fails to identify an object that is present in the scene; (vi) false positives (FP); and (vii) false negatives (FN).

[2] https://github.com/aleju/imgaug
[3] https://github.com/dbolya/tide/

4. Results and Findings

4.1. `InterAug` consistently produces superior performance across different distribution shifts

In Figure 3, we present detailed performance results of the three architectures, Faster-RCNN, RetinaNet, and DETR, across various BDD-OOD benchmarks. We make a number of interesting observations. Firstly, we find that that `InterAug` provides significant improvements over the bbox-level baseline across all three distribution shifts, with average boosts of 10.1%, 6.7%, and 6.9% for Faster-RCNN, RetinaNet, and DETR, respectively. Next, across the different architectures `InterAug` produces gains on average 3.2%, 2.5%, and 6.03% compared to the image-level augmentation policy. Furthermore, `InterAug` not only produces significantly lesser false positives thus improving AP, but also achieves fewer false negatives.

These improvements can be directly attributed to the efficacy of our proposed augmentation policy which enables the detectors to leverage the effective context of the object while avoiding shortcut decision rules. Finally, we also include qualitative examples obtained using Faster-RCNN, and we notice that `InterAug` produces a better-calibrated model compared to the other two methods. This is demonstrated by the reduced amount of false positives and hallucinations (detecting objects that are not present in the scene), which was the case in image-level and bbox-level policies.

4.2. `InterAug` is effective under limited training data sizes

In Figure 4, we present detailed results of Faster-RCNN and RetinaNet models trained on 10% and 20% of the Pascal VOC training data, utilizing the three augmentation policies. As expected, we notice a monotonous increase in performance in all cases, as the amount of training data increases. Strikingly, `InterAug` provides non-trivial performance improvements compared to the two other baselines. For example, when trained using only 10% of the available data, both Faster-RCNN and RetinaNet improve upon the bbox-level baseline by 3.47% and 3.14% and produces gains of 2.6% and 3.1% over the best-performing image-level policy respectively. From the fine-grained analysis, we notice that the proposed augmentation policy shows particularly strong performance in reducing localization and background errors, the two main contributors to the false positives. In the 20% case, RetinaNet trained with `InterAug` achieves an 1.5% improvement in localization error over Image-level and 1.2% improvement in background error over bbox-level augmentation policy. Interestingly, the image-level policy is reasonably effective at reducing false positives, it tends to produce higher false negatives. In contrast, bbox-level conservatively reduces the number of false negatives at the

OOD Setting	Aug. Policy	Model: F-RCNN			Model: RetinaNet			Model: DETR		
		AP50	FP	FN	AP50	FP	FN	AP50	FP	FN
Scene	Image-level	36.36	3.83	54.58	45.69	20.25	22.71	22.82	10.22	57.13
	Bbox-level	29.51	5.63	61.52	42.47	20.6	23.86	21.71	9.73	58.68
	InterAug	**39.5**	**2.87**	**51.62**	**48.27**	**19.08**	**21.62**	**30.34**	**7.56**	**52.22**
Weather	Image-level	37.36	3.83	52.55	44.23	18.14	24.85	27.42	9.58	53.12
	Bbox-level	31.15	3.31	58.62	41.37	18.53	26.27	26.84	10.49	52.79
	InterAug	**40.73**	**3.2**	**51.6**	**47.03**	**17.27**	**22.49**	**32.19**	**8.22**	**49.71**
Time	Image-level	29.16	5.7	52.19	38.4	23.42	21.78	24.51	14.13	51.42
	Bbox-level	21.63	5.6	65.29	31.9	24.28	22.28	23.71	12.64	53.77
	InterAug	**32.16**	**2.83**	**51.9**	**40.56**	**22.85**	**19.2**	**30.32**	**10.92**	**48.88**

