Contrastive View Design Strategies to Enhance Robustness to Domain Shifts in Downstream Object Detection

Anonymous submission

Abstract

Contrastive learning has emerged as a competitive pretraining method for object detection. Despite this progress, there has been minimal investigation into the robustness of contrastively pretrained detectors when faced with domain shifts. To address this gap, we conduct an empirical study of contrastive learning with out-of-domain object detection, studying how the design of contrastive views affects robustness. In particular, we perform a case study of the detection-focused pretext task Instance Localization and propose strategies to augment views and enhance robustness in appearance-shifted and context-shifted detection scenarios. Amongst these strategies, we propose changes to cropping such as altering the percentage used, adding IoU constraints, and integrating saliency-based object priors. We also explore the addition of shortcut-reducing augmentations such as Poisson blending, texture flattening, and elastic deformation. We benchmark these strategies on abstract, weather, and context-based domain shifts and illustrate ways to combine them for enhanced domain robustness, in both pretraining on single-object and multi-object image datasets. Our results and insights demonstrate how to ensure robustness through the choice of views in contrastive learning.

Introduction

Self-supervised learning has been rising in popularity in computer vision, with many top methods using contrastive learning (Hadsell, Chopra, and LeCun 2006), a form of learning that optimizes feature representations for positive samples to be close together and negative samples to be far apart. Contrastive models such as SimCLR (Chen et al. 2020a) and MoCo (He et al. 2020) have been shown to approach or surpass the performance of supervised models when representations are transferred to image classification, object detection, and semantic segmentation tasks (Ericsson, Gouk, and Hospedales 2021). This success is in part due to strategic data augmentation pipelines that create effective positive and negative *views* (samples) for learning.

Despite this progress, the out-of-distribution robustness of contrastive representations has been minimally studied, especially with regards to object detection. We hypothesize that existing data augmentation pipelines in contrastive learning may result in representations that lack robustness to various detection domain shifts. For example, as shown in Fig. 1, state-of-the-art pipelines (Chen et al. 2020a; Mo et al. 2021) create positive views with aggressive random cropping of a single image. This augmentation's use can lead to features for objects being made similar to the background or co-occurring objects, potentially causing contextual bias and hurting performance in out-of-context detection scenarios. Random cropping may also result in texture-biased (rather than shape-biased) features, since object shapes may not be consistent across crops. A lack of shape in representations can lead to degraded performance when texture is shifted (*e.g.* appearance changes due to rain or snow).

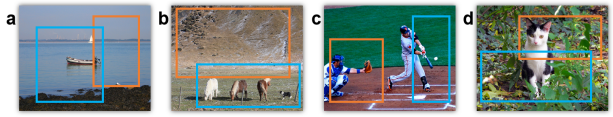

Figure 1: **What properties of contrastive views enable robust downstream object detection?** Existing contrastive view pipelines may cause detectors to lack robustness in domain-shifted settings (*e.g.* appearance, context). For instance, state-of-the-art pipelines use random cropping, which may cause contextual bias as objects can be aligned to common backgrounds (*e.g.* boat in **a**, horses in **b**) or co-occurring objects (*e.g.* player, glove, ball, and bat in **c**). Due to a lack of spatial consistency between views, its use may also discourage the learning of object shape (*e.g.* **d**).

In this work, we investigate various strategies to improve contrastive view design for enhanced robustness in such domain-shifted detection. In particular, we conduct an empirical study of a state-of-the-art detection pretext task InsLoc (Yang et al. 2021). We vary key components of the contrastive data augmentation pipeline (adjusting cropping, adding shortcut-reducing appearance augmentations, integrating saliency-based object priors) and evaluate the effectiveness of such changes in downstream appearance-shifted and context-shifted detection settings. We thoroughly evaluate these strategies in two practical settings of pretraining on single-object and multi-object image datasets.

We present these insights into contrastive view design:

- Increasing the minimum percentage of an image used as a crop or adding an IoU constraint between crops causes the model to use larger object parts to match views. This encourages the model to learn object shape and improves

domain robustness to appearance shifts.

- Combining non-aggressive crops and shortcut-reducing augmentations results in robust object features overall, exemplified by improvements over InsLoc in-domain (up to +2.73 AP) and out-of-domain (up to +3.07 AP).
- The use of saliency priors in views is effective for out-of-context robustness. Their use is best in a mechanism that removes background and tightens crops to object regions.
- Using shortcut-reducing augmentation on the non-salient regions in views, in combination with crop tightening and shape strategies, is an effective method for enhancing robustness in both appearance and context-shifted settings.

Background and Related Work

Self-supervised and contrastive learning Many top self-supervised methods use contrastive learning and the instance discrimination pretext task (Wu et al. 2018), where each image is its own class, and the goal is to discern that two positive samples (or *views*) are from the same image when considered versus a set of negative samples. Positives are generated through aggressive data augmentation of a single image, and they are compared versus a large number of negatives from other images. While the representations for negatives have been stored in large memory banks (Wu et al. 2018), recent methods optimize the learning pipeline to rather use large batch sizes (Chen et al. 2020a) or a dynamic dictionary (He et al. 2020). Alternatively, the online clustering approach of (Caron et al. 2020) avoids the need for pairwise comparisons entirely, and the iterative approach of (Grill et al. 2020) learns without negatives at all. In general, contrastive methods are evaluated by transferring representations to downstream tasks such as object detection, and *domain shifts are not usually considered with detection*. General detectors have been shown to lack robustness to shifts, retaining only 30-60% of performance when tested on natural corruptions (Michaelis et al. 2019), and contrastive detectors may similarly lack robustness. In this work, we fill in the need to more broadly characterize and improve the generalizability of contrastive representations through our study of view design strategies in domain-shifted detection.

