Quantifying Context Bias in Domain Adaptation for Object Detection

Anonymous authors
Paper under double-blind review

Abstract

Domain adaptation for object detection (DAOD) has become essential to counter performance degradation caused by distribution shifts between training and deployment domains. However, a critical factor influencing DAOD—context bias resulting from learned foreground-background (FG–BG) associations—has remained underexplored. In this work, we present the first comprehensive empirical and causal analysis specifically targeting context bias in DAOD. We address three key questions regarding FG–BG associations in object detection: (a) are FG–BG associations encoded during the training, (b) is there a causal relationship between FG–BG associations and detection performance, and (c) is there an effect of FG–BG association on DAOD. To examine how models capture FG–BG associations, we analyze class-wise and feature-wise performance degradation using background masking and feature perturbation, measured via change in accuracies (defined as drop rate). To explore the causal role of FG–BG associations, we apply do-calculus on FG–BG pairs guided by class activation mapping (CAM). To quantify the causal influence of FG–BG associations across domains, we propose a novel metric—domain association gradient—defined as the ratio of drop rate to maximum mean discrepancy (MMD). Through systematic experiments involving background masking, feature-level perturbations, and CAM, we reveal that convolution-based object detection models encode FG–BG associations. These associations substantially impact detection performance, particularly under domain shifts where background information significantly diverges. Our results demonstrate that context bias not only exists but also causally undermines the generalization capabilities of object detection models across domains. Furthermore, we validate these findings across multiple models and datasets, including state-of-the-art architectures such as ALDI++. This study highlights the necessity of addressing context bias explicitly in DAOD frameworks, providing insights that pave the way for developing more robust and generalizable object detection systems.

1 Introduction

Domain adaptation for object detection (DAOD) has been studied extensively to enable object detectors to perform well on datasets with distribution shifts from the training data (Kay et al., 2024; Chen et al., 2022; Deng et al., 2021; Hoyer et al., 2023; Li et al., 2022b; Koh et al., 2021; Kalluri et al., 2023). It is well known that there's an entanglement between background and foreground features in object detection, leading to a phenomenon called context bias in DAOD (Torralba & Efros, 2011; Divvala et al., 2009; Khosla et al., 2012; Zhang et al., 2024; Choi et al., 2012; Shetty et al., 2019). Here, significant differences in background features between the source and target domains can cause a notable decline in the quality and number of detections, even when the foreground features remain unchanged. Recent studies in image classification (Li et al.; Aniraj et al., 2023) and segmentation (Zhu et al., 2024; Chen et al., 2021; Dreyer et al., 2023) have attempted to mitigate context bias by minimizing this association. Oliva & Torralba (2007) demonstrated that context bias could result in the corruption of foreground objects by contextually correlated backgrounds, substantially degrading detection quality. However, there has been no prior work specifically analyzing the impact of context bias in DAOD. This work aims to address this gap.

In the realm of human cognition, the brain can accurately and instantly recognize foreground-background (FG–BG) associations without extensive training (Papale et al., 2018). Several studies, including Zhang et al. (2023); Poort et al. (2016); Papale et al. (2018); Huang et al. (2020), have investigated the processes of background suppression and foreground representation to understand the scene and temporal dynamics of foreground and background modulation in the brain. These insights can be applied to the field of computer vision for DAOD through a comprehensive analysis of the representation of FG–BG associations.

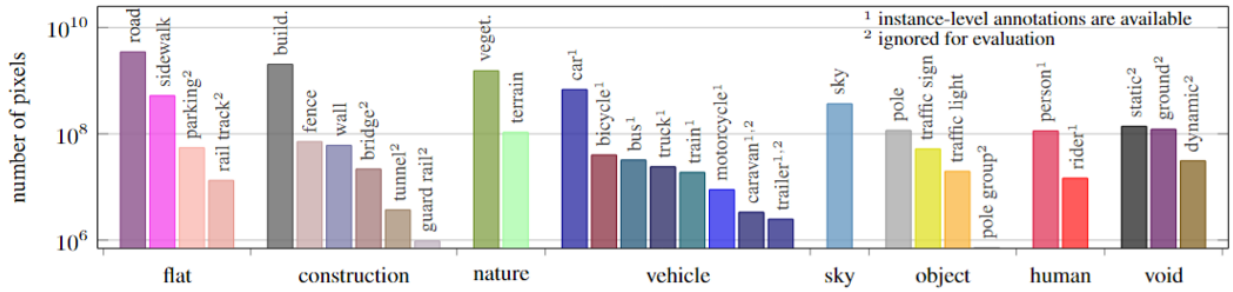


Figure 1: The proportion of background pixels in Cityscapes Cordts et al. (2016) are the highest of all classes. The image is from Cityscapes publication.

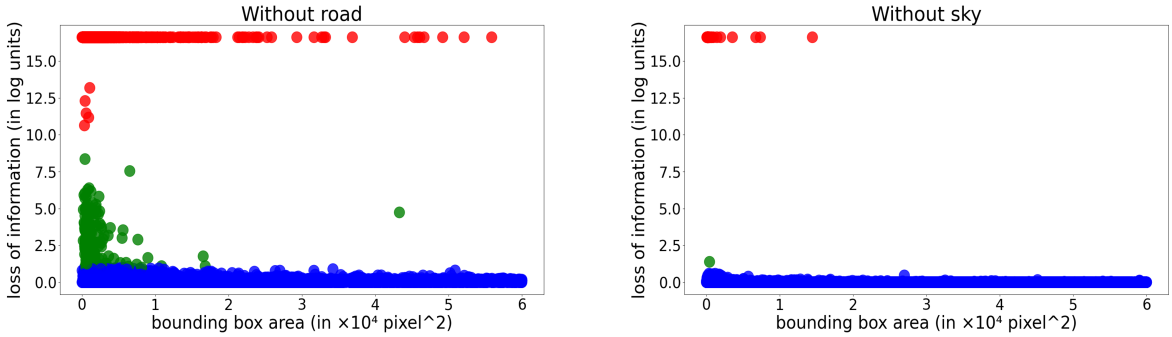


Figure 2: **Loss of information as a function of the bounding box area of the object.** Left side figure shows the suppression of "road" while right side figure shows the suppression of "sky". The dots are grouped into three clusters - red indicates missed detections (maximum information loss), green indicates partial matches (significant loss). Blue dots indicate no change. Label of "road" is more associated with "car" than "sky.

1.1 Our observations on context bias

To motivate our problem, we first looked at the proportion of background features in autonomous driving datasets, as an example. For Cityscapes dataset (Cordts et al., 2016), the number of pixels from built-up features (such as "road" and "sidewalk") are much greater than the foreground object pixels (see Fig. 1). Based on semantic segmentation outcomes (Alonso et al., 2021; Wang et al., 2020; Erisen, 2024), "road" has the highest accuracy and lowest variability. As a motivating experiment, we aimed to quantify the change in performance as a function of the background masking for a real dataset. We used the second layer ($res2.2$) of ResNet-50 backbone in the Detectron2 (Wu et al., 2019), trained on the Cityscapes dataset for object detection. We hypothesized that $res2.2$ effectively balances low and high-level features on FG–BG. The loss of information was computed when activated features for the specific background regions were zeroed out using semantic labels as a function of the ground truth bounding box area of the foreground objects. We defined performance drop ΔIoU as $0 \leq \Delta IoU \leq 1$ and the amount of loss of information as the negative log of the complement of ΔIoU ($-log(\Delta IoU)$). It computed change of intersection over union (IoU) with background masking. Figure 2 shows the performance drop of the removal of "road" as opposed to "sky".

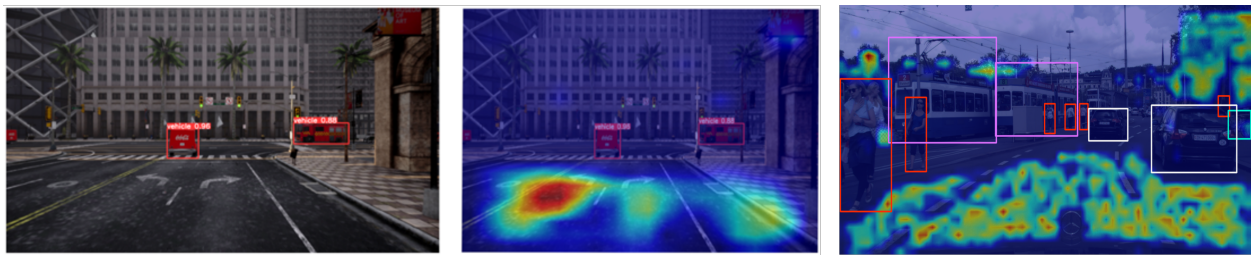


Figure 3: The left figure shows 2D inference of CARLA dataset using YOLOv4 model while the center figure shows the CAM attention map of the inference using EigenCAM (Muhammad & Yeasin, 2020). The right figure is EigenCAM result of YOLOv11 trained on Cityscapes. Road and vegetation are significantly enhanced.

We found that the loss of information is much higher with the "road" suppressed as compared to "sky", which means that "road" has more contextual association with vehicles, particularly when the vehicle size is small.

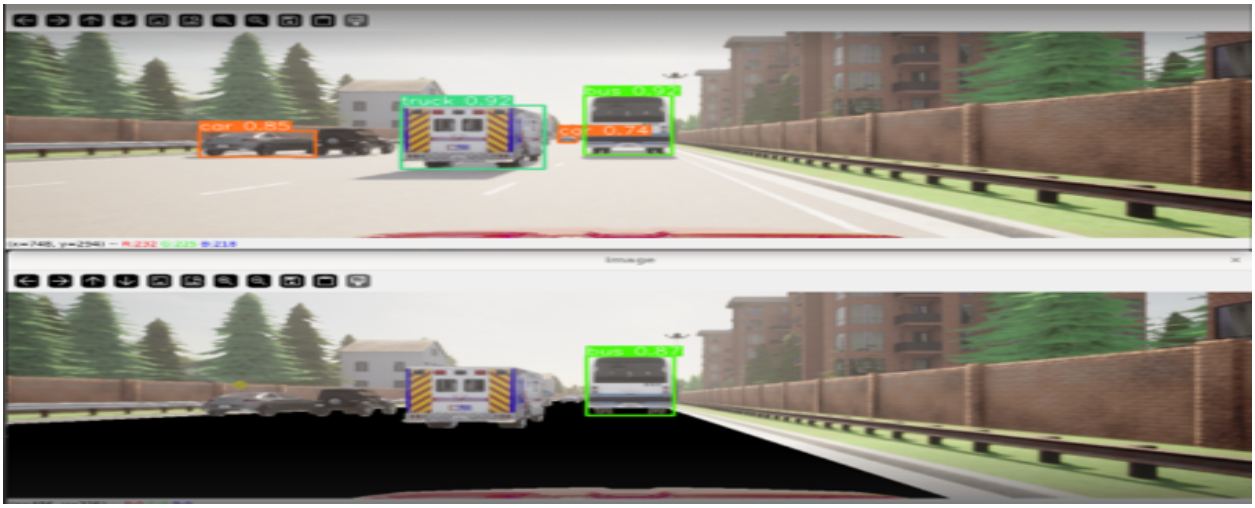


Figure 4: Masking the road on CARLA image and generating inference using YOLOv8 model - top figure shows that 3 out of 4 vehicles get detected correctly while in the bottom figure with road masking, only one vehicle is detected.