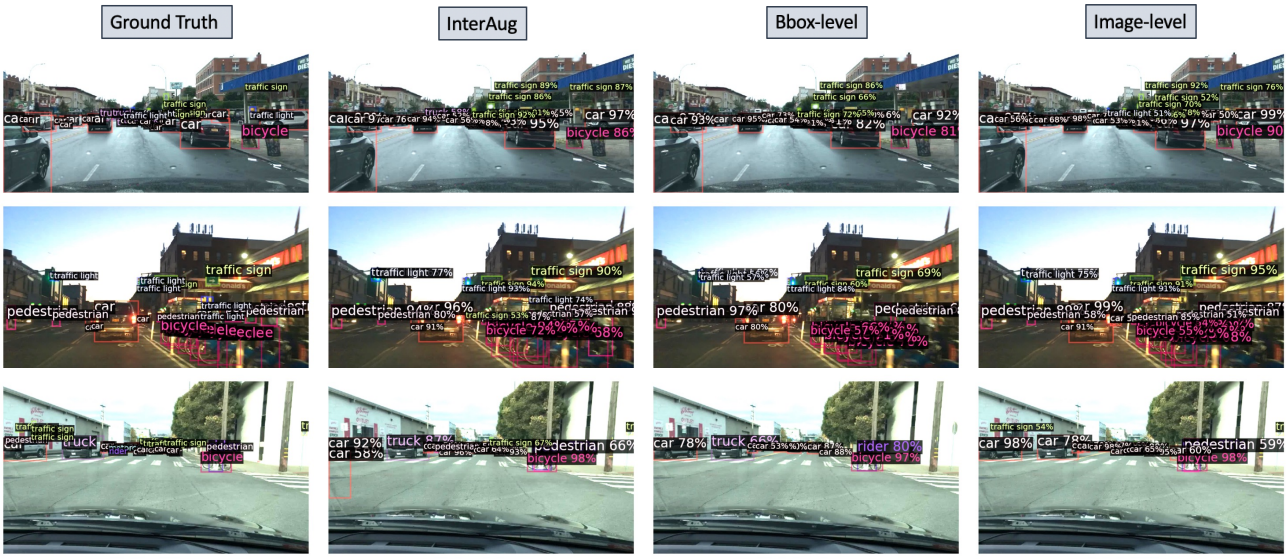

Figure 3: **Robustness.** Performance obtained by training with different augmentation policies on three real-world shifts from DetectBench [22]. We conduct experiments with three different model architectures (Faster-RCNN, RetinaNet and DETR) and report mAP@0.5 along with false positives and false negatives. We observe that `InterAug` consistently produces more robust detectors across all model architectures. Finally, we also show qualitative results obtained using Faster-RCNN.

cost of much higher false positive rates. In comparison, `InterAug` is the best performing across both error types.

4.3. `InterAug` can handle high context diversity scenes

Through Figure 5, we illustrate the performance on the Synthetic fruits dataset which is a challenging benchmark with a large number of object categories (65) and limited data (only 1000 training samples). The observations here are similar to the Pascal VOC and BDD-OOD benchmarks and `InterAug` provides non-trivial gains of 4% over bbox-level and 2.7% over image-level policies respectively for Faster-RCNN. Similar improvements can be observed even with RetinaNet, thus demonstrating the benefits of `InterAug` even under high background diversity.

5. Related Work

Data augmentation is routinely used when training deep models for computer vision [19], due to its utility in improving generalization and reducing overfitting. By leveraging synthetic data obtained via pre-defined manipulations, *e.g.* geometric transformations or corruptions [10, 23], one can build models that generalize better to unseen test data, even under distribution shifts. State-of-the-art techniques go beyond conventional image manipulations, and adopt inter-