Data augmentations and views in contrastive learning Recent work has explored how to construct positive and negative views for contrastive learning. In particular, the use of many negatives has led to research into how to handle false negatives (Chuang et al. 2020) and how to mine hard negatives (Kalantidis et al. 2020; Robinson et al. 2021). Positives have also been studied with regards to intra-image augmentations and instance discrimination. For instance, SimCLR (Chen et al. 2020a) finds that creating positives with random cropping, color distortion, and Gaussian blur is effective for ImageNet classification. We also explore augmentations for positive views, but consider those specifically targeting domain robustness in detection, *which include some previously unexplored in contrastive learning: Poisson blending, texture flattening, and elastic deformation*. We also alternatively find random cropping to only be domain-robust when used non-aggressively or with IoU constraints.

Robustness in contrastive learning Neural network representations have been shown to struggle to generalize to various domain shifts (*e.g*. pose (Alcorn et al. 2019), corruptions (Hendrycks and Dietterich 2019)) and to suffer from biases (*e.g*. texture (Geirhos et al. 2019), contextual (Singh et al. 2020), background (Xiao et al. 2021)). Contrastive representations face similar issues, for instance versus viewpoint shifts (Purushwalkam and Gupta 2020) and texture-shape conflicts (Geirhos et al. 2020). *Most works that strive to explicitly improve contrastive robustness either focus on image recognition ((Ge et al. 2021), (Khosla et al. 2020)), proxy recognition tasks like the Background Challenge (Xiao et al. 2021), or evaluate only when transferring representations to object detection, but not on domain shifts in detection.* We alternatively consider contrastive robustness with respect to object detection and relevant domain shifts. One shift we consider is *in context*, as some works have identified contextual bias as an issue in contrastive pretraining on multi-object scene images. For example, (Selvaraju et al. 2021) addresses contextual bias in COCO pretraining through constraining views to overlap with saliency maps and using a Grad-CAM attention loss, leading to performance gains over MoCo-v2 on COCO and VOC detection. Similarly, (Mo et al. 2021) proposes two augmentations, object-aware cropping (OA-Crop) and background mixup (BG-Mixup), to reduce a contrastive model's contextual biases. Notably, these cropping strategies have minimally been tested with detection (just in-domain), so it is unclear how such strategies perform in out-of-domain detection. We evaluate these strategies in out-of-domain scenarios and show that they do not always result in improvements. We thus propose a hybrid strategy of the methods and show that it substantially improves out-of-context robustness.

Unsupervised pretraining for object detection Our work also fits with recent approaches tailoring pretraining to downstream tasks besides image classification, such as object detection. In particular, we explore InsLoc, a pretext task in which detection-focused representations are built in contrastive learning through integration of bounding box information (Yang et al. 2021). Other notable approaches exist which leverage selective search region proposals (Wei et al. 2021), global and local views (Xie et al. 2021a), spatial consistency representation learning (Roh et al. 2021), fine-grained pretext tasks at the pixel level (Wang et al. 2021; Xie et al. 2021b), and transformers (Dai et al. 2021). *Our work is orthogonal to such works as we provide contrastive view design insights that can guide future detection-focused pretext tasks to additionally consider out-of-domain robustness.*

Experimental Approach

In this study, our goal is to analyze how strategic changes to contrastive views impact downstream object detection robustness. In particular, we consider two families of domain shifts that cause drops in object detection performance: appearance and context. *Appearance shift* in our study is defined as change in the visual characteristics of objects (such as color brightness and texture). *Context shift* in our study is defined as when an object appears with different objects or

in different backgrounds during train and test time.

Formally, we view contrastive learning from a knowledge transfer perspective. There is a source (pretext) task s and a downstream object detection task d. Unsupervised, contrastive pretraining is performed in s on dataset $\mathbf{s}$, then representations are transferred to d for supervised finetuning on dataset $\mathbf{d}$. We explore strategies to adapt s to s', which represents a task with views strategically augmented to be robust to families of domain shifts (*e.g.* appearance, context). To evaluate the effectiveness of s', representations from s' (after finetuning on $\mathbf{d}$) are evaluated on $\mathbf{d}$ and various $\mathbf{t_i}$, which represent target datasets domain-shifted from $\mathbf{d}$.

Base Pretext Task We select s to be Instance Localization (InsLoc) (Yang et al. 2021), a state-of-the-art detection-focused, contrastive pretext task. In particular, InsLoc is designed as an improvement over MoCo-v2 (Chen et al. 2020b) for the Faster R-CNN detector (Ren et al. 2015). Views are created through pasting random crops from an image at various aspect ratios and scales onto random locations of two random background images. Instance discrimination is then performed between foreground features obtained with RoI-Align (He et al. 2017) from the two composited images (positive *query* and *key* views) and from negatives maintained in a dictionary. This task is optimized with the InfoNCE loss (Van den Oord, Li, and Vinyals 2018).