We trained a YOLOv4 detection model (Bochkovskiy et al., 2020) on a sample CARLA (Dosovitskiy et al., 2017) dataset collected under sunny conditions and provided inference on a separate CARLA dataset collected under cloudy conditions. We found that the model was focused on the road in front of the vehicles rather than the vehicles themselves (see Fig. 3) using class activation mapping (CAM). Additionally, to capture whether the same issue arises across different types of models, we performed an analogous experiment where we transformed the road pixels by masking them and found that YOLOv8 model was unable to detect most of the vehicles otherwise detected in the normal image (see Fig. 4). This outcome suggests that convolution-based neural network model may have implicitly learned to associate vehicles with road environments, leading to poor performance in detecting vehicles when a different background is present.

In order to understand the spread of feature pattern across domains, we plotted the foreground and background feature distributions of different domains using UMAP (McInnes et al., 2018). Figure 5 presents the visualization of the foreground and background features from different domains in 2D. We used the features of "Car" at the 5th ResNet layer ($res.5.2$) from different data distributions. We assumed that $res.5.2$ captures high-level features of FG and BG. The interesting finding was the differences of background alignment across the comparisons. It was immediately apparent that as the target domain shifts away from the source

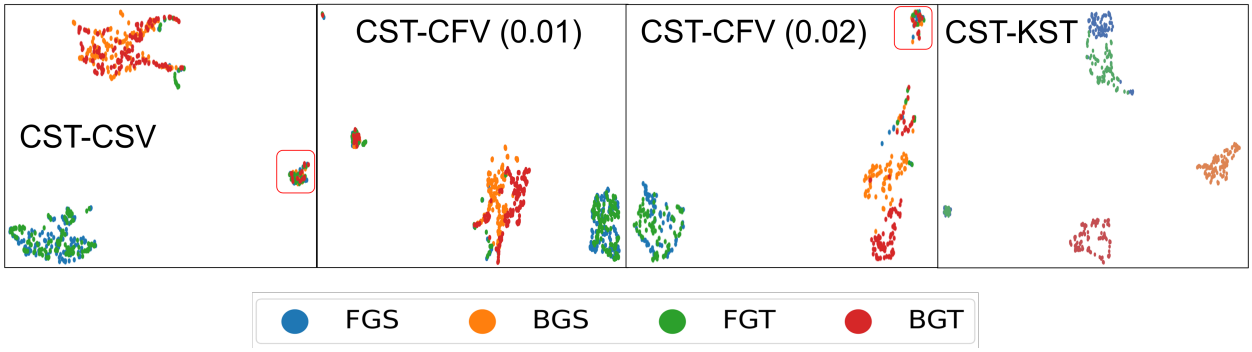


Figure 5: **UMAP feature embedding results.** "CST" is Cityscapes train, "CSV" is Cityscapes validation, "CFV (0.0X)" are Cityscapes foggy validation with different parameters for fogginess. "KST" means KITTI semantic train. We used the "car" label, as it is the most common category across the datasets. "FGS" is foreground in source domain and "BGS" is background in source domain. "FGT" and "BGT" are foreground and background in target domain. Similar colors should be intertwined, indicating similar feature distribution.

domain, the background became more separable than the foreground. For the Cityscapes training (CST) and validation (CSV) dataset, foreground and background features were distinguishable from each other but appeared intermingled between the source and target domains. In the CST-CFV panel, foreground features remained clustered together while the background features were separable but overlapping. We saw an extreme case in the CST-KST where the foreground features between CST and KST were next to each other but were non-overlapping while the background features were very distant from each other. The process to extract features are illustrated in Method section 3.

Prior studies (Choi et al., 2012; Torralba, 2003) have researched context bias for object detection and classification. The studies pointed out that relying only on local features (foreground features in our case) has limitations, including degraded quality due to noise and ambiguity in the target search space. They extended the likelihood to incorporate context information surrounding the foreground, which enhances object classification and detection by providing a stronger conditional probability like the Equation 1. The conditional probability of the object (O) given the features (f) was given as:

$$P(O|f) = P(O|F, B) = \frac{P(F|O, B)P(O|B)}{P(F|B)} \tag{1}$$

where F and B are the foreground and background features. However, it did not address DAOD issues like sim-to-real transfer and the root causes of FG–BG associations during training and inference remain unclear, especially given the causal relationships imposed post-detection.

The challenge with using a convolutional neural network (CNN) assuming identical and independent distribution (i.i.d) to estimate likelihood is the inability to explicitly teach the model to learn each factor in a specific order (Schölkopf et al., 2021; Agrawal et al., 2019). In other words, it means that parameters of CNN can be different depending on how it can be trained like the equation 2. The modeling can also be interpreted as:

$$P(O|f) = P(O|B, F) = \frac{P(B|O, F)P(O|F)}{P(B|F)} \tag{2}$$

In CNNs, likelihood estimation is a process to find the mean of a distribution with proper priors, which necessitates more samples to accurately estimate the true mean. This aligns with the principle that a more extensive and refined dataset, achieved through data augmentation, is crucial for better performance Taylor & Nitschke (2018). However, such datasets typically do not account for FG–BG associations, which is subtle

to capture during data collection. In summary, FG–BG associations can disrupt the trained estimation process for each probability estimate, leading to performance degradation in target domains due to these broken associations.

From these observations and hypotheses, our fundamental questions are as follows:

Q1. Are FG–BG associations being inadvertently learned during the training process?

Deep learning identifies latent patterns that optimize objective functions, typically by maximizing data likelihood. During the feature extraction process, models may learn spurious or unexpected features if such features improve predictive performance without any understanding of causality by following the previous reasoning (Bishop, 2006; Goodfellow et al., 2016; MacKay, 2003; Murphy, 2012). This underscores the importance of incorporating causal reasoning into deep learning frameworks to improve robustness and generalization (Schölkopf et al., 2021). From our motivation (see Fig. 2, 3, and 4), we conducted two experiments for Q1 by performing class-wise and feature-wise background removal.

These experiments were designed to capture the existence of FG–BG associations under fair conditions. From these findings about FG–BG associations, we pose the following question: **To what extent do FG–BG associations affect model accuracy?** To address this, we conducted a series of experiments aimed at capturing, representing, and quantifying the impact of FG–BG associations across domains, leading into the next set of questions (Q2 and Q3).

Q2. Is there a causal relationship between FG–BG associations and object detection?

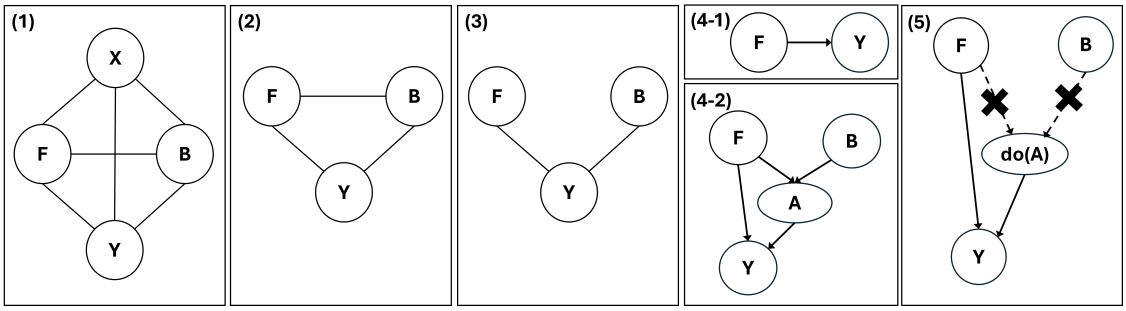


Figure 6: **Identification process to generate causal graph.** (1) is the first step from a complete undirected graph where "X" is an image, "F" is foreground, "B" is background, and "Y" is an outcome. "X" is removed because the object detection model will engineer on the input. So it generates (2). (3) is induced because "F" and "B" are independent. (4-1) is the optimal factor graph we derive that the outcome is solely related to "F". (4-2) indicates we captured from preliminary experiments (see Introduction 1). "A" indicates FG–BG associations. (5) is the factor graph with do-calculus to quantify the causal effect of "A" on "Y". By applying the backdoor adjustment, we can quantify the causal effect.

We employed graph-based causal analysis to investigate the causal effect of FG–BG associations on object detection performance. As illustrated in Figure 6, a causal model was constructed using the PC algorithm Glymour et al. (2019) to infer causal relationships within the object detection pipeline. In the context of a graphical causal model (F →Y ←B), it represents the joint distribution P(Y, F, B), which can be decomposed as either $P(Y|F,B)P(F|B)P(B)$ or $P(Y|F,B)P(B|F)P(F)$ (see the panel (3) in Figure 6, Equation 1 and 2). This is a spurious factor leading to FG–BG associations in CNN training. Causal identification further enables through intervention analysis (do-calculus). The causal effect of FG–BG associations (A) on object detection performance (Y) can be expressed as $P(Y|\text{do}(A)) = \sum_F P(Y|A,F), P(F)$, following the backdoor adjustment formula. This effect can be quantified by the expected difference $\mathbb{E}[Y|A=0,F] - \mathbb{E}[Y|A=1,F]$, which captures the impact of FG–BG association (A) on detection outcomes (Y), conditioned on contextual features (F) (see Fig. 6). To this end, we designed an experiment through intervention via CAM and instance masks. By combining CAM with ground-truth instance masks, we performed do-calculus interventions

incurring backdoor adjustment to model the causal influence of FG–BG associations on object detection accuracy. The combination of CAM and instance masks controlled activated background regions for each instance with different thresholds.

Q3. What is the impact of FG–BG associations on DAOD and how can the effect be quantified?

While the causal association with detection outcomes (Y) learned in source domains typically remains stable, the conditional distribution $P(Y|A)$ may shift due to background distribution changes. This can weaken the causal strength in target domains, resulting in performance degradation. We quantified this effect using the domain association gradient (which we refer to as *Gradient* in the rest of the paper), which captures the impact of FG–BG associations and using summation of intertwined features within and across domains respectively. We note that prior domain adaptation research emphasizes FG feature alignment exclusively (for e.g. adversarial classifiers and moment matching) and thus our effort is to showcase the gap in literature. For intra-domain analysis, we applied the maximum mean discrepancy (MMD) metric to compare contextual features of instances clustered by associated and non-associated background features. For cross-domain evaluation, we measured the feature discrepancies using MMD of associated and non-associated groups across domains. We note that we termed it *Gradient* since it quantifies the change of the response due to the change in FG–BG association.

Our experimental procedure is described in detail in Section 3.

1.2 Contributions

Our main contributions are as follows:

- We highlight a crucial gap, suggesting that considering alleviation of context bias is essential for enhancing the generalization and robustness of models across various environments by quantifying its effect. None of the current approaches investigate how context bias can manifest across various domains. We examine the issue of DAOD in relation to context bias.
- We analyze FG–BG associations and the causal relationship through the drop rate and do-calculus. Additionally, we employ distance-based metrics to measure the association between foreground and background under domain shifts.
- We also propose an additional metric, domain association gradient, to quantify the context bias on the source and target domain, respectively. To our knowledge, this is the first work to quantify this phenomenon.
- We provide a novel and practical research perspective by framing context bias as a critical factor in cross-domain object detection. Our study follows a logical progression from empirical observation to theoretical analysis, followed by quantitative and qualitative evaluations resulting in convincing evidence that FG–BG associations significantly affect domain adaptation performance.