Model	Train Size	Aug. Policy	AP50	Cls. Error	Loc. Error	CE Error	Bck. Error	Missed	FP	FN
F-RCNN	10%	Image-level	72.62	**2.34**	7.13	1.27	3.84	5.11	17.12	10.87
		Bbox-level	71.73	2.43	6.77	1.35	4.73	**4.35**	18.47	**10.17**
		InterAug	**75.2**	2.85	**6.29**	**0.92**	**2.41**	7.09	**12.55**	13.34
	20%	Image-level	75.3	**2.05**	6.5	1.13	4.11	4.13	16.23	8.91
		Bbox-level	74.14	2.09	6.53	1.22	4.65	**3.98**	17.48	**8.73**
		InterAug	**77.71**	2.32	**5.52**	**0.86**	**2.94**	5.7	**12.25**	10.81
RetinaNet	10%	Image-level	75.35	3.06	4.36	0.97	4.15	1.24	21.02	4.84
		Bbox-level	75.73	2.95	**3.84**	0.91	4.7	**0.96**	21.54	3.79
		InterAug	**78.49**	**2.49**	4.0	**0.87**	**3.63**	1.01	**18.83**	**3.69**
	20%	Image-level	77.22	2.28	4.21	1.04	4.19	1.04	19.79	3.88
		Bbox-level	77.01	2.09	4.34	0.99	4.5	**0.81**	20.8	3.02
		InterAug	**80.28**	**2.02**	**3.72**	**0.88**	**3.32**	0.98	**17.61**	**2.93**

Figure 4: **Data-efficient Training.** We report the data-constrained dtector performance obtained using two different architectures (Faster RCNN, RetinaNet) and three different augmentation policies on the Pascal VOC benchmark. In both cases, we report the average mAP@0.5 scores, when trained with 10% and 20% of the training data. Furthermore, we show the fine-grained evaluation using TIDE metrics. We find that `InterAug` achieves significant improvements over the baselines. Finally, we provide example detections for the RetinaNet model trained using different augmentation policies.

polation techniques such as Mixup [25] and CutMix [20], or compositional strategies such as AugMix [11], TrivialAug [17], AugMax, ALT [9] etc.

In practice, augmentation design typically requires dataset-specific tuning and may rely on knowledge about the task-relevant invariances. In order to simplify this process, AutoAugment strategies [6], which pose augmentation design for a given dataset as a search problem, and learn an optimal policy through reinforcement learning, have also been proposed. In practice, they can be computationally expensive and can even be impractical when the design space becomes large. Interestingly, a recent study [17] showed that, in object recognition models, an augmentation policy drawn in random can achieve similar performance as that of AutoAugment methods.

In addition to geometric or color space transformations, mixing and copy-paste style augmentations, which copy an object from one image and paste in another image, have gained popularity for object detection tasks [7, 8]. More recently, AutoAugment techniques specifically designed for object detection have emerged [26, 5]. In [5], Chen *et al*. proposed an auto augment approach to exploit the relative size of objects in a given frame and advocated for bounding box-level augmentations, which many off-the-shelf policies do not leverage. However, in our experiments, we observed that a naïve adoption of bbox-level augmentations yields consistently poor results compared to the standard image-level policy.

We hypothesize that, the inconsistent nature of bounding box labels can be one of the reasons for this behavior. Annotating a large number of examples for object detection tasks is expensive and error-prone. While multiple annotators are often required to obtain high quality annotations in many real-world applications practitioners routinely collect data from less expensive data resources, including social media/crowd-sourcing platforms, or use fewer annotators to save costs. This often results in imprecise bounding box labels. To address this challenge of noisy labels, recent works [24, 4, 16, 1] have developed sophisticated training methods that typically require large amounts of data, computationally expensive optimization strategies and multiple additional objectives. In contrast `InterAug` works out of

Model	Aug. Policy	AP50	Cls. Error	Loc. Error	CE Error	Bck. Error	Missed	FP	FN
F-RCNN	Image-level	38.25	13.51	**4.37**	1.32	2.95	12.42	**21.01**	24.9
	Bbox-level	36.96	17.95	5.41	1.24	**2.66**	12.76	23.12	25.78
	InterAug	**40.92**	**11.42**	4.6	**1.14**	2.95	**6.62**	21.5	**19.54**
RetinaNet	Image-level	37.82	18.35	**2.23**	**0.57**	2.49	8.37	32.86	17.91
	Bbox-level	35.65	17.31	2.81	0.74	3.31	6.05	33.68	16.36
	InterAug	**39.31**	**13.08**	**2.22**	0.79	**2.25**	**4.18**	**32.64**	**13.21**