Outlining the specifics of InsLoc's augmentation pipeline further, the crops used for views are uniformly sampled to be between 20-100% of an image. Crops are then resized to random aspect ratios between 1/2 and 2 and width and height scales between 128 and 256 pixels. Composited images have size 256×256 pixels. Other augmentations used include random applications of Gaussian blurring, horizontal flipping, color jittering, and grayscale conversion. Notably, InsLoc's appearance augmentations and cropping are characteristic of state-of-the-art contrastive methods (Chen et al. 2020a; He et al. 2020), *making InsLoc a fitting contrastive case study*.

General Strategies to Enhance InsLoc We propose multiple ways to augment the InsLoc view pipeline (*s'*) for enhanced robustness in appearance-shifted and out-of-context detection scenarios. First, as noted in Fig. 1, we consider cropping since InsLoc, like other contrastive methods (Chen et al. 2020a; He et al. 2020), uses aggressive random cropping to create positive views. As random cropping has been shown to bias a model towards texture (Hermann, Chen, and Kornblith 2020), we reason that InsLoc may struggle when texture shifts in detection such as when it is raining or snowing. Models that learn shape on the other hand can be effective in such situations (Geirhos et al. 2019). We thus explore simple strategies to encourage InsLoc to learn shape. In particular, we experiment with *geometric changes to crops* (increasing the minimum % of an image used in crops *m*, enforcing an IoU constraint *t* between views). We expect such changes to increase the spatial consistency between crops and encourage the model to learn object parts and shapes. In turn, the model can become more robust to texture shifts.

Furthermore, we consider using *shortcut-reducing appearance augmentations*, as we find that InsLoc may not adequately discourage the model from attending to shortcuts

Figure 2: Our shortcut-reducing augmentations of study, shown within InsLoc. We perform each augmentation on the crop in the query view (left), but not in the key view (right).

that are non-robust in appearance-shifted scenarios, such as color histograms, high-frequency noise, or texture. One strategy we explore is *Poisson blending* (Pérez, Gangnet, and Blake 2003), which is a method to seamlessly blend crops into a background image. We use Poisson blending instead of random copy-pasting in InsLoc to reduce contrast between foreground and background regions, making the pretext harder as the model cannot use contrast as a shortcut to solve the task. It is also found that Poisson blending can introduce random illumination effects from the background image, which may be desirable to learn invariance towards for appearance shifts. Second, we experiment with *texture flattening*, another application of using the Poisson equation, as its use results in washed out texture and brightness changes due to the preservation of gradients at only edge locations. We reason that this augmentation can be effective to teach the model to not overfit to high-frequency texture shortcuts. Last, we investigate *elastic deformation* (Simard et al. 2003), an augmentation that alters images by moving pixels with displacement fields. This augmentation can help make features more invariant to local changes in edges and thus noise shortcuts. We illustrate our proposed use of these augmentations in Fig. 2. Augmentations are notably applied 100% of the time, unless otherwise noted. Implementation-wise, we use Poisson blending and texture flattening as provided in the OpenCV library (Bradski 2000) and the algorithm of (Simard et al. 2003) for elastic deformation.

Lastly, we note that random cropping may result in the aligning of context (background and objects or objects and objects), which can lead representations that are contextually biased and not robust in out-of-context detection. To address this problem, we experiment with *strategies that leverage saliency-based object priors*, as they can enable crops to refer to specific object regions rather than to background or co-occurring objects. In particular, we experiment with two state-of-the-art approaches (Mo et al. 2021; Selvaraju et al. 2021), as well as a hybrid of such approaches. We compare each strategy out-of-domain, and also consider combining saliency strategies with shape and appearance strategies.

Pretraining and Finetuning Datasets We identify two pretraining scenarios $\mathbf{s}$ to evaluate the robustness of contrastive view design strategies. First is ImageNet pretraining (Krizhevsky, Sutskever, and Hinton 2012), a standard scenario for many contrastive approaches (Chen et al. 2020a; He et al. 2020). In ImageNet, the majority of images are *iconic*, containing a single, large, centered object. With the

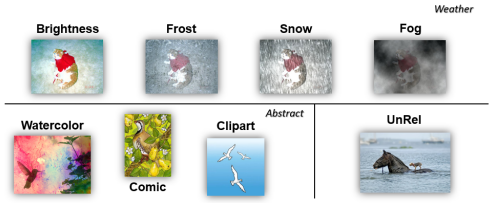

Figure 3: Our domain-shifted test sets. Weather and Abstract are for appearance shifts, and UnRel is for context shifts.

heavy computational nature of contrastive pretraining, along with our goal to repeat and conduct multiple experiments, we sample ImageNet from over 1 million to 50,000 images and call this set *ImageNet-Subset*. For a dataset with different properties, we also consider pretraining on COCO (Lin et al. 2014), which contains more *scene* imagery, having multiple, potentially small objects. With the goal of self-supervised learning to learn robust representations on large, uncurated datasets, which are likely to contain scene imagery, COCO is a practical dataset to study. Also its multi-object nature makes it an appropriate test case for benchmarking contextual bias downstream. We pretrain specifically with COCO2017train and do not sample since its size is relatively small (118,287 images).

We explore two finetuning datasets **d**: VOC (Everingham et al. 2010) when **s** is ImageNet-Subset and both VOC and COCO2017train when **s** is COCO. For VOC, we specifically use VOC0712train+val (16,551 images). We evaluate on COCO2017val (5,000 images) and VOC07test (4,952 images). In our sampling of ImageNet, we ensure semantic overlap with VOC by choosing 132 images for each of 379 classes from 13 synset classes that are related to VOC's classes: aircraft, vehicle, bird, boat, container, cat, furniture, ungulate, dog, person, plant, train, and electronics.