2 Related Work

2.1 FG–BG Associations and Context Bias

There have been a number of studies aimed at improving performance in tasks such as classification, object recognition, and object localization researching background influence (Xiao et al., 2020; Liang et al., 2023; Zhang et al., 2007; Ribeiro et al., 2016; Zhu et al., 2016; Rosenfeld et al., 2018; Barbu et al., 2019; Sagawa et al., 2019) and context bias (Torralba & Efros, 2011; Khosla et al., 2012; Choi et al., 2012; Shetty et al., 2019). Xiao et al. (2020) and Zhu et al. (2016) studied effect of background on classification accuracy by modifying images with different combinations of foreground and background. Choi et al. (2012) proposed a graphical model that constructed FG–BG associations using conditional probability serving as a methodological inspiration for us. Several studies have addressed context bias using techniques such as

data augmentation to generate out-of-distribution samples in the background, combination of naturally unmatched backgrounds and foregrounds (e.g., an elephant in aroom), and applying background removal during training. Torralba (2003) demonstrated that background effects can be factorized into object priming, focus of attention, and scale selection by modeling the FG–BG associations in a probabilistic model. Liang et al. (2023) studied background influence using fashion datasets (Jia et al., 2020; Takagi et al., 2017). These studies (Zhai et al., 2024; Wu et al., 2022) localized foreground objects better than CAM-based algorithms without using bounding box information and with only classification labels. These prior works focused on context bias within the same domain and used datasets with smaller variations such as centered objects or single objects. Thus, we found that there remains a gap in understanding how context bias affects DAOD.

2.2 Domain Adaptation for Object Detection

Different variations of DAOD methods have been proposed using feature alignment, synthetic images, and self-training or self-distillation. Feature alignment finds transformations between source and target domains to reduce distribution shifts with adversarial training (He & Zhang, 2019; Chen et al., 2021; Ganin et al., 2016; Zhu et al., 2019). It can be helpful to extract common latent features from different domains. Progressive Domain Adaptation for Object Detection (Hsu et al., 2020) synthesized new dataset by using CycleGAN (Zhu et al., 2017) which enables to bridge domain gaps and Self-Adversarial Disentangling for Specific Domain Adaptation (Zhou et al., 2023) achieved 45.2 mAP on Cityscapes to Cityscapes foggy dataset using synthetic images. Gong et al. (2022) utilized transformers to focus on aligning features across backbone and decoder networks. However, combining multiple sources into a single dataset and performing single-source domain adaptation for feature alignment does not guarantee better performance compared to using the best individual source domain (Zhao et al., 2020).

Self-training uses a teacher model to predict pseudo labels on target domains to gradually understand domain shift (Caron et al., 2021; Pham et al., 2022; Cai et al., 2019; Chen et al., 2022; Cao et al., 2023). MIC (Hoyer et al., 2023) employed masked images on teacher-student model and MRT (Zhao et al., 2023) suggested modified masked based retraining approach on the teacher-student model. Kay et al. (2024) performed alignment and distillation to enforce invariance across domains to reduce feature discrepancies.

Finding common features from multiple domains is critical for DAOD. They have summarily demonstrated that the foreground features in latent space can be aligned using dimension reduction methods such as UMAP (McInnes et al., 2018) and t-SNE (Van der Maaten & Hinton, 2008). These studies did not address how to manage context bias when adapting across different domains. Instead, they proposed and validated their methods within the DAOD framework using accuracy metrics. Thus, we focus on analyzing the root causes of domain discrepancy in object detection both qualitatively and quantitatively.

3 Method

The following abbreviations are used throughout this paper to refer to the datasets and models in Table 1:

Table 1: Dataset and model abbreviations

Abbreviation	Meaning
CST	Cityscapes Train
CSV/ CFV / CRV	Cityscapes Validation / Foggy / Rainy
KST	KITTI Semantic Train
BG-20K	Background 20K Dataset
VKC / VKF / VKM / VKO / VKR / VKS	Virtual KITTI Clone / Fog / Morning / Virtual KITTI Overcast / Rain / Sunset
ALDI++	ResNet-50 FPN with ALDI++ best
Res	ResNet-50 FPN
Eff	EfficientNet-B0 FPN
Yo	YOLOv11

3.1 Models

We employed ResNet-50 ("Res") and EfficientNet-B0 ("Eff") as backbones for FPN models implemented in Detectron2, as well as YOLOv11 ("Yo") (Khanam & Hussain, 2024), an anchor-free detection model. "Res" represents a backbone dominantly used in different architectures. "Eff" was chosen for its lightweight architecture. Additionally, we included the state-of-the-art DAOD method ALDI++ with a ResNet-50 backbone to evaluate its effectiveness in mitigating FG–BG associations.

3.2 Datasets

We used multiple datasets for training and evaluation, including Cityscapes, KITTI Semantic, and various subsets of Virtual KITTI. Additionally, BG-20K, a collection of 20,000 images containing non-salient objects, was utilized to generate randomized background images. The Cityscapes and "KST" sets share 8 foreground and 11 background object categories. The Virtual KITTI subsets contain 3 foreground and 10 background object classes.

The dataset sizes are as follows:

- **Cityscapes**: 2,950 training images, 500 validation images, 1,500 foggy validation images, and 1,188 rainy validation images.
- **KITTI Semantic Train**: 200 images.
- **Virtual KITTI Semantic**: 2,126 images across 6 simulated weather conditions. This dataset is synthetic and based on object tracking in diverse environments.

3.3 Training and Tests

We trained "Res" and "Eff" on the "CST", "KST", and "VKC" datasets. "Yo" was trained with the same conditions using Ultralytics framework. ALDI++ was trained on ("KST","CSV") and ("VKC","VKF") as source and target domain pairs respectively. We used the pre-trained ALDI++ model provided by the official repository without additional training for Cityscapes. For training, we used a learning rate of 0.02 for "Res" with an input resolution of 1024×2048 for Cityscapes and 375×1242 for KITTI-related datasets. For "Eff", we used 1024×1024 resolution for Cityscapes and the same KITTI resolution, with a learning rate of 0.01. All models used identical data augmentation: resizing and cropping, color jitter, and horizontal flipping. Each model was trained with a batch size of 8. Training ran for approximately 100 epochs for ALDI++, "Res" and "Yo", and 200 epochs for "Eff". During evaluation, we used 1024×2048 images for Cityscapes-related datasets and 375×1242 for KITTI-related datasets on models except "Yo". "Yo" was trained and evaluated on 512x1024 and 320x1024 image resolution respectively. The best model checkpoint was selected based on the highest mean Average Precision with 0.5 IoU threshold (mAP@50) with respect to DAOD perspective. For example, "Res" achieved a mAP@50 of 67.758 on the Cityscapes validation set and 54.617 on the "CFV" at epoch 7799, and 67.597 and 57.131 at epoch 7999, respectively. We selected the model from epoch 7999 for subsequent experiments. For "KST", the model with the highest performance on "CFV" was chosen. For "VKC", we selected the model with the highest mAP@50 on "VKF", which represents the largest domain shift among the Virtual KITTI variants. All models except "Yo" were trained using the standard loss functions provided by Detectron2 and Ultralytics was utilized for "Yo". Training was conducted using an NVIDIA RTX A4500 GPU.

3.4 Q1 - Exp1. Class-wise Background Removal Experiments in Image Space

The initial experiment evaluates the effect of background variation by performing inference on foreground objects placed over random background images fixed across different domains. Foreground regions were preserved and composited with backgrounds randomly sampled from the BG-20K dataset. The same sequence of background images was applied consistently across domains, thereby reducing the learned association between foreground and background. To ensure statistical validity, the experiment was repeated 6 times using different random sequences of background selections. Algorithm 1 describes the experiment process. The results, summarized using the mAP@50 metric and standard deviation, are presented in Table 4. Figure

Figure 7: **Example of synthetic image.** We visualize the annotated bounding boxes for foreground objects.

7 shows an example of superimposing the foreground objects with a random background image from the BG-20K dataset.

Algorithm 1: Class-wise Background Removal Experiments in Image Space

Input:

- FG: Set of foreground object instances
- $BG - 20K$: Set of 20,000 random background images
- D: Set of target domains for inference

Output: Mean and standard deviation of mAP@50 across 6 repeated trials

foreach *domain* $d \in D$ **do**
    foreach *foreground object* $f \in FG$ **do**
        Randomly sample a fixed set of background images $BG_i \subset BG_{20K}$;
        Synthesize image $I_{f,d}$ by placing f onto a background from BG_i;
    Apply the trained model to all synthesized images $\{I_{f,d}\}$ for domain d;
    Measure detection performance using mAP@50;
Compute mean and standard deviation of mAP@50 across all 6 trials;

3.5 Q1-Exp2. Feature-wise Background Removal Experiments in Feature Space

The second experiment investigates feature-wise FG–BG associations by selectively suppressing specific background labels in the feature space during inference. Using ground-truth semantic annotations, a particular background class (e.g., "road") was removed in repeated inference runs. This was achieved by zeroing out activation values in the corresponding background regions at shallow network layers: $res2.2$ in "Res" and ALDI++, $backbone.bottom_up._blocks.0$ in "Eff", and $model.1$ in "Yo". Due to the hierarchical nature of deep learning models, this targeted suppression weakened the FG–BG associations for the removed background label, potentially affecting detection outcomes. To evaluate this effect, we measured the number of detections from unmodified models and compared them to detections after modification. The drop rate,

calculated for each FG–BG class pair, reflects the sensitivity of foreground object detection to the presence of specific background labels.

Definition of detection drop: A detection was counted as dropped under either of the following conditions:

1. The predicted class changed due to background label removal.
2. The loss of information we defined was greater than 1.0, which means IoU with ground truth significantly decrease (less than 0.1 IoU).
3. The prediction matched a different ground-truth object not originally considered a true positive.

Importantly, drop rates were computed only for true positive cases. To ensure statistical rigor, the process was repeated 6 times. Since drop rate distributions did not satisfy normality assumptions, we employed the Wilcoxon signed-rank test (Woolson, 2005) to assess statistical significance. Algorithm 2 describes the experiment process.

Algorithm 2: Feature-wise Background Removal Experiments in Feature Space

Input:

- FG: Set of foreground object instances
- BG: Set of background regions with semantic labels
- L_{remove}: Background label to be removed (e.g., "road")
- M: Deep learning models (e.g., "Res")

Output: Drop rate statistics for each FG–BG pair and Wilcoxon test results

```
for i = 1 to 6 do
    foreach model m ∈ M do
        foreach image x with semantic ground truth do
            Perform standard inference on x with model m, store number of detections D_std;
            Remove BG activated pixels in shallow feature maps corresponding to label
              L_remove (e.g., res2.2 for "Res");
            Perform modified inference on x, store number of detections D_mod;
            Compute detection drop ΔD = D_std − D_mod for each FG–BG pair;
    Store all ΔD values for statistical analysis;
Aggregate drop rates for each FG–BG pair across all trials;
Conduct Wilcoxon signed-rank test to assess significance of drop rate distributions;
```

3.5.1 Q2-Exp1. FG–BG Association with Respect to Activated Background Region

Using CAM masks with varying thresholds, we measured the mAP@50 drop rate to investigate the causal influence of activated background regions on object detection performance. Smooth-GradCAM++ (Omeiza et al., 2019) generates contextually meaningful instance masks by gradient backward on each object's score with 0.85 confidence threshold. The CAM masks were binarized by applying threshold values that decreased by 0.1 with each bin increase. The extent of the activated background region was controlled by the chosen threshold, while the masked foreground region remained fixed throughout the experiment, regardless of background variation. Through statistical analysis, the causality of association and accuracy were demonstrated. Algorithm 3 describes the experiment process. Figure 8 illustrates the contextual masks depending on different layers and bins. We define the hit ratio as the ratio of foreground and background pixels captured by CAM in the activation maps, normalized by the number of ground truth pixels from the instance masks. "FG mean" is the number of pixels hit by CAM and ground truth of instance mask. "BG mean" is activated background region over the number of foreground pixels. We averaged all instances hit ratio to compute "FG mean" and "BG mean" respectively. It indicates CAM captures properly contextual information for each instance. Definition of associated and non-associated is in Section 3.6.1.