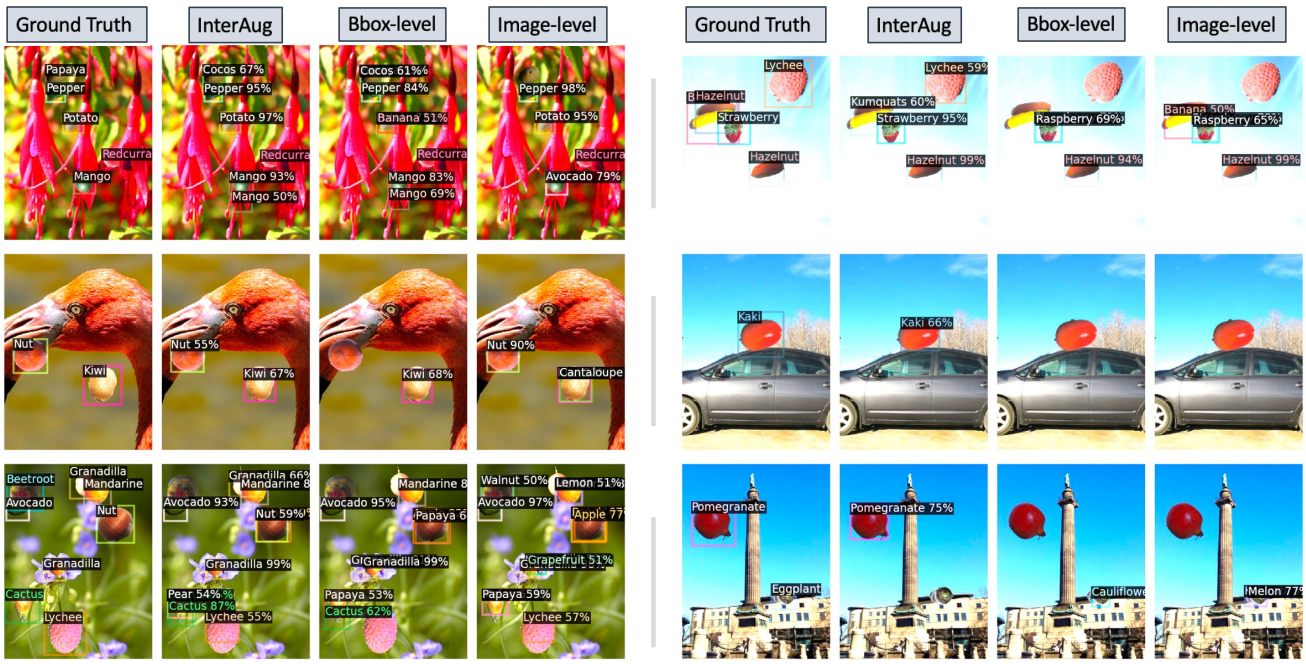

Figure 5: **Handling high context diversity.** We present object detection results of two architectures on a practically relevant and challenging task that involves limited training data and scenes characterized by high diversity in background context. We use Synthetic Fruits dataset, which contains only 1000 natural scenes with fruits from 65 categories synthetically added to them. We find that `InterAug` outperforms both baselines by significant margins. Furthermore, `InterAug` produces considerably fewer localization and background errors, as well as a reduced number of missed objects compared to the baselines. Finally, we also include qualitative visual examples.

the box, without requiring any modifications to the training loop or the model architecture, and provides significantly robust detectors and is effective even under scarce training data scenarios.

6. Conclusion

We introduced a new augmentation policy for training object detectors, referred to as `InterAug`. Importantly, `InterAug` is simple to implement and can be utilized with any off-the-shelf augmentation policy. In our study, we implemented `InterAug` with TrivialAug, originally designed for object recognition, for object detection. `InterAug` considers the effective context of an object and achieves invariance to the local context. Our experiments on three popular benchmarks demonstrated that `InterAug` consistently produces robust object detectors, outperforming current practices, and leading to improved generalization in limited training data settings and scenarios with high context diversity. On closer look, the models trained with `InterAug` reduce the number of false positives without compromising on the false negatives. In summary, our work clearly emphasizes the benefits of utilizing bounding box annotations in augmentation policies, for producing reliable and data-efficient object detectors.

