CAT: Exploiting Inter-Class Dynamics for Domain Adaptive Object Detection

Mikhail Kennerley[1,2], Jian-Gang Wang[2], Bharadwaj Veeravalli[1], and Robby T. Tan[3,1]
[1]National University of Singapore, Department of Electrical and Computer Engineering
[2]Institute for Infocomm Research, A*STAR
[3]ASUS Intelligent Cloud Services
mikhailk@u.nus.edu, jgwang@i2r.a-star.edu.sg, elebv@nus.edu.sg, robby.tan@nus.edu.sg

Abstract

Domain adaptive object detection aims to adapt detection models to domains where annotated data is unavailable. Existing methods have been proposed to address the domain gap using the semi-supervised student-teacher framework. However, a fundamental issue arises from the class imbalance in the labelled training set, which can result in inaccurate pseudo-labels. The relationship between classes, especially where one class is a majority and the other minority, has a large impact on class bias. We propose Class-Aware Teacher (CAT) to address the class bias issue in the domain adaptation setting. In our work, we approximate the class relationships with our Inter-Class Relation module (ICRm) and exploit it to reduce the bias within the model. In this way, we are able to apply augmentations to highly related classes, both inter- and intra-domain, to boost the performance of minority classes while having minimal impact on majority classes. We further reduce the bias by implementing a class-relation weight to our classification loss. Experiments conducted on various datasets and ablation studies show that our method is able to address the class bias in the domain adaptation setting. On the Cityscapes → Foggy Cityscapes dataset, we attained a 52.5 mAP, a substantial improvement over the 51.2 mAP achieved by the state-of-the-art method. [1]

1. Introduction

Domain adaptive object detection (DAOD) has been proposed as a solution for object detection in domains where no annotated data is available. This need is due to the increasing demands of data tied with annotation being both cost-prohibitive and potentially inaccurate in challenging domains. DAOD has been progressing with the introduction of adversarial learning [5, 33, 43, 48], style transfer [50, 51], and notably, student-teacher frameworks [1, 8, 16, 24, 26, 29, 37, 49]. Yet, these methods ignore the critical issue of class imbalance, which is a problem in many real-life scenarios, such as autonomous driving. For instance, in the Cityscapes dataset [7], the 'car' class dominates the dataset with 26,963 instances while classes such as 'train' contain only 168 instances.

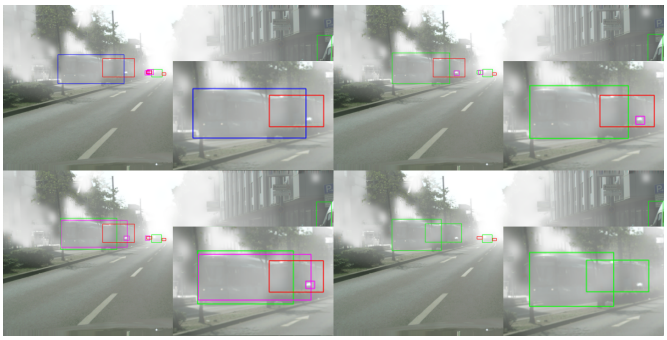

Figure 1. **Performance of Class-Aware Teacher (CAT).** AT [29] (top left), with Inter-Class Loss, ICL, (top-right), with Class Relation Augmentation, CRA, (bottom-left), and CAT (bottom-right). CAT is able to address misclassification and false positives, blue and red boxes, respectively, in minority classes such as 'train'. The combination of ICL and CRA further boosts performance by reducing the number of false positives shown as pink boxes.

Previous work to mitigate class imbalance in DAOD has applied class-specific discriminators [48] to align classes in distinct domains. Additionally, class weights has been proposed to boost minority categories while aligning domain features [3]. Recently, many DAOD methods employ the student-teacher framework, leading to improved performance. Despite their effectiveness, these student-teacher methods suffer from the class imbalance problem, resulting in poor performance for minority classes.

In the student-teacher framework, class-specific thresholding, which provides more lenient thresholds for minor-

[1] www.github.com/mecarill/cat

ity classes, has been proposed [23, 27, 46]. Yet, this approach does not address the fundamental class imbalance. Even with perfectly accurate pseudo-labels guiding the student, the model's bias would at best align with the biases present in the dataset, rather than providing an unbiased view. Moreover, inter-class dynamics play a crucial role in addressing class imbalance, especially when minority classes share high similarities with majority classes, increasing the likelihood of misclassification.

To address these challenges, in this paper, we introduce our Class-Aware Teacher (CAT), specifically designed to tackle class imbalance in the DAOD setting. CAT implements an Inter-Class Relation module (ICRm) that approximates the model's existing class biases as well as inter-class dynamics. With the knowledge of these biases, CAT applies Class-Relation Augmentation (CRA) to the training images. CRA increases the representation of minority classes by blending them with highly similar majority classes at the instance-level. To aid in this augmentation, a Cropbank [46] is used to store a collection of cropped instances. Furthermore, this augmentation is not just confined to the source domain but is also applied across domains. By allowing cross-domain augmentation, we are able to address the domain gap more holistically. To further address the inter-class bias, we propose an Inter-Class Loss (ICL). ICL utilises the insights from the ICRm to prioritise the model's attention towards minority classes. This priority is particularly focused on cases where minority classes are prone to being misclassified as majority classes.