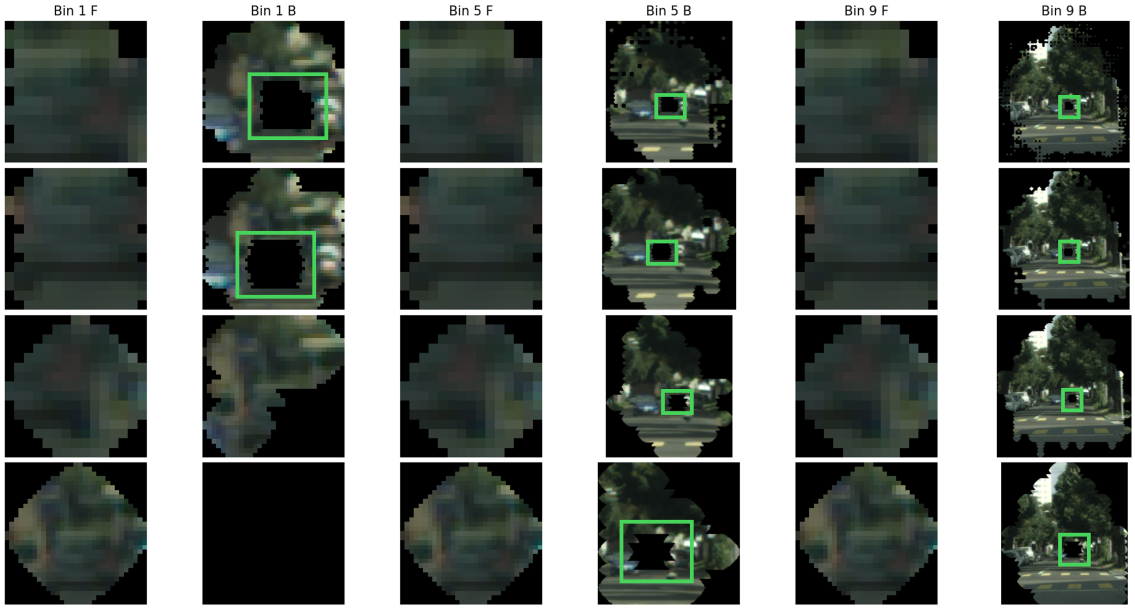


Figure 8: **"Res" foreground and background images using CAM and instance masks.** Each column corresponds to bin 1, 5, and 9 with foreground and background orders. Each row is different layers. The green bounding boxes highlight where the foreground object gets located. Foreground regions are maintained across layers. The blurriness of the images is due to them being scaled up.

Algorithm 3: Causality Analysis via Smooth-GradCAM++ Mask Thresholding

Input:

- FG: Set of foreground object instances with 0.85 confidence threshold
- x: Input image
- T: Set of CAM thresholds (e.g., maximum of activation value to $1e^{-9}$ in 0.1 decrements)

Output: Drop rates under different CAM thresholds

foreach *foreground instance* $f \in FG$ **do**
  Generate Smooth-GradCAM++ map H_f using prediction confidence of f from image x;
  foreach *threshold* $t \in T$ **do**
    Binarized CAM mask: $M_f^{(t)} = \mathbb{1}(H_f \geq t)$;
    (Note: foreground region remains fixed; only background area changes);
    Remove partial BG activations in shallow feature maps corresponding to the Masks (e.g., $res2.2$ for "Res") and compute detection result D_{mod};
    Compute drop rate: $\Delta D_t = D_{std} - D_{mod}$ where D_{std} is detection without CAM masking;
Aggregate drop rates and hit ratios across all instances;
Analyze drop rate trend across thresholds to infer FG–BG causality;

3.6 Q3-Exp1 and Exp2. Quantification the Causal Effect on DAOD

The results from our experiments confirmed the existence and causal effect of FG–BG associations. However, these findings did not directly quantify their influence on DAOD. To address this, we introduced a new metric, defined as domain association gradient (*Gradient*), which measures performance perturbation in response to FG–BG associations strength. For the association strength, we add the MMD between FG and BG and

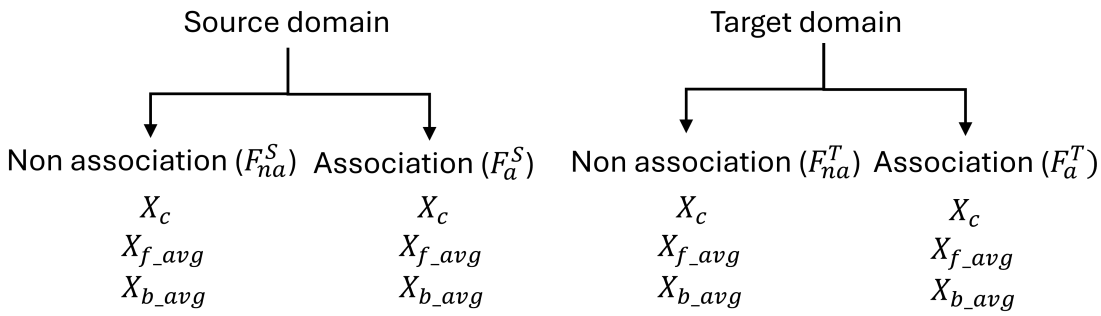


Figure 9: **Feature extraction graph.** We extracted features from all true positives with different FG–BG pairs. The process to generate each feature is described in Algorithm 4.

subtract the MMD in FG and BG. Figure 9 illustrates the feature extraction breakdown. For each domain (D), we can separate the features into association (F_a^D) and non-association features (F_{na}^D). These features can then be further broken down into CAM activated features (X_c), 2D pooled averaged FG features within X_c (X_{f_avg}) and 2D pooled averaged BG features within X_c (X_{b_avg}) using instance masks to make FG and BG separable. The detailed process to extract each feature from FG and BG of different domains (e.g. X_c, X_{f_avg}, X_{b_avg}) is described in Section 3.6.1.

The intuition is that if a strong FG–BG association is learned, the MMD between features that preserve this association should be small, due to shared contextual dependencies. Besides, drop rate should be large with a strong FG–BG associations. Consequently, the *Gradient* value that computed using these features will be larger than those derived from features lacking FG–BG associations. We calculated *Gradient* values using the drop rate for each FG–BG pair and analyzed their patterns via statistical testing using equations 3. Only FG–BG pairs with statistically significant drop rate differences identified in Experiment 3.5 were included in this analysis. Based on our hypothesis, we expected Gradient^S to exceed Gradient^T, as models trained on the source domain tend to capture stronger FG–BG associations. Each superscript (S and T) denotes the source and target domain, respectively.

$$\text{Gradient}^{\text{S}} = \frac{\text{Source domain drop rate}}{\text{f2b}^{\text{S}} + \text{b2f}^{\text{S}} - \text{f2f}^{\text{S}} - \text{b2b}^{\text{S}}}, \quad \text{Gradient}^{\text{T}} = \frac{\text{Target domain drop rate}}{\text{f2b}^{\text{T}} + \text{b2f}^{\text{T}} - \text{f2f}^{\text{T}} - \text{b2b}^{\text{T}}} \tag{3}$$

Equation 4 defines how various feature combinations, drawn from different contexts, are used in computing this metric. Specifically, cross-context feature comparisons, such as `f2b` and `b2f`, are expected to yield higher MMD values than within-context comparisons like `f2f` and `b2b`, since the latter reflect similar contextual structures.

$$\begin{aligned}
\text{f2f}^{\text{S}} &= \text{MMD}(X_{f_avg} \in F_{na}^S,\, X_{f_avg} \in F_a^S) & \text{f2f}^{\text{T}} &= \text{MMD}(X_{f_avg} \in F_{na}^T,\, X_{f_avg} \in F_a^T) \\
\text{f2b}^{\text{S}} &= \text{MMD}(X_{f_avg} \in F_{na}^S,\, X_{b_avg} \in F_a^S) & \text{f2b}^{\text{T}} &= \text{MMD}(X_{f_avg} \in F_{na}^T,\, X_{b_avg} \in F_a^T) \\
\text{b2f}^{\text{S}} &= \text{MMD}(X_{b_avg} \in F_{na}^S,\, X_{f_avg} \in F_a^S) & \text{b2f}^{\text{T}} &= \text{MMD}(X_{b_avg} \in F_{na}^T,\, X_{f_avg} \in F_a^T) \\
\text{b2b}^{\text{S}} &= \text{MMD}(X_{b_avg} \in F_{na}^S,\, X_{b_avg} \in F_a^S) & \text{b2b}^{\text{T}} &= \text{MMD}(X_{b_avg} \in F_{na}^T,\, X_{b_avg} \in F_a^T)
\end{aligned} \tag{4}$$

Furthermore, we conducted a focused analysis on `f2b` and `b2f` MMD values for both associated and non-associated features across domains using equation 5 and 6. We computed the sum of these values and compared them using pairwise T-test (O'Mahony, 2017) after Shapiro-Wilk test (Shapiro & Wilk, 1965).

$$\begin{aligned}
f2b_{asso} &= \text{MMD}(X_{f_avg} \in F_a^S,\, X_{b_avg} \in F_a^T) & f2b_{no-asso} &= \text{MMD}(X_{f_avg} \in F_{na}^S,\, X_{b_avg} \in F_{na}^T) \\
b2f_{asso} &= \text{MMD}(X_{b_avg} \in F_a^S,\, X_{f_avg} \in F_a^T) & b2f_{no-asso} &= \text{MMD}(X_{b_avg} \in F_{na}^S,\, X_{f_avg} \in F_{na}^T)
\end{aligned} \tag{5}$$

$$\text{sum}_{\text{a-a}} = \text{f2b}_{\text{asso}} + \text{b2f}_{\text{asso}} \qquad \text{sum}_{\text{na-na}} = \text{f2b}_{\text{no-asso}} + \text{b2f}_{\text{no-asso}} \tag{6}$$

A higher sum_{na-na} than sum_{a-a} indicates that FG–BG association persists across domains and may influence DAOD performance when the model unintentionally leverages such cross-domain dependencies during inference. In other words, FG–BG association was not impact on across DAOD then sum_{na-na} and sum_{a-a} should be no statistical significance different. We randomly sampled non-associated features to correspond to the number of associated features. We divided all datasets into two groups. The first group includes Cityscapes related datasets with "KST" and the second group contains only Virtual KITTI related dataset due to unmatched label configurations.

3.6.1 Associated Features and Non-associated Features Extraction

With drop rate experiment (Algorithm 3) as pre-processing, we defined associated features when the extracted features of each instance was not detected without background context. In other words, it indicates the features encode association between FG–BG. Conversely, non-associated features indicate that FG–BG association is not included in the extracted feature. We extracted features and clustered them per FG–BG pair. We used 4 different layers for each model to extract features in different scales. For "Res" and ALDI++, we used $res2.2, res3.3, res4.5, res5.2$ and for "Eff", we used $backbone.block.0, 1, 3$, and 5. From each domain, we saved associated features and non-associated features. Algorithm 4 describes steps to process foreground-related features and background-related features. Table 2 defines F_a^D and F_{na}^D meaning FG–BG associated features and FG–BG non-associated features from each domain D.

Algorithm 4: Feature Extraction from CAM and Ground Truth Instance Mask

Input:

- C: CAM mask from Algorithm 3.
- G: Ground truth instance binary mask
- A: Activation maps from different layers (e.g. $res2.2, res3.3, res4.5$, and $res5.2$ for "Res")

Output: X_c, X_{f_avg}, X_{b_avg}

Compute features X from CAM mask C;
    $X = A \cdot \mathbb{1}(C = 1)$;
Compute Normalized features X_c from X;
    X_c =Normalize(X);
Separate normalized features using ground truth mask G:
    X_{f_avg} = Adaptive pool 2d($X_c \cdot \mathbb{1}(G = 1)$);
    X_{b_avg} = Adaptive pool 2d($X_c \cdot \mathbb{1}(G = 0)$);

Table 2: **Definition of FG–BG associated features and non-associated features at each domain D.** "1" means BG removal is performed and detection fails for BG removal and Detection, respectively. "0" indicates BG removal is not applied and the success of detection.