References

[1] Maximilian Bernhard and Matthias Schubert. Correcting imprecise object locations for training object detectors in re-

mote sensing applications. *Remote Sensing*, 13(24):4962, 2021. 7

[2] Daniel Bolya, Sean Foley, James Hays, and Judy Hoffman. Tide: A general toolbox for identifying object detection errors. In *European Conference on Computer Vision*, pages 558–573. Springer, 2020. 2, 5

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 4

[4] Simon Chadwick and Paul Newman. Training object detectors with noisy data. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pages 1319–1325. IEEE, 2019. 7

[5] Yukang Chen, Yanwei Li, Tao Kong, Lu Qi, Ruihang Chu, Lei Li, and Jiaya Jia. Scale-aware automatic augmentation for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9563–9572, 2021. 1, 7

[6] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 113–123, 2019. 7

[7] Debidatta Dwibedi, Ishan Misra, and Martial Hebert. Cut, paste and learn: Surprisingly easy synthesis for instance detection. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 1310–1319. IEEE Computer Society, 2017. 7

[8] Hao-Shu Fang, Jianhua Sun, Runzhong Wang, Minghao Gou, Yong-Lu Li, and Cewu Lu. Instaboost: Boosting instance segmentation via probability map guided copy-pasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 682–691, 2019. 7

[9] Tejas Gokhale, Rushil Anirudh, Jayaraman J Thiagarajan, Bhavya Kailkhura, Chitta Baral, and Yezhou Yang. Improving diversity with adversarially learned transformations for domain generalization. *arXiv preprint arXiv:2206.07736*, 2022. 7

[10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 6

[11] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019. 1, 7

[12] Alexander B. Jung, Kentaro Wada, Jon Crall, Satoshi Tanaka, Jake Graving, Christoph Reinders, Sarthak Yadav, Joy Banerjee, Gábor Vecsei, Adam Kraft, Zheng Rui, Jirka Borovec, Christian Vallentin, Semen Zhydenko, Kilian Pfeiffer, Ben Cook, Ismael Fernández, François-Michel De Rainville, Chi-Hung Weng, Abner Ayala-Acevedo, Raphael Meudec, Matias Laporte, et al. imgaug. https://github.com/aleju/imgaug, 2020. Online; accessed 01-Feb-2020. 5

[13] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, 2017. 4

[14] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 4

[15] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2014. cite arxiv:1405.0312Comment: 1) updated annotation pipeline description and figures; 2) added new section describing datasets splits; 3) updated author list. 5

[16] Chengxin Liu, Kewei Wang, Hao Lu, Zhiguo Cao, and Ziming Zhang. Robust object detection with inaccurate bounding boxes. In *European Conference on Computer Vision*, pages 53–69. Springer, 2022. 7

[17] Samuel G Müller and Frank Hutter. Trivialaugment: Tuning-free yet state-of-the-art data augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 774–782, 2021. 1, 2, 3, 7

[18] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. 4

[19] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019. 1, 6

[20] Ryo Takahashi, Takashi Matsubara, and Kuniaki Uehara. Data augmentation using random image cropping and patching for deep cnns. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(9):2917–2931, 2019. 1, 7

[21] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Perric Cistac, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-Art Natural Language Processing. pages 38–45. Association for Computational Linguistics, 10 2020. 5

[22] Fan Wu, Nanyang Ye, Lanqing HONG, Chensheng Peng, Bikang Pan, Huaihai Lyu, and Heyuan Shi. Detectbench: An object detection benchmark for OOD generalization algorithms, 2023. 2, 4, 6

[23] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 5, 6

[24] Youjiang Xu, Linchao Zhu, Yi Yang, and Fei Wu. Training robust object detectors from noisy category labels and imprecise bounding boxes. *IEEE Transactions on Image Processing*, 30:5782–5792, 2021. 7

[25] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. 1, 7

[26] Barret Zoph, Ekin D Cubuk, Golnaz Ghiasi, Tsung-Yi Lin, Jonathon Shlens, and Quoc V Le. Learning data augmentation strategies for object detection. In *European conference on computer vision*, pages 566–583. Springer, 2020. 1, 7