Domain Shift Datasets We select various datasets $\mathbf{t_i}$ for out-of-domain evaluation. First, when **d** is VOC, we test on the challenging, abstract Clipart, Watercolor, and Comic object detection datasets (Inoue et al. 2018), as they represent significant domain shifts in appearance. Clipart has the same 20 classes as VOC and 1,000 samples, while Watercolor and Comic share 6 classes with VOC and have 2,000 samples each. We take the average performance across the three sets and describe the overall set as *Abstract*. When **d** is COCO, we test on the out-of-context UnRel dataset (Peyre et al. 2017). This set captures relations that are "ununsual" between objects (*e.g.* car under elephant), making this set useful for evaluating out-of-context robustness. We evaluate on 29 classes which overlap with COCO thing classes (1,049 images). Lastly, for both VOC and COCO, we consider Pascal-C and COCO-C (Michaelis et al. 2019), a collection of sets that are synthetically domain-shifted on natural corruption types. In particular, we explore the appearance-based Weather split at severity level 5, which consists of brightness, fog, frost, and snow shifts. We refer to the overall set for VOC and COCO as VOC-Weather and COCO-Weather, respectively. Examples for the test sets are shown in Fig. 3.

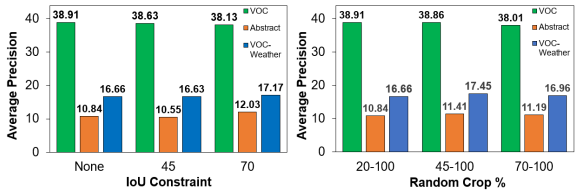

Figure 4: In-domain (VOC) vs. out-of-domain (Abstract, VOC-Weather) AP following InsLoc pretraining with various values for IoU constraint t and min % crop m, averaged over 3 trials. Note that the top-performing settings for domain robustness (45-100% crop, 0.70 IoU constraint) are different from the settings of the InsLoc baseline (20-100% crop and no IoU constraint).

Training Setup Pretraining is performed with the provided InsLoc implementation (Yang et al. 2021). Faster R-CNN (Ren et al. 2015), with a ResNet-50 backbone and FPN, serves as the trained detector. With high computational costs for contrastive pretraining, these experiments consider a fixed pretraining budget of 200 epochs. For COCO, pretraining is performed with per-GPU batch size 64 and learning rate 0.03 on 4 NVIDIA Quadro RTX 5000 GPUs with memory 16 GB. For ImageNet-Subset, pretraining is performed with per-GPU batch size 32 and learning rate 0.015 on 2 NVIDIA GeForce GTX 1080 Ti GPUs with memory 11 GB. All pretraining uses a dictionary size of K=8,192. Full finetuning of all layers is performed within the Detectron2 (Wu et al. 2019) framework with a 24k iteration schedule, a learning rate of 0.02, and a batch size of 4 on 2 NVIDIA GeForce GTX 1080 Ti GPUs, unless otherwise noted.

Experiments and Analysis

In this section, we outline various strategies for contrastive view design and evaluate their effectiveness in the InsLoc pretext task, considering both pretraining on ImageNet-Subset and COCO. Evaluation metrics are AP and AP_{50}.

Pretraining on ImageNet-Subset

How can we encourage contrastive learning to use shape to become more robust to appearance domain shifts? First, we consider appearance domain shifts in detection, where object shapes are preserved and texture is distorted. We wish to encourage InsLoc to capture object shapes for such scenarios and thus propose two simple strategies: (1) increasing the minimum % of an image sampled as a crop m and (2) adding an IoU contraint t, such that query and key crops must have at least such IoU. To evaluate these strategies, we pretrain InsLoc on ImageNet-Subset using two different values of m, 45% and 70%, in addition to InsLoc's default of 20%, while keeping the maximum crop bound as 100%. We also pretrain InsLoc with two IoU constraint values of t, 45% and 70%. We then perform finetuning on VOC and evaluate in-domain on VOC and out-of-domain on Abstract and VOC-Weather (two sets with distorted texture).

Results over three trials are shown in Fig. 4. Notably, we find that the default InsLoc settings (20-100% crops, no IoU

Method	VOC AP	Abstract AP	Weather AP
InsLoc, m=20 (Baseline)	38.91	10.84	16.66
+Poisson Blending, m=20	38.49	11.95	16.98
InsLoc, m=45	38.86	11.41	17.45
+Poisson Blending, m=45	40.22	12.96	19.23
InsLoc, m=70	38.01	11.19	16.96
+Poisson Blending, m=70	**40.54**	**13.00**	**19.73**
InsLoc, t=70	38.13	12.03	17.17
+Poisson Blending, t=70	39.31	11.97	18.36

Table 1: In-domain (VOC) vs. out-of-domain (Abstract, VOC-Weather) AP following InsLoc pretraining with and without Poisson blending InsLoc's query crop, for various values of min % crop m and IoU threshold t. Note that the use of our Poisson blending strategy results in substantial in-domain and out-of-domain gains, especially with m=70.

constraint) result in the best in-domain AP, but *not the top* out-of-domain AP. In particular, m=45 and t=70 have the highest out-of-domain AP for their respective value comparisons (up to +1.19 AP on Abstract, +0.79 AP on VOC-Weather). These results show that substantial overlap helps domain robustness, and there is a sweet spot for crop % that helps robustness. *Therefore, including more object parts and shapes in crops and enforcing spatial consistency are effective for improving robustness to appearance shifts*. We note that the sweet spot for m may be related to an observed tradeoff with in-domain AP, which drops as m or t increases. We reason that while we are encouraging the model to learn shape features, we are also increasing the probability that the model can attend to natural, high-frequency shortcuts in images since crops that have more area and overlap more might share more of these signals. We next consider augmentations to remove potential shortcuts and improve these strategies.