		Detection	
		0	**1**
BG removal	**1**	When "road" is removed, detection fails → car feature with association (F_a^D)	Without "road", detection succeeds → car feature without association (F_{na}^D)
	0	False Negative. Unknown association impact on prediction.	True Positive. No association impact on prediction.

4 Experiments

Q1. Are FG–BG associations being inadvertently learned during the training process?

Model Evaluation

We evaluated the trained models using the mAP@50 metric. Table 3 summarizes the evaluation results. Among models trained on the Cityscapes dataset, ALDI++ outperformed others on Cityscapes-related datasets and "KST", likely due to longer training epochs on target domain and the use of multiple domain datasets via DAOD algorithms. Although "Yo" achieved strong performance on "CST", its domain adaptation capability was weaker than that of the baseline "Res". "Eff" consistently showed the lowest performance on both Cityscapes and "KST" but outperformed "Yo" on "KST". For models trained on "KST", ALDI++ outperformed other models, while "Eff" and "Yo" showed significantly worse results. The "KST" dataset, with only 200 images, introduced a domain shift that limited generalization to larger datasets. Consequently, "Res" and "Yo" also exhibited weakened performance due to insufficient data. With VKC-trained models, "Res" and "Yo" demonstrated reasonable domain generalization compared with "Eff". However, on the "VKF" validation set, "Yo"'s performance dropped significantly relative to "Res", despite its overall strong results on other datasets. "Eff" also experienced a notable performance decline on "VKF". ALDI++ demonstrated the effectiveness of domain generalization methods.

Table 3: **Model evaluation across different training and validation sets.** "-" is not measurable because of unmatched class category.

Dataset	Cityscapes Train				KST Train				VKC Train			
	Res	Eff	Yo	ALDI++	Res	Eff	Yo	ALDI++	Res	Eff	Yo	ALDI++
CST	79.14	41.12	**88.26**	87.97	-	-	-	-	-	-	-	-
CSV	67.59	42.90	59.56	**70.08**	43.23	2.74	21.69	**51.09**	-	-	-	-
CFV	57.13	20.58	44.49	**67.45**	35.48	0.61	12.53	**43.62**	-	-	-	-
CRV	58.65	23.18	48.77	**69.78**	37.42	0.86	17.57	**47.82**	-	-	-	-
KST	46.25	28.92	23.91	**47.96**	86.17	10.42	21.53	**92.44**	-	-	-	-
VKC	-	-	-	-	-	-	-	-	81.67	50.09	**85.99**	81.96
VKF	-	-	-	-	-	-	-	-	61.14	5.80	34.27	**72.60**
VKM	-	-	-	-	-	-	-	-	79.72	29.52	79.55	**80.27**
VKO	-	-	-	-	-	-	-	-	75.14	30.41	**81.93**	78.58
VKR	-	-	-	-	-	-	-	-	71.66	25.53	75.93	**78.36**
VKS	-	-	-	-	-	-	-	-	76.02	26.18	**78.65**	77.59

Q1-Exp1. Class-wise Background Removal Experiments in Image Space

This experiment evaluated the role of FG–BG associations by replacing background regions with non-salient object images while preserving foreground objects. We measured mAP@50 over six repetitions using randomly generated images. Table 4 presents the mean and standard deviation across the six evaluation runs. Models trained on Cityscapes showed substantial performance drops compared to their original evaluation. Notably, ALDI++ exhibited greater degradation than "Res" on "KST", suggesting that ALDI++ strongly relies on FG–BG associations learned from both source and target domains. Similarly, models trained on "KST" and "VKC" also experienced considerable performance declines, indicating that FG–BG associations were learned during training and utilized during inference.

Q1-Exp2. Feature-wise Background Removal Experiments in Feature Space

In addition to background perturbation in image space, we performed background removal in feature space in Table 5. Specifically, we zeroed out activated background regions in the shallow layers of each model architecture, preventing background information from propagating to deeper layers. This effectively disables the FG–BG association during inference. The Cityscapes-related datasets and "KST" contain 88 distinct FG–BG combinations, while the Virtual KITTI datasets include 30 combinations. We report only the combinations that resulted in a statistically significant performance drop of at least 8%. Note that additional combinations exhibited smaller drops and are not included in the table. Figure 10 illustrates the example

Table 4: **Mean ± standard deviation of mAP@50 across synthetic datasets and models trained on Cityscapes, "KST", and "VKC".** The bolded values are the highest mAP@50 for each train dataset-model pair.

Cityscapes Trained

Dataset + BG	Res	Eff	Yo	ALDI++
CSV	44.5 ± 6.1	10.02 ± 6.4	32.78 ± 0.6	**47.96** ± 6.3
CFV	34.6 ± 10.6	8.99 ± 4.5	15.70 ± 0.3	**43.17** ± 9.8
CRV	29.1 ± 4.4	14.24 ± 8.4	17.37 ± 0.3	**38.18** ± 5.1
KST	**41.4** ± 4.4	25.33 ± 6.4	19.92 ± 1.4	33.98 ± 6.7

KST Trained

Dataset + BG	Res	Eff	Yo	ALDI++
CSV	19.2 ± 5.2	0.16 ± 0.1	21.34 ± 0.3	**23.01** ± 0.7
CFV	17.0 ± 5.8	0.22 ± 0.1	10.35 ± 0.1	**24.55** ± 0.1
CRV	13.6 ± 3.4	2.25 ± 2.1	14.29 ± 0.1	**35.29** ± 0.7

VKC Trained

Dataset + BG	Res	Eff	Yo	ALDI++
VKF	38.7 ± 10.7	1.17 ± 0.5	14.61 ± 0.5	**51.04** ± 0.5
VKM	61.8 ± 5.1	11.44 ± 6.4	35.26 ± 0.6	**64.17** ± 0.3
VKO	60.0 ± 4.7	10.70 ± 5.5	33.66 ± 0.4	**64.44** ± 0.5
VKR	58.4 ± 3.5	7.99 ± 3.5	27.56 ± 0.4	**59.81** ± 0.4
VKS	59.8 ± 6.5	12.92 ± 4.6	35.93 ± 0.8	**64.26** ± 0.2

of significant performance drop on "CST" with "Res" model. It indicates that the models learned notable FG–BG associations during training, which can cause performance degradation.

Table 5: **The number of FG–BG pairs statistically significant different across models**. Only FG–BG pairs more than 8% drop rate are denoted. The bold values indicate stronger FG–BG associations for each model across the respective datasets.

	Res	**ALDI++**	**Eff**	**Yo**
CST	14/88	**18/88**	15/88	12/88
CSV	7/88	**17/88**	7/88	3/88
CFV	13/88	**20/88**	11/88	7/88
CRV	15/88	**21/88**	12/88	6/88
KST	2/88	2/88	**4/88**	2/88
VKC	8/30	7/30	**9/30**	4/30
VKF	**7/30**	**7/30**	4/30	2/30
VKM	**9/30**	7/30	7/30	3/30
VKO	**9/30**	7/30	8/30	3/30
VKR	7/30	**9/30**	**9/30**	6/30
VKS	8/30	8/30	**9/30**	4/30

Q2. Is there a causal relationship between FG–BG associations and object detection?

In the previous section, we validated the existence of FG–BG associations. To further investigate this phenomenon, we computed CAM masks for each object and analyzed background-region-based associations. To explore the causal relationship, we applied do-calculus using CAM-derived masks and ground truth instance masks. We also note that "Yo" contains non-differentiable non-maximum suppression (NMS) which prevents us from using gradient based CAM to capture specific objects' contextual masks. Thus, we are not able to answer Q2 and Q3 experiments using this particular model.

Q2-Exp1. FG–BG Association with Respect to Activated Background Regions

With do-calculus, we computed the mean of drop rate of all classes per bin. While foreground region were maintained regardless of bin (see Fig. 8 and Table 6), the drop rate significantly increased with bin 1 which has small amount of activated background (see Table 7). While enlarging the background regions, the drop rate converged toward 0 which means the objects were detected correctly. Through the experiment, we confirmed a causal relationship between FG–BG associations and outcomes, as the accuracy of foreground objects changed notably across bins, particularly across from bin 1 to bin 5.

Table 6: **Definition of FG–BG associated features and non-associated features for each domain D.** The FG mean of 1.0 indicates that all foreground pixels are captured by the CAM. The BG mean represents the ratio of captured background pixels to the total number of foreground pixels, reflecting the extent of background activation relative to the foreground.

Layer	Hit ratio			
	Associated		Non-associated	
	FG mean	BG mean	FG mean	BG mean
$res2.2$	1.0	14.81	1.0	20.04
$res3.3$	1.0	14.33	1.0	18.79
$res4.5$	1.0	12.66	1.0	27.2
$res5.2$	1.0	21.0	1.0	12.33

Table 7: **Drop rate per bin.** 5 numbers in each cell are bin 1, 3, 5, 7, and 9 ("B"). The lower means fewer or no drop rates were measured. The bolded values highlight the significant performance drop.

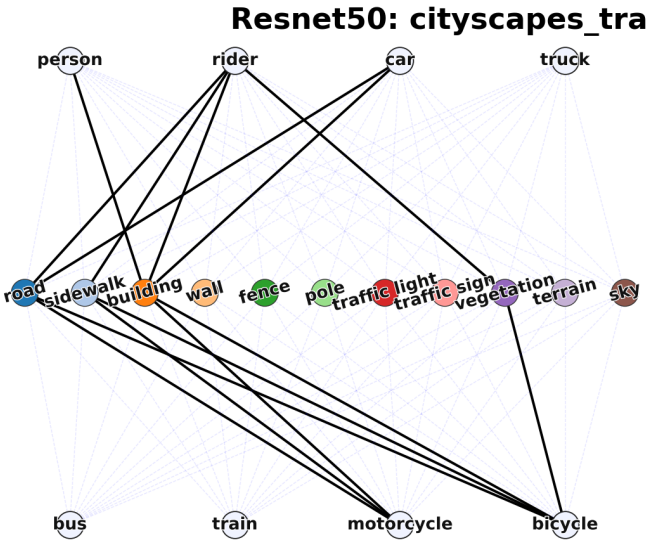
	Res					**ALDI++**					**Eff**				
	B1	B3	B5	B7	B9	B1	B3	B5	B7	B9	B1	B3	B5	B7	B9
CST	**0.65**	0.14	0.02	0.01	0.00	**0.66**	0.18	0.02	0.00	0.00	**0.74**	0.31	0.25	0.26	0.18
CSV	**0.62**	0.10	0.02	0.00	0.00	**0.64**	0.11	0.01	0.00	0.00	**0.71**	0.18	0.14	0.13	0.10
CFV	**0.67**	0.12	0.03	0.00	0.00	**0.68**	0.10	0.02	0.01	0.00	**0.79**	0.19	0.09	0.11	0.11
CRV	**0.67**	0.27	0.07	0.02	0.01	**0.64**	0.20	0.05	0.01	0.01	**0.65**	0.12	0.06	0.07	0.09
KST	**0.84**	0.53	0.22	0.01	0.00	**0.70**	0.31	0.00	0.02	0.02	**0.82**	0.71	0.36	0.69	0.68
VKC	**0.57**	0.08	0.02	0.00	0.00	**0.33**	0.03	0.01	0.00	0.00	**0.71**	0.19	0.07	0.04	0.03
VKF	**0.75**	0.37	0.12	0.04	0.01	**0.42**	0.05	0.02	0.01	0.00	**0.92**	0.15	0.04	0.05	0.02
VKM	**0.63**	0.10	0.03	0.00	0.00	**0.42**	0.05	0.02	0.01	0.00	**0.77**	0.25	0.11	0.06	0.05
VKO	**0.56**	0.11	0.03	0.00	0.00	**0.33**	0.05	0.02	0.00	0.00	**0.76**	0.23	0.05	0.03	0.05
VKR	**0.62**	0.13	0.04	0.01	0.00	**0.33**	0.05	0.02	0.00	0.00	**0.80**	0.26	0.11	0.08	0.07
VKS	**0.64**	0.16	0.07	0.02	0.01	**0.34**	0.05	0.02	0.00	0.00	**0.77**	0.21	0.06	0.03	0.03

Q3. Is FG–BG Associations Impact on DAOD and How to Quantify the Effect?