By integrating these methods, our results indicate an improvement in the accuracy of minority class predictions, with a quantifiable increase in performance on benchmarks such as Cityscapes $\rightarrow$ Foggy Cityscapes by +1.3 mAP. Figure 1 demonstrates the performance of our method. We summarise the contributions of this paper as follows:

- We propose our Class-Aware Teacher (CAT) model, supported by our inter-class relation module (ICRm), which is able to map the model's existing class biases.
- We present Class-Relation Augmentation which emphasises augmentation between related classes across domains, coupled with Inter-Class Loss to further prioritise the performance of minority classes.
- Thorough experimental analysis that confirms the capabilities of CAT. Our experiments demonstrate significant improvements in performance compared with the state-of-the-art methods in DAOD benchmarks.

2. Previous Work

UDA for Object Detection Unsupervised Domain Adaptation (UDA) is designed to adapt a model trained on a labelled source domain to an unlabelled target domain. In object detection tasks, methods like adversarial training coupled with domain classifiers [5, 33, 43, 48] are prevalent for cultivating domain-invariant image feature representations. Other strategies, such as image-to-image translation, use generative models [50, 51] or frequency-based methods [44] to bridge the gap between domains. Recent approaches have applied the mean-teacher (MT) framework [1, 8, 16, 24, 26, 29, 37], initially conceived for semi-supervised learning, to UDA challenges. For instance, the UMT [8] leverages CycleGAN-generated images to train the student-teacher model, aiming to diminish domain bias. AT [29] employs strong-weak image augmentation, intentionally degrading the student's input compared to the teacher's, and incorporates adversarial training to further reduce the domain gap. 2PCNet [24] takes a two-stage approach to provide more diverse pseudo-labels with domain specific augmentation. Despite significant improvements over their predecessors, these methods often overlook the class imbalance issue prevalent in benchmark datasets. This oversight can lead to suboptimal performance on minority classes, some of which may appear up to 20 times less frequently than majority classes [7].

Class-Imbalanced Object Detection The issue of imbalance in object detection largely stems from an overrepresentation of background over foreground classes in predictions [31]. Our research, however, addresses the imbalance among foreground classes themselves, which often suffer from unequal frequency within datasets. A challenge here is the risk of overfitting to minority classes, particularly when their instances are sparse [14]. Over-sampling of minority classes, a strategy borrowed from classification tasks, has been adapted for object detection as well [46]. In student-teacher frameworks, a static hard threshold for pseudo-label generation has evolved into a dynamic, class-specific threshold to mitigate teacher bias [23, 27]. Although this approach can enhance the quality of pseudo-labels, it does not necessarily balance the sample distribution between majority and minority classes. In the DAOD, methods such as class-specific discriminators [48] and weighted losses [3] have been proposed to address class imbalance alongside domain adaptation. A critical aspect that remains underexplored is the inter-class relationship, particularly between majority and minority classes with similar features. In this paper, we aim to explore inter-class dynamics to effectively tackle class imbalance.

3. Preliminaries

Problem Definition In this paper, we propose a method for class balanced domain adaptive object detection, employing a labelled source dataset $D_s = \{I_s; Y_s\}$ and an unlabelled target dataset $D_t = \{I_t\}$. I_s and $Y_s = \{b_s; c_s\}$ denote the images and their corresponding ground-truth labels, which include bounding box and class information, indicated as b_s and c_s, respectively.