Can stronger appearance augmentations reduce shortcuts and make shape-based strategies more effective? Though contrastive view pipelines typically have significant appearance augmentations like Gaussian blur, grayscale conversion, and color jitter (Chen et al. 2020b), we reason that even more aggressive augmentations may be beneficial with our min % crop and IoU constraint strategies to further limit shortcuts and learn shape better. SimCLR (Chen et al. 2020a) serves as motivation, as the authors note that the model can use color histograms as a shortcut. We explore Poisson blending, texture flattening, and elastic deformation as augmentations to reduce similar shortcuts and enable InsLoc to learn robust object features more effectively.

To first test how augmentations interact with the % crop and IoU strategies, we perform Poisson blending with various values of m and t, shown in Table 1. Interestingly, different from Fig. 4, we find that the top domain robustness setting is m=70 (rather than t=70) and that *significant* out-of-domain gains are achieved in such setting (+2.16 AP on Abstract, +3.07 AP on VOC-Weather). Moreover, we find that *in-domain* AP also increases in this setting (+1.63 AP), indicating that *shortcut augmentations can enable the learning of shape without tradeoffs in-domain*. Note also that Poisson blending at m=20 is not effective in-domain and is less effective out-of-domain. We reason that shortcut augmentation is better at higher crop % since robust features like shape can be better captured there. The results at m=20 also indicate that *shortcut-reducing augmentations may not be effective unless learning robust features like shape is encouraged*.

Method	VOC AP	Abstract AP	Weather AP
InsLoc, m=20 (Baseline)	38.91	10.84	16.66
InsLoc, m=70	38.01	11.19	16.96
+Poisson Blending, m=70	40.54	13.00	**19.73**
+Elastic Deformation, m=70	40.94	13.26	18.70
+Texture Flattening, m=70	40.45	**13.57**	19.58
+Apply 25% of Time, m=70	**41.64**	12.53	19.70

Table 2: In-domain (VOC) vs. out-of-domain (Abstract, VOC-Weather) AP when pretraining InsLoc with shortcut-reducing augmentations at min % crop m=70. "Apply 25% of Time" means we either apply Poisson blending, elastic deformation, or texture flattening, or use the baseline setting, each with probability 25%.

We further use the top setting of m=70 to test each of Poisson blending, elastic deformation, and texture flattening, shown over three trials in Table 2. We find that all augmentations help both in-domain and out-of-domain. In particular, Poisson blending is the top for VOC-Weather (+3.07 AP) and texture flattening is for Abstract (+2.73 AP). We reason that texture flattening simulates the flattened texture of digital Abstract imagery well, while Poisson blending's random illumination effects are helpful for the texture changes seen with weather. Also shown in Table 2 is a scenario where we either apply one of the three augmentations or just use the default InsLoc setting, each with probability 25%. We find even more substantial gains in-domain (+2.73 AP) in such scenario. *These results indicate that creating views with shape-encouraging and shortcut-reducing appearance augmentations together leads to more robust object features overall, helping both in-domain and out-of-domain.* To our knowledge, we are the first to demonstrate such effectiveness of Poisson blending, texture flattening, and elastic deformation as augmentations for contrastive views.

Pretraining on COCO

How do saliency-based view strategies compare on out-of-context and appearance domain shifts? With pretraining on multi-object images (*e.g.* COCO), it has been noted that the intra-image cropping of contrastive pipelines is inadequate due to the potential to align different objects (Purushwalkam and Gupta 2020). To reduce the effects of contextual bias, object priors have been leveraged, with OA-Crop (Mo et al. 2021) and CAST (Selvaraju et al. 2021) representing two cropping methods. OA-Crop uses an initial pretraining of MoCo-v2 to gather Contrastive Class Activa-

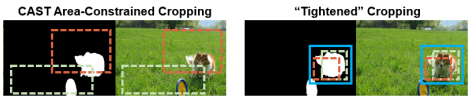

Figure 5: CAST (Selvaraju et al. 2021) vs. DeepUSPS-Tightened crops. The blue box shows a randomly chosen object crop that is the source for query and key views. Green and orange boxes show query and key crops.

Cropping Method	COCO AP_{50}	UnRel AP_{50}	Weather AP_{50}
InsLoc	26.16	22.60	10.53
+OACrop	25.55	23.10	9.82
+CAST	26.70	22.36	10.58
+DeepUSPS-Tight, m=8	27.76	24.59	12.20
+DeepUSPS-Tight, m=20	**28.39**	**26.11**	**12.33**

Table 3: Comparison of saliency-based view strategies, within InsLoc, pretrained/finetuned on COCO (24k sched.). For InsLoc, OACrop, and CAST, results are with the default or optimal cropping values if reported (m=20, 8, and 20 respectively). DeepUSPS-Tightened is tested at m=8 and 20.

tion Maps, and creates a number of object crops for an image from bounding boxes around salient regions. During training, one randomly selected object crop, rather than the entire image, is used to crop views. CAST, alternatively, ensures that crops overlap with saliency maps gathered with DeepUSPS (Nguyen et al. 2019), an "unsupervised" saliency detector (it still uses ImageNet-supervised weights).