We computed *Gradient* values and analyzed the MMDs between associated and non-associated features across different domains. This allowed us to quantify the impact of FG–BG associations on domain shifts.

Q3-Exp1. Domain Association Gradient

To validate our hypothesis that $Gradient^S$ should be lower than $Gradient^T$ due to learned FG–BG associations, we categorized the comparisons into three cases: **(1) $Gradient^S$ significantly lower than $Gradient^T$, (2) $Gradient^S$ significantly higher than $Gradient^T$**, and **(3) no statistically significant difference.** Table 8 summarizes the results. Overall, the findings support our hypothesis. However, ALDI++ showed opposite results on the CST-CSV pair, possibly due to DAOD training strategy using target domain information (CFV). In the Virtual KITTI-related datasets, particularly for "Eff", some results contradicted expectations. This may be attributed to strong spatial and temporal correlations inherent in the dataset, which is derived from object-tracking video sequences, or to biases introduced by a small number



BG	FG	Drop rate
building	person	0.22
road	rider	0.21
sidewalk	bicycle	0.21
building	bicycle	0.20
building	rider	0.19
building	motorcycle	0.18
road	bicycle	0.18
road	motorcycle	0.16
sidewalk	motorcycle	0.16
sidewalk	rider	0.14
vegetation	rider	0.12
building	car	0.11
road	car	0.10
vegetation	bicycle	0.09

Figure 10: **Feature-wise drop rates of "Res" on "CST".** Bold edges indicate statistically significant difference on the FG–BG pair. The table on the right in the figure illustrates the drop rate in decreasing order.

of detections and drops. These results elaborates quantification of class-wise causal effects in DAOD. Figure 11 presents the results of the *Gradient* comparisons as box plots to present the visual comparison.

Table 8: *Gradient* **comparison across different domains.** Each number's denominator is the number of FG–BG association in common across two domains. Cases 1, 2 and 3 are given as C1, C2 and C3. The bolded values represent the dominant case for each dataset-model pair.

	Res			**ALDI++**			**Eff**		
	C1	C2	C3	C1	C2	C3	C1	C2	C3
CST - CSV	**4/5**	0/5	1/5	4/10	**6/10**	0/10	**3/7**	**3/7**	1/7
CST - CFV	**11/11**	0/11	0/11	**15/15**	0/15	0/15	**7/11**	2/11	2/11
CST - CRV	**10/11**	1/11	0/11	**8/10**	2/10	0/10	**7/10**	3/10	0/10
CST - KST	**1/1**	0/1	0/1	**1/1**	0/1	0/1	**2/3**	1/3	0/3
VKC - VKF	**5/5**	0/5	0/5	**4/4**	0/4	0/4	**1/2**	**1/2**	0/2
VKC - VKM	**5/7**	2/7	0/7	**4/4**	0/4	0/4	1/5	**4/5**	0/5
VKC - VKO	**4/7**	3/7	0/7	**4/4**	0/4	0/4	1/5	**4/5**	0/5
VKC - VKR	**6/6**	0/6	0/6	**5/5**	0/5	0/5	1/5	**4/5**	0/5
VKC - VKS	2/6	**4/6**	0/6	**4/5**	1/5	0/5	1/5	**4/5**	0/5

Q3-Exp2. Associated and Non-associated Features

To understand the class- and feature-wise impact across different domains, we compared the gap between the MMD of associated features and that of non-associated features. Similar to previous *Gradient* analysis, we categorized the comparisons into three cases: **(1) The summation of MMD of `f2b` and `b2f` in associated features significantly lower than that of non-associated features (2) The summation of MMDs of `f2b` and `b2f` for associated features significantly higher than that of non-associated features**, and **(3) no statistically significant difference.** Overall, associated features sharing the same FG–BG association across domains exhibited lower MMD than non-associated features, indicating stronger FG–BG association consistency in cross-domain. However, "Eff" showed reversed outcomes on the CST-CRV and CST-KST pairs. This may be due to a limited number of detections, resulting in insufficient feature representations or overall poor model performance. Table 12 presents the results, while Figure 12 depicts

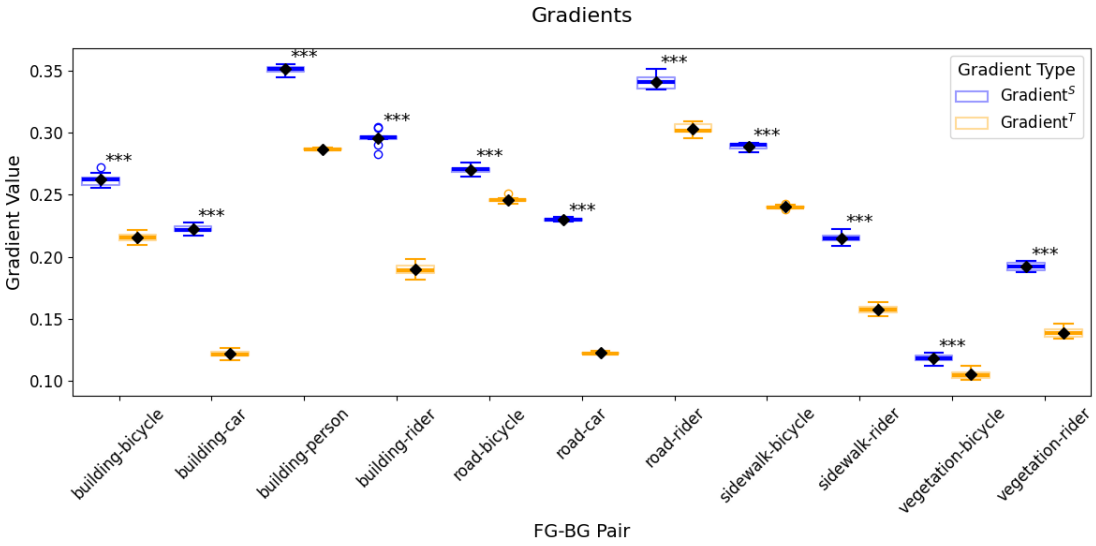


Figure 11: ***Gradient^S* and *Gradient^T* comparison of "Res" model with "CST" and "CFV"**. The number of "*" indicates the level of statistical significance of a p-value.

the results of associated and non-associated feature comparison, with statistical analysis annotated in the box plots.

Table 9: **Associated and non-associated features comparison across different domains.** Each number's denominator is the number of FG–BG association in common across two domains. Cases 1, 2 and 3 are given as C1, C2 and C3. "-" is not measurable statistically. The bolded values represent the dominant case for each dataset-model pair.

	Res			**ALDI++**			**Eff**		
	C1	C2	C3	C1	C2	C3	C1	C2	C3
CST - CSV	**5/5**	0/5	0/5	**9/10**	0/10	1/10	**3/7**	2/7	2/7
CST - CFV	**11/11**	0/11	0/11	**9/15**	6/15	0/15	**6/11**	5/11	0/11
CST - CRV	**8/11**	3/11	0/11	**7/10**	3/10	0/10	2/10	**6/10**	2/10
CST - KST	**1/1**	0/1	0/1	**1/1**	0/1	0/1	0/3	**2/3**	1/3
VKC - VKF	**2/5**	1/5	**2/5**	**4/4**	0/4	0/4	**2/2**	0/2	0/2
VKC - VKM	**4/7**	2/7	1/7	**4/4**	0/4	0/4	**3/5**	1/5	1/5
VKC - VKO	**5/7**	2/7	0/7	**4/4**	0/4	0/4	**4/5**	1/5	1/5
VKC - VKR	**4/6**	1/6	1/6	**5/5**	0/5	0/5	**4/5**	1/5	0/5
VKC - VKS	**5/6**	1/6	0/6	**5/5**	0/5	0/5	**3/5**	1/5	1/5

5 Discussion and Conclusion

In this work, we present a comprehensive empirical and theoretical investigation into the role of context bias in domain-adaptive object detection (DAOD). While context bias has previously been studied in classification and segmentation tasks, our work is the first to formally identify, quantify, and causally analyze this phenomenon within the context of DAOD.

Our findings show that modern object detection models frequently rely on foreground–background (FG–BG) associations that often do not generalize well across domains. Through systematic background masking and feature-level perturbations, we demonstrate that removing or altering background information can lead to substantial drops in detection performance even when the foreground remains intact. These effects are

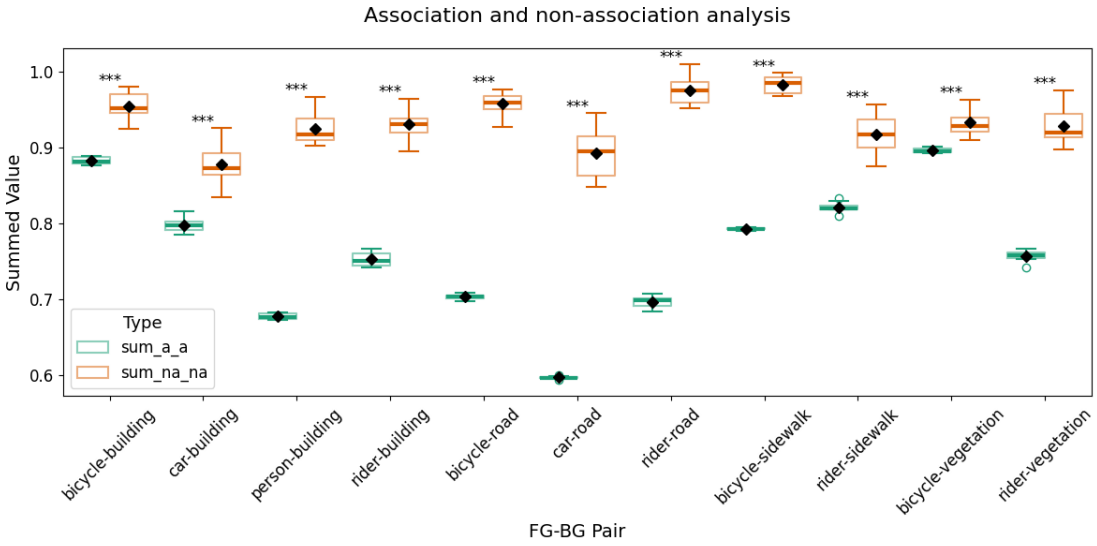


Figure 12: **MMD of associated features and non-associated comparison of "Res" model with "CST" and "CFV"**. The number of "*" indicates the level of statistical significance of a p-value.

consistent across various model architectures and domain pairs, including ALDI++, a state-of-the-art DAOD model on the Cityscapes dataset.

Furthermore, we present that FG–BG associations are not only empirically observable but also causally linked to detection outcomes. Using a combination of do-calculus, Smooth-GradCAM++, and layer-wise feature analysis, we construct and validate a causal model that quantifies the influence of context bias. We introduce a novel domain association gradient metric and find that domain shifts exacerbate performance disparities when models rely on FG–BG associations.

In Appendix (Section A), we provide additional clarifications of our experimental results. Specifically, we include (i) a fixed color background replacement experiment analyzed from a counterfactual perspective, (ii) ViTDet (Li et al., 2022a) attention maps and drop rates under background replacement in image space, and (iii) *Gradient* and association and no association analysis using a Radial Basis Function (RBF) kernel. Together, these supplementary analyses help address several open questions and limitations of our main experiments.

Limitations

Despite the strength of our analysis, we acknowledge that extracting foreground and background features separately across large datasets is computationally expensive. This limits the scalability of some of the proposed methods. Additionally, our study does not explore transformer-based architectures, which may inherently reduce FG–BG dependency due to their global receptive fields; however, interpreting FG–BG associations in such architectures remains ambiguous. Some outliers might be derived from imbalanced foreground objects of each dataset. For example, "Car" and "Person" are dominant but other foreground objects are rare. There are also certain neural network architectures such as in "Yo" that prevent us from computing the CAM masks and deriving the FG and BG activation features. Thus, while our method is robust, it is only applicable for architectures where we can run gradient-based CAM methods.

Future Work

Importantly, our results suggest that current DAOD methods may unintentionally reintroduce context bias from the target domain. This highlights a new dimension of the domain adaptation problem and points to the need for bias-aware adaptation strategies that explicitly consider FG–BG association. Our focus

through this work is to show that robust DAOD requires dissociating FG from BG in order to maintain the performance level even in the divergent domains. The FG–BG association may act as a spurious or beneficial factor, depending on the stage of the pipeline. During the feature extraction process (done mostly by the backbone network), if the features get entangled, that can lead to large performance drops or complete non-identification of known FG. However, once the FG has been identified, it helps to provide BG cues to understand whether such as association is viable or not. For example, a car being in sky can account for such an example where we want to identify the car and then flag the association between the car and sky as unusual.

Since the entanglement occurs during the feature aggregation process, we believe it is due to the pooling mechanism which aggregates features without any thought of object boundaries. To mitigate the entanglement of FG and BG features, we believe pooling steps may be redesigned for DAOD. Therefore, building a mask for object boundary is crucial before the feature extraction process. However, that leads to a chicken-and-egg problem. We are actively exploring several potential remedies, including mask-based pooling to emphasize foreground regions, EM-based mask generation, and graph-matching techniques. We believe these approaches can mitigate the identified performance drops, and we plan to extend our study in this direction in future work.

We believe our work opens a novel research direction in DAOD by emphasizing the need to go beyond feature alignment and focus on understanding and mitigating causal biases introduced by background context. Future work may explore efficient integration of bias-awareness into end-to-end training pipelines and investigate connections between FG–BG associations and broader issues such as spurious correlations and fairness in DAOD.

References

Ajay Agrawal, Joshua S Gans, and Avi Goldfarb. Artificial intelligence: the ambiguous labor market impact of automating prediction. *Journal of Economic Perspectives*, 33(2):31–50, 2019.

Inigo Alonso, Alberto Sabater, David Ferstl, Luis Montesano, and Ana C Murillo. Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8219–8228, 2021.

Ananthu Aniraj, Cassio F Dantas, Dino Ienco, and Diego Marcos. Masking strategies for background bias removal in computer vision models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4397–4405, 2023.

Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in neural information processing systems*, 32, 2019.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 2006. ISBN 9780387310732.

Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

Qi Cai, Yingwei Pan, Chong-Wah Ngo, Xinmei Tian, Lingyu Duan, and Ting Yao. Exploring object relation in mean teacher for cross-domain detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11457–11466, 2019.

Shengcao Cao, Dhiraj Joshi, Liang-Yan Gui, and Yu-Xiong Wang. Contrastive mean teacher for domain adaptive object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 23839–23848, 2023.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Meilin Chen, Weijie Chen, Shicai Yang, Jie Song, Xinchao Wang, Lei Zhang, Yunfeng Yan, Donglian Qi, Yueting Zhuang, Di Xie, et al. Learning domain adaptive object detection with probabilistic teacher. *arXiv preprint arXiv:2206.06293*, 2022.

Yuhua Chen, Haoran Wang, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Scale-aware domain adaptive faster r-cnn. *International Journal of Computer Vision*, 129(7):2223–2243, 2021.

Myung Jin Choi, Antonio Torralba, and Alan S Willsky. Context models and out-of-context objects. *Pattern Recognition Letters*, 33(7):853–862, 2012.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4091–4101, 2021.

Santosh K Divvala, Derek Hoiem, James H Hays, Alexei A Efros, and Martial Hebert. An empirical study of context in object detection. In *2009 IEEE Conference on computer vision and Pattern Recognition*, pp. 1271–1278. IEEE, 2009.

Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pp. 1–16. PMLR, 2017.

Maximilian Dreyer, Reduan Achtibat, Thomas Wiegand, Wojciech Samek, and Sebastian Lapuschkin. Revealing hidden context bias in segmentation and object detection through concept-specific explanations. 2023 ieee. In *CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3829–3839, 2023.

Serdar Erisen. Sernet-former: semantic segmentation by efficient residual network with attention-boosting gates and attention-fusion networks. *arXiv preprint arXiv:2401.15741*, 2024.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

Kaixiong Gong, Shuang Li, Shugang Li, Rui Zhang, Chi Harold Liu, and Qiang Chen. Improving transferability for domain adaptive detection transformers. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 1543–1551, 2022.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, Cambridge, MA, 2016. ISBN 9780262035613.

Zhenwei He and Lei Zhang. Multi-adversarial faster-rcnn for unrestricted object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6668–6677, 2019.

Lukas Hoyer, Dengxin Dai, Haoran Wang, and Luc Van Gool. Mic: Masked image consistency for context-enhanced domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11721–11732, 2023.

Han-Kai Hsu, Chun-Han Yao, Yi-Hsuan Tsai, Wei-Chih Hung, Hung-Yu Tseng, Maneesh Singh, and Ming-Hsuan Yang. Progressive domain adaptation for object detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 749–757, 2020.

Ling Huang, Lijuan Wang, Wangming Shen, Mengsha Li, Shiyu Wang, Xiaotong Wang, Leslie G Ungerleider, and Xilin Zhang. A source for awareness-dependent figure–ground segregation in human prefrontal cortex. *Proceedings of the National Academy of Sciences*, 117(48):30836–30847, 2020.

Menglin Jia, Mengyun Shi, Mikhail Sirotenko, Yin Cui, Claire Cardie, Bharath Hariharan, Hartwig Adam, and Serge Belongie. Fashionpedia: Ontology, segmentation, and an attribute localization dataset. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pp. 316–332. Springer, 2020.

Tarun Kalluri, Wangdong Xu, and Manmohan Chandraker. Geonet: Benchmarking unsupervised adaptation across geographies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15368–15379, 2023.

Justin Kay, Timm Haucke, Suzanne Stathatos, Siqi Deng, Erik Young, Pietro Perona, Sara Beery, and Grant Van Horn. Align and distill: Unifying and improving domain adaptive object detection. *arXiv preprint arXiv:2403.12029*, 2024.

Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*, 2024.

Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A Efros, and Antonio Torralba. Undoing the damage of dataset bias. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part I 12*, pp. 158–171. Springer, 2012.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pp. 5637–5664. PMLR, 2021.

Wuyang Li, Jie Liu, Bo Han, and Yixuan Yuan. Adjustment and alignment for unbiased open set domain adaptation (supplementary material).

Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *European conference on computer vision*, pp. 280–296. Springer, 2022a.

Yu-Jhe Li, Xiaoliang Dai, Chih-Yao Ma, Yen-Cheng Liu, Kan Chen, Bichen Wu, Zijian He, Kris Kitani, and Peter Vajda. Cross-domain adaptive teacher for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7581–7590, 2022b.

Junhui Liang, Ying Liu, and Vladimir Vlassov. The impact of background removal on performance of neural networks for fashion image classification and segmentation. In *2023 Congress in Computer Science, Computer Engineering, & Applied Computing (CSCE)*, pp. 1960–1968. IEEE, 2023.

David J.C. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, Cambridge, 2003. ISBN 9780521642989.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Mohammed Bany Muhammad and Mohammed Yeasin. Eigen-cam: Class activation map using principal components. In *2020 international joint conference on neural networks (IJCNN)*, pp. 1–7. IEEE, 2020.

Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, Cambridge, MA, 2012. ISBN 9780262018029.

Aude Oliva and Antonio Torralba. The role of context in object recognition. *Trends in cognitive sciences*, 11(12):520–527, 2007.

Michael O'Mahony. *Sensory evaluation of food: statistical methods and procedures*. Routledge, 2017.

Daniel Omeiza, Skyler Speakman, Celia Cintas, and Komminist Weldermariam. Smooth grad-cam++: An enhanced inference level visualization technique for deep convolutional neural network models. *arXiv preprint arXiv:1908.01224*, 2019.

Paolo Papale, Andrea Leo, Luca Cecchetti, Giacomo Handjaras, Kendrick N Kay, Pietro Pietrini, and Emiliano Ricciardi. Foreground-background segmentation revealed during natural image viewing. *eneuro*, 5(3), 2018.

Minh Pham, Minsu Cho, Ameya Joshi, and Chinmay Hegde. Revisiting self-distillation. *arXiv preprint arXiv:2206.08491*, 2022.

Jasper Poort, Matthew W Self, Bram Van Vugt, Hemi Malkki, and Pieter R Roelfsema. Texture segregation causes early figure enhancement and later ground suppression in areas v1 and v4 of visual cortex. *Cerebral cortex*, 26(10):3964–3976, 2016.

Harish Guruprasad Ramaswamy et al. Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In *proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 983–991, 2020.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

Amir Rosenfeld, Richard Zemel, and John K Tsotsos. The elephant in the room. *arXiv preprint arXiv:1808.03305*, 2018.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5): 612–634, 2021.

Samuel Sanford Shapiro and Martin B Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4):591–611, 1965.

Rakshith Shetty, Bernt Schiele, and Mario Fritz. Not using the car to see the sidewalk–quantifying and controlling the effects of context in classification and segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8218–8226, 2019.

Moeko Takagi, Edgar Simo-Serra, Satoshi Iizuka, and Hiroshi Ishikawa. What makes a style: Experimental analysis of fashion prediction. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 2247–2253, 2017.

Luke Taylor and Geoff Nitschke. Improving deep learning with generic data augmentation. In *2018 IEEE symposium series on computational intelligence (SSCI)*, pp. 1542–1547. IEEE, 2018.

Antonio Torralba. Contextual priming for object detection. *International journal of computer vision*, 53: 169–191, 2003.

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

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Wenfu Wang, Yongjian Fu, Zhijie Pan, Xi Li, and Yueting Zhuang. Real-time driving scene semantic segmentation. *IEEE Access*, 8:36776–36788, 2020.

Robert F Woolson. Wilcoxon signed-rank test. *Encyclopedia of biostatistics*, 8, 2005.

Pingyu Wu, Wei Zhai, and Yang Cao. Background activation suppression for weakly supervised object localization. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14228–14237. IEEE, 2022.

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

Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. *arXiv preprint arXiv:2006.09994*, 2020.

Wei Zhai, Pingyu Wu, Kai Zhu, Yang Cao, Feng Wu, and Zheng-Jun Zha. Background activation suppression for weakly supervised object localization and semantic segmentation. *International Journal of Computer Vision*, 132(3):750–775, 2024.

Baoqiang Zhang, Saisai Hu, Tingkang Zhang, Min Hai, Yongchun Wang, Ya Li, and Yonghui Wang. Different patterns of foreground and background processing contribute to texture segregation in humans: an electrophysiological study. *PeerJ*, 11:e16139, 2023.