Notably, these works have not been evaluated in out-of-domain detection, so we fill in this gap by comparing them within InsLoc. We also consider a hybrid approach called *DeepUSPS-Tightened* crops, where DeepUSPS saliency maps, rather than ContraCAMs, are used to create object crops like OA-Crop, as we observe DeepUSPS's maps are higher quality. We emphasize that the difference between between CAST and DeepUSPS-Tightened crops is that maps are used with CAST to ensure that crops *overlap* with objects, rather than to *reduce* background area and *tighten* crops to objects, which is the goal of the hybrid that we propose. We exemplify these differences in Fig. 5.

In Table 3, we compare each strategy following COCO pretraining and finetuning in terms of AP_{50} on COCO, the out-of-context UnRel, and the appearance-shifted COCO-Weather. We use the saliency maps provided by (Mo et al. 2021) and (Selvaraju et al. 2021), as well as their top reported settings (or default if not reported). We also test the hybrid DeepUSPS-Tightened crops at m=8 and m=20. A first observation is that there is a *significant* domain gap (-3.56 AP_{50}) between COCO and UnRel without incorporating any saliency-based strategy, *indicating that contextual bias exists in downstream object detection*. The OACrop strategy improves AP_{50} on UnRel, while CAST does not. Alternatively, CAST improves on COCO and COCO-Weather (slightly) while OA-Crop does not. Notably, *our hybrid DeepUSPS-Tightened leads to the top gains* vs. InsLoc (+2.23 AP_{50} on COCO, +3.51 AP_{50} on UnRel, +1.80 AP_{50} on COCO-Weather). The domain gap is much smaller than InsLoc as well (-2.28 vs. -3.56 AP_{50}).

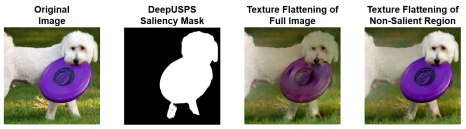

Figure 6: Our proposed texture flattening augmentation.

Method	COCO AP_{50}	UnRel AP_{50}	Weather AP_{50}
InsLoc	26.16	22.60	10.53
+DeepUSPS-Tight/TF	27.73	25.27	11.94
+DeepUSPS-Tight/TFNS	**28.74**	**25.71**	**12.44**

Table 4: In-domain (COCO) vs. out-of-domain (UnRel, COCO-Weather) AP_{50} for full image (TF) versus non-salient region texture flattening (TFNS) of InsLoc query crops in combination with DeepUSPS-Tightened strategy (at m=70), with respect to InsLoc baseline (at m=20).

We note that the high quality of DeepUSPS maps along with the *removal* of background (through cropping) are potential reasons for DeepUSPS-Tightened to have the top AP_{50}. In detection pretraining, we reason that it may not be important to include background in crops, and that *strategic use of a quality saliency detector can enable detectors to be less biased and more robust out-of-context.*

Can shape and appearance strategies further help robustness with saliency strategies? While tightened crops remove many non-salient pixels, some still remain inside crops. The pretext task can be solved by matching positives based on the background, rather than objects, which can hurt out-of-context robustness. Moreover, aggressive cropping of even object-based crops can still lead to representations that do not capture shape well, hurting appearance shift robustness. With inspiration from the ImageNet-Subset experiments, we explore shape and appearance strategies with COCO pretraining as well. In particular, we consider m=70 as a shape strategy and texture flattening as an appearance strategy. Since we have saliency maps, we also propose one more strategy: texture flattening of non-salient regions only (TFNS). We specifically propose to distort the background (non-salient image regions marked by DeepUSPS) of one view in InsLoc to encourage background invariance between views during pretraining. We reason that TFNS could be effective for out-of-context robustness as shortcuts may come more significantly from the background, rather than salient regions. We illustrate our proposed augmentation in Fig. 6.

In Table 4, we show results with m=70 and texture flattening (full and just non-salient regions) with DeepUSPS-Tightened crops. We observe that both strategies are effective, and TFNS leads to larger gains across all sets. We reason that learning some level of texture, even with shape, is important, and TFNS preserves important texture (of ob-

Method	Crop%	Aug	*VOC Finetuning*			*COCO Finetuning*		
			VOC AP	**Abstract** AP	**Weather** AP	**COCO** AP_{50}	**UnRel** AP_{50}	**Weather** AP_{50}
InsLoc	20-100	Default	39.79	11.31	18.47	26.16	22.60	10.53
	70-100	Default	38.45	12.19	17.37	24.70	20.87	9.56
	20-100	TFNS	40.43	12.45	19.42	26.94	23.00	11.37
	70-100	TFNS	40.09	12.56	19.05	27.12	23.03	11.25
InsLoc +DeepUSPS-Tightened	8-100	Default	40.87	11.03	19.85	27.76	24.59	12.20
	20-100	Default	41.56	11.78	19.86	28.39	26.11	12.33
	70-100	Default	39.30	12.11	17.69	25.46	22.66	10.67
	8-100	TFNS	41.08	12.48	**20.64**	28.44	**27.31**	12.27
	20-100	TFNS	41.70	12.50	19.54	28.66	24.93	12.12
	70-100	TFNS	**41.80**	**13.18**	20.46	**28.74**	25.71	**12.44**

Table 5: InsLoc+DeepUSPS-Tightened vs. InsLoc at various min crop % and with application of texture flattening of non-salient regions (TFNS) (24k schedule for both VOC and COCO finetuning). Underlined=top per strategy, bold=best overall.

jects) while removing unimportant texture (high-frequency shortcuts which come from the background).