Jianguo Zhang, Marcin Marszałek, Svetlana Lazebnik, and Cordelia Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *International journal of computer vision*, 73:213–238, 2007.

Kexuan Zhang, Qiyu Sun, Chaoqiang Zhao, and Yang Tang. Causal reasoning in typical computer vision tasks. *Science China Technological Sciences*, 67(1):105–120, 2024.

Sicheng Zhao, Bo Li, Pengfei Xu, and Kurt Keutzer. Multi-source domain adaptation in the deep learning era: A systematic survey. *arXiv preprint arXiv:2002.12169*, 2020.

Zijing Zhao, Sitong Wei, Qingchao Chen, Dehui Li, Yifan Yang, Yuxin Peng, and Yang Liu. Masked retraining teacher-student framework for domain adaptive object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19039–19049, 2023.

Qianyu Zhou, Qiqi Gu, Jiangmiao Pang, Xuequan Lu, and Lizhuang Ma. Self-adversarial disentangling for specific domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7): 8954–8968, 2023.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

Lanyun Zhu, Tianrun Chen, Jianxiong Yin, Simon See, and Jun Liu. Addressing background context bias in few-shot segmentation through iterative modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3370–3379, 2024.

Xinge Zhu, Jiangmiao Pang, Ceyuan Yang, Jianping Shi, and Dahua Lin. Adapting object detectors via selective cross-domain alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 687–696, 2019.

Zhuotun Zhu, Lingxi Xie, and Alan L Yuille. Object recognition with and without objects. *arXiv preprint arXiv:1611.06596*, 2016.

A Appendix

A.1 Additional BG replacement experiments with simple changes

We conducted fixed BG replacement to ensure a diverse and unbiased distribution of conditions. Specifically, we experimented with different solid background colors and noise-added backgrounds. While we did not perform a detailed statistical analysis, we observed noticeable accuracy variations across these settings. Our experimental design was therefore structured to progress from simple perturbations to more controlled and feasible analyses to make more generalized and practical results. Figure 13 illustrates the outcome.

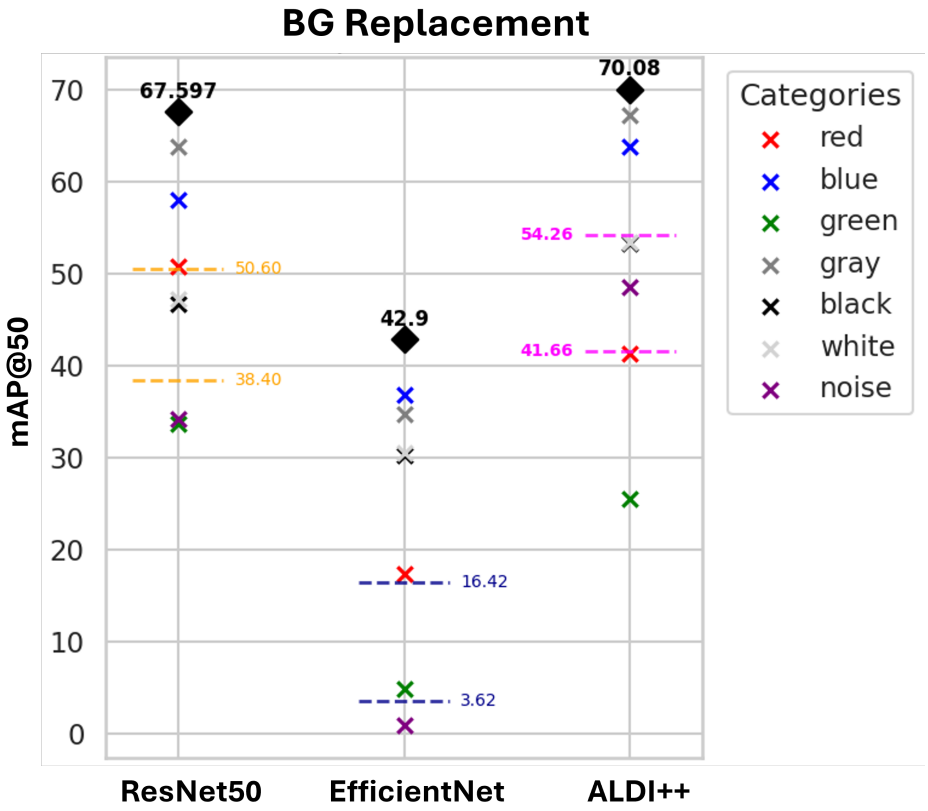


Figure 13: The black diamonds are mAP@50 of each model on the Cityscapes train set. The different categories ("x") mean types of background (e.g., changed in red). The dashed lines represent the range of random BG experiments we performed. Each number on the figure is mAP@50.

A.2 ViTDeT experiments

Transformer-based detectors can strengthen our experiment, but it is difficult to apply the same process due to their architecture. They reorganize spatial information through a pose embedding step, which incurs obstacles to track spatial information to remove background and capture foreground-background related features. So we performed simple experiments using VitDet (Li et al., 2022a). Gradient-based CAM is not suitable, we applied Ablation CAM (Ramaswamy et al., 2020) like method and measured the attention map for each object. For example, we zeroed out for each token for a "car" and computed the weight of the token as the difference Intersection over Union (IoU) between the original bounding box and the new bounding box. If a current target car is not detected by removing the activation of a token, the weight of the token will be 1.0. We visualize an attention map based on the computed weights and the normalized activation map (see Figure 14).

Figure 14: **Left panel**: An attention map (white) is overlaid on the RGB input, with the green box marking the target object. **Right panel**: Attention map with thresholding values. The red bounding box is the same target object.

Table 10: Per-background Δ mAP@50 for Cityscapes and Virtual KITTI. mAP@50 on "CSV" is 74.82 and on "VKC" is 79.42. It shows significant drop even with "CST", which was training set.

	road	sidewalk	building	wall	fence	pole	traffic light	traffic sign	vegetation	terrain	sky
CFV	-16.22	-15.62	-9.88	-1.43	-1.52	-4.92	-0.75	-0.98	-7.80	-3.25	-0.79
CRV	-14.45	-11.13	-10.21	-0.52	-0.39	-3.19	-0.04	-0.71	-5.93	-1.23	-0.04
CST	-8.75	-10.94	-6.04	-0.50	-0.96	-2.34	-0.03	-0.35	-5.03	-1.83	-0.22
CSV	-10.15	-10.51	-4.78	-0.63	-0.69	-3.22	-0.22	-0.39	-5.02	-1.95	-0.34
KST	-4.77	-7.01	-1.01	-0.95	-0.52	-2.96	+1.58	+0.11	-7.07	-6.29	-1.29

	terrain	sky	tree	vegetation	building	road	guardRail	traffic sign	traffic light	pole
VKC	-1.44	-0.93	-1.60	-0.54	-0.78	-0.96	-0.11	-0.38	+0.00	-0.11
VKF	-10.29	-7.05	-12.11	-8.86	-4.47	-7.57	-5.60	-2.36	+0.10	-0.80
VKM	-1.04	-1.88	-3.89	-0.08	-1.17	-1.33	-0.07	-0.34	-0.01	-0.23
VKR	-3.08	-0.40	-2.91	-3.17	-2.08	-2.48	-0.65	-1.31	+0.00	-0.59
VKO	-0.88	-1.10	+1.12	-1.25	+0.96	-3.02	+0.45	-0.36	-0.11	+0.86
VKS	-2.25	-1.76	-3.80	-0.56	-1.20	-1.56	-0.21	-0.59	-0.06	-0.13

Because the ViT-based model uses global attention, the map often activates across the entire image. After suppressing the top 20% of the highest responses, the "road" and "buildings" remain prominently highlighted.

Besides, we measured the drop rate of mAP@50 by removing background region in image space. Table 10 demonstrates the drop rate results on Cityscapes and Virtual KITTI datasets. Still, it expresses significant drops on BG labels. Interesting finding is that the drop rate rarely increases on few labels such as "traffic sign" and "traffic light". Due to its global attention feature, it may exhibit different FG–BG associations. It could be potentially token-wise, but globally. However, this investigation falls beyond our current scope and we aim to analyze transformer-based detectors in future work.

A.3 MMD with different kernel results

Different from the multiscale kernel used, we utilize RBF kernel to compute MMD. The number of pairs with statistically significant differences differs slightly from the main results, but the findings support the same overall conclusions.

Table 11: *Gradient* **comparison across different domains with RBF kernel** Each number's denominator is the number of FG–BG association in common across two domains. Cases 1, 2 and 3 are given as C1, C2 and C3. The bolded values represent the dominant case for each dataset-model pair. $*$ indicates that it differs from the result in the main text.

	Res			**ALDI++**			**Eff**		
	C1	C2	C3	C1	C2	C3	C1	C2	C3
CST - CSV	**2/5**$*$	1/5	**2/5**$*$	4/10	**6/10**	0/10	**3/7**	**3/7**	1/7
CST - CFV	**11/11**	0/11	0/11	**15/15**	0/15	0/15	**8/11**$*$	2/11	1/11$*$
CST - CRV	**10/11**	1/11	0/11	**8/10**	2/10	0/10	**6/10**	4/10	0/10
CST - KST	**1/1**	0/1	0/1	**1/1**	0/1	0/1	**2/3**	1/3	0/3
VKC - VKF	**5/5**	0/5	0/5	**4/4**	0/4	0/4	**2/2c**	0/2	0/2
VKC - VKM	**5/7**	2/7	0/7	**4/4**	0/4	0/4	1/5	**4/5**	0/5
VKC - VKO	**3/7**$*$	**3/7**$*$	1/7$*$	**4/4**	0/4	0/4	1/5	**3/5**$*$	1/5$*$
VKC - VKR	**6/6**	0/6	0/6	**5/5**	0/5	0/5	2/5$*$	**3/5**$*$	0/5
VKC - VKS	**4/6**$*$	0/6$*$	1/6$*$	**4/5**	1/5	0/5	1/5	**4/5**	0/5

Table 12: **Associated and non-associated features comparison across different domains with RBF kernel.** Each number's denominator is the number of association in common across two domains. Cases 1, 2 and 3 are given as C1, C2 and C3. "-" is not measurable statistically. The bolded values represent the dominant case for each dataset-model pair. $*$ indicates that it differs from the result in the main text.

	Res			**ALDI++**			**Eff**		
	C1	C2	C3	C1	C2	C3	C1	C2	C3
CST - CSV	**5/5**	0/5	0/5	**9/10**	0/10	1/10	**3/7**	2/7	2/7
CST - CFV	**11/11**	0/11	0/11	**9/15**	6/15	0/15	**6/11**	5/11	0/11
CST - CRV	**8/11**$*$	1/11$*$	2/11$*$	**7/10**	3/10	0/10	2/10	**6/10**	2/10
CST - KST	**1/1**	0/1	0/1	**1/1**	0/1	0/1	0/3	**2/3**	1/3
VKC - VKF	**3/5**$*$	1/5	1/5$*$	**4/4**	0/4	0/4	**2/2**	0/2	0/2
VKC - VKM	**4/7**	2/7	1/7	**4/4**	0/4	0/4	**3/5**	1/5	1/5
VKC - VKO	**4/7**$*$	2/7	1/7$*$	**4/4**	0/4	0/4	**4/5**	1/5	1/5
VKC - VKR	**4/6**	1/6	1/6	**4/5**$*$	1/5$*$	0/5	**4/5**	1/5	0/5
VKC - VKS	**5/6**	1/6	0/6	**4/5**$*$	1/5$*$	0/5	**3/5**	1/5	1/5