For a more thorough evaluation of TFNS and minimum crop %, in Table 5, we present results testing more changes to *m* for the InsLoc baseline and DeepUSPS-Tightened strategy, with finetuning on both VOC and COCO. We find that the combination of DeepUSPS-Tightened, *m*=70, and TFNS results in the top AP on VOC and Abstract and the top AP_{50} on COCO and COCO-Weather. *These results show that combining shape, appearance, and saliency strategies are effective both in-domain and out-of-domain.* We do observe that the top performance on UnRel (+4.71 AP_{50}) is achieved at *m*=8, along with TFNS and DeepUSPS-Tightened. We reason that since context is a "natural" domain shift, where texture of objects is preserved, shape may less useful, and aggressive cropping at *m*=8 of *crops with mostly salient pixels* can result in effective texture features. Such texture features may also be strong enough even when appearance is shifted, explaining the high performance on VOC-Weather. A last note is that we find TFNS to be most effective when applied at *m*=70, which makes sense as more non-salient pixels exist in such views.

How does texture flattening of non-salient regions compare to other background strategies? We gain further understanding of the effectiveness of TFNS through evaluating it versus replacing the query crop's non-salient pixels with a random grayscale value, a top background debiasing strategy (Ryali, Schwab, and Morcos 2021; Zhao et al. 2021). Results are shown for COCO pretraining and VOC finetuning in Table 6. We find that TFNS outperforms the grayscale strategy on all sets. We reason that TFNS is more beneficial for background debiasing as in distorting the background, it maintains continuity between an image's salient and non-salient pixels, making images seen in pretraining more natural and closer to those seen at test time.

How are results at a longer training schedule? In Table 7, we lastly show that our strategies maintain effectiveness with COCO at a longer training schedule (2x). Gains can notably be observed on *both* out-of-domain test sets: +1.58 AP on UnRel and +0.27 AP on COCO-Weather.

Method	**VOC** AP	**Abstract** AP	**Weather** AP
RandGrayBG	40.99	12.97	18.36
TFNS	**41.80**	**13.18**	**20.46**

Table 6: InsLoc+DeepUSPS-Tightened (*m*=70), using query crops with texture-flattened non-salient regions (TFNS) or random grayscale backgrounds (RandGrayBG).

Method	**COCO**		**UnRel**		**Weather**	
	AP	AP_{50}	AP	AP_{50}	AP	AP_{50}
InsLoc	29.63	46.50	25.88	41.68	14.25	23.32
Ours	**30.08**	**47.30**	**27.46**	**42.93**	**14.52**	**23.81**

Table 7: InsLoc+DeepUSPS-Tightened (*m*=70) with texture flattening of non-salient regions vs. InsLoc baseline, pretrained and finetuned on COCO with 2x schedule.

Conclusion

In this work, we present contrastive view design strategies to improve robustness to domain shifts in object detection. We show that we can make the contrastive augmentation pipeline more robust to domain shifts in appearance through encouraging the learning of shape (with higher minimum crop % and IoU constraints). Furthermore, combining these shape strategies with shortcut-reducing appearance augmentations is shown to lead to more robust object features overall, demonstrated by both in-domain and out-of-domain performance improvements. Finally, when pretraining on multi-object image datasets with saliency map priors, we find that tightening crops to salient regions, along with texture flattening the remaining non-salient pixels in a view, is an effective strategy to achieve out-of-context detection robustness. Overall, these strategies can serve to guide view design in future detection-focused, contrastive pretraining methods.

References

Alcorn, M. A.; Li, Q.; Gong, Z.; Wang, C.; Mai, L.; Ku, W.-S.; and Nguyen, A. 2019. Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4845–4854.

Bradski, G. 2000. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.

Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; and Joulin, A. 2020. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33: 9912–9924.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020a. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, 1597–1607. PMLR.

Chen, X.; Fan, H.; Girshick, R.; and He, K. 2020b. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.

Chuang, C.-Y.; Robinson, J.; Lin, Y.-C.; Torralba, A.; and Jegelka, S. 2020. Debiased contrastive learning. *Advances in Neural Information Processing Systems*, 33: 8765–8775.

Dai, Z.; Cai, B.; Lin, Y.; and Chen, J. 2021. UP-DETR: Unsupervised pre-training for object detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1601–1610.

Ericsson, L.; Gouk, H.; and Hospedales, T. M. 2021. How Well Do Self-Supervised Models Transfer? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5414–5423.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The Pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2): 303–338.

Ge, S.; Mishra, S.; Li, C.-L.; Wang, H.; and Jacobs, D. 2021. Robust Contrastive Learning Using Negative Samples with Diminished Semantics. *Advances in Neural Information Processing Systems*, 34.

Geirhos, R.; Narayanappa, K.; Mitzkus, B.; Bethge, M.; Wichmann, F. A.; and Brendel, W. 2020. On the surprising similarities between supervised and self-supervised models. *arXiv preprint arXiv:2010.08377*.

Geirhos, R.; Rubisch, P.; Michaelis, C.; Bethge, M.; Wichmann, F. A.; and Brendel, W. 2019. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *7th International Conference on Learning Representations, ICLR*.

Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; et al. 2020. Bootstrap your own latent - A new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33: 21271–21284.

Hadsell, R.; Chopra, S.; and LeCun, Y. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*. IEEE.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9729–9738.

He, K.; Gkioxari, G.; Dollar, P.; and Girshick, R. 2017. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Hendrycks, D.; and Dietterich, T. 2019. Benchmarking neural network robustness to common corruptions and perturbations. In *7th International Conference on Learning Representations, ICLR*.

Hermann, K.; Chen, T.; and Kornblith, S. 2020. The origins and prevalence of texture bias in convolutional neural networks. *Advances in Neural Information Processing Systems*, 33: 19000–19015.

Inoue, N.; Furuta, R.; Yamasaki, T.; and Aizawa, K. 2018. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5001–5009.

Kalantidis, Y.; Sariyildiz, M. B.; Pion, N.; Weinzaepfel, P.; and Larlus, D. 2020. Hard negative mixing for contrastive learning. *Advances in Neural Information Processing Systems*, 33: 21798–21809.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33: 18661–18673.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25.

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.

Michaelis, C.; Mitzkus, B.; Geirhos, R.; Rusak, E.; Bringmann, O.; Ecker, A. S.; Bethge, M.; and Brendel, W. 2019. Benchmarking Robustness in Object Detection: Autonomous Driving when Winter is Coming. *CoRR*, abs/1907.07484.

Mo, S.; Kang, H.; Sohn, K.; Li, C.-L.; and Shin, J. 2021. Object-aware contrastive learning for debiased scene representation. *Advances in Neural Information Processing Systems*, 34.

Nguyen, T.; Dax, M.; Mummadi, C. K.; Ngo, N.; Nguyen, T. H. P.; Lou, Z.; and Brox, T. 2019. DeepUSPS: Deep robust unsupervised saliency prediction via self-supervision. *Advances in Neural Information Processing Systems*, 32.

Pérez, P.; Gangnet, M.; and Blake, A. 2003. Poisson image editing. In *ACM SIGGRAPH 2003 Papers*, 313–318.

Peyre, J.; Sivic, J.; Laptev, I.; and Schmid, C. 2017. Weakly-supervised learning of visual relations. In *Proceedings of the IEEE International Conference on Computer Vision*, 5179–5188.

Purushwalkam, S.; and Gupta, A. 2020. Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases. *Advances in Neural Information Processing Systems*, 33: 3407–3418.

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.

Robinson, J. D.; Chuang, C.; Sra, S.; and Jegelka, S. 2021. Contrastive Learning with Hard Negative Samples. In *9th International Conference on Learning Representations, ICLR*.

Roh, B.; Shin, W.; Kim, I.; and Kim, S. 2021. Spatially consistent representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1144–1153.

Ryali, C.; Schwab, D. J.; and Morcos, A. S. 2021. Learning Background Invariance Improves Generalization and Robustness in Self-Supervised Learning on ImageNet and Beyond. *Workshop on ImageNet: Past, Present, and Future, held in conjunction with NeurIPS*.

Selvaraju, R. R.; Desai, K.; Johnson, J.; and Naik, N. 2021. CASTing your model: Learning to localize improves self-supervised representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11058–11067.

Simard, P. Y.; Steinkraus, D.; Platt, J. C.; et al. 2003. Best practices for convolutional neural networks applied to visual document analysis. In *ICDAR*, volume 3.

Singh, K. K.; Mahajan, D.; Grauman, K.; Lee, Y. J.; Feiszli, M.; and Ghadiyaram, D. 2020. Don't judge an object by its context: Learning to overcome contextual bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11070–11078.

Van den Oord, A.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv e-prints*, arXiv–1807.

Wang, X.; Zhang, R.; Shen, C.; Kong, T.; and Li, L. 2021. Dense contrastive learning for self-supervised visual pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3024–3033.

Wei, F.; Gao, Y.; Wu, Z.; Hu, H.; and Lin, S. 2021. Aligning pretraining for detection via object-level contrastive learning. *Advances in Neural Information Processing Systems*, 34.

Wu, Y.; Kirillov, A.; Massa, F.; Lo, W.-Y.; and Girshick, R. 2019. Detectron2. https://github.com/facebookresearch/detectron2.

Wu, Z.; Xiong, Y.; Yu, S. X.; and Lin, D. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3733–3742.

Xiao, K.; Engstrom, L.; Ilyas, A.; and Madry, A. 2021. Noise or signal: The role of image backgrounds in object recognition. In *9th International Conference on Learning Representations, ICLR*.

Xie, E.; Ding, J.; Wang, W.; Zhan, X.; Xu, H.; Sun, P.; Li, Z.; and Luo, P. 2021a. DetCo: Unsupervised Contrastive Learning for Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 8392–8401.

Xie, Z.; Lin, Y.; Zhang, Z.; Cao, Y.; Lin, S.; and Hu, H. 2021b. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16684–16693.

Yang, C.; Wu, Z.; Zhou, B.; and Lin, S. 2021. Instance localization for self-supervised detection pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3987–3996.

Zhao, N.; Wu, Z.; Lau, R. W.; and Lin, S. 2021. Distilling localization for self-supervised representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 10990–10998.