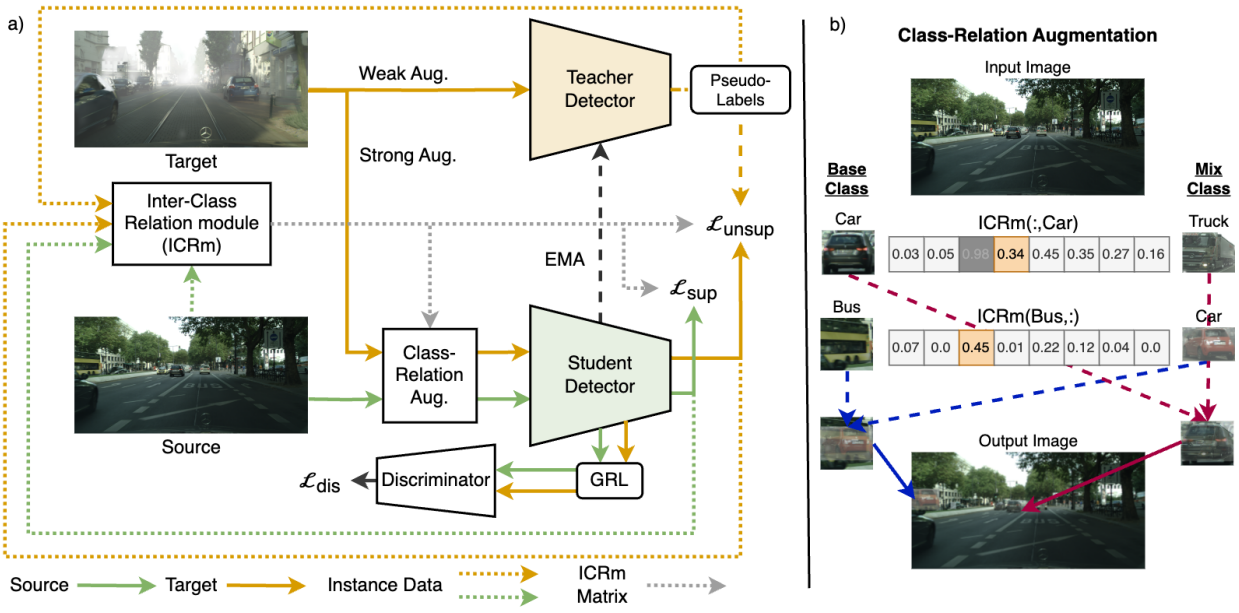

Figure 2. (a) **Class-Aware Teacher (CAT)** consists of: a student-teacher network; Inter-Class Relation module (ICRm), which estimates inter-class biases; Class-Relation Augmentation, which augments images to reduce the inter-class biases by mixing the cropped instances of related classes; and Inter-Class Loss, which emphasises the loss on highly misclassified minority classes. (b) Class-Relation Augmentation demonstrated on majority (Car) and minority (Bus) classes.

Mean Teacher We utilise the mean-teacher framework, comprising of a student and teacher network that shares identical architectures and network parameters. The teacher's network parameters, denoted as θ_t, are not updated through backpropagation but are instead updated using the Exponential Moving Average (EMA) of the student's parameters θ_s, following:

$$\theta_t \leftarrow \alpha\theta_t + (1-\alpha)\theta_s, \tag{1}$$

where α is the decay rate that controls the update momentum.

The teacher network generates pseudo-labels , $Y_t = \{b_t, c_t\}$ from weakly augmented unlabelled images. These pseudo-labels are utilised by the student network to calculate the unsupervised loss in conjunction with the strongly augmented inputs. The student's inputs are intentionally degraded compared to the teacher's inputs to challenge the student network further. The supervised loss is consistent with the Faster R-CNN framework [32], while the unsupervised loss is formulated as:

$$\mathcal{L}_{\text{unsup}} = \mathcal{L}^{\text{rpn}}_{\text{obj}}(I_t, b_t) + \mathcal{L}^{\text{roi}}_{\text{cls}}(I_t, b_t, c_t). \tag{2}$$

notably excluding the regression losses in the unsupervised context. We implement a hard threshold, τ, on the classification scores generated by the teacher to ensure that only pseudo-labels with high confidence are utilised by the student network, thereby encouraging more reliable learning outcomes. Following [29], a discriminator is added to encourage domain invariant feature representations with an associated loss, $\mathcal{L}_{\text{dis}}$.

4. Proposed Method

Our method, Class-Aware Teacher (CAT), as depicted in Figure 2.a, builds on the mean-teacher framework. Central to CAT is our Inter-Class Relation module (ICRm), designed to quantify the class biases inherent in the model. Unlike traditional approaches that address class bias in a broad sense, ICRm maps the dynamic relationships between classes. It particularly focuses on minority classes that are disproportionately misclassified as dominant majority classes. This mapping is achieved by constructing a confusion matrix for each batch, which is normalised against the ground truth, allowing for real-time bias estimation. A global matrix, updated continually with batch-level data, serves as a robust representation of the model's class biases. The ICRm is integral to our methodology, underpinning two components: the Class-Relation Augmentation and the Inter-Class Loss.

Class-Relation Augmentation addresses class imbalance at the image-level. Minority classes that share a high similarity to majority classes in an image are identified using ICRm. MixUp [47], which has been shown to address the imbalanced class problem [6, 11], is then used to merge the related minority and majority classes thus increasing the

representation of these minority classes. Our process not only boosts the number of minority class samples but also encourages the model to distinguish between closely associated classes, as illustrated in Figure 2.b.

Additionally, ICRm informs the distribution of weights within our Inter-Class Loss. This weighted loss function emphasises the loss on minority classes, especially those frequently mislabelled as majority classes. By doing so, we provide a counterbalance to the model's learned biases, nudging it towards a more balanced classification performance.

4.1. Inter-Class Relation module

Prior research addressing class imbalance within domain adaptation [17, 21, 36] has significantly advanced the performance on imbalanced classes. However, these methods often overlook the inter-class relationships and their impact on class imbalance. Our experimental observations suggest that the likelihood of misclassification between minority and majority classes is heavily influenced by their similarity. For instance, minority vehicle classes are more prone to be misclassified as 'car', a majority class, rather than as 'person', another majority class, due to their inherent resemblance.

Our approach aims to leverage these observed relationships through the Inter-Class Relation module (ICRm). Distinct from general class bias, inter-class dynamics cannot be directly inferred from the dataset but must be extrapolated from the model during training. We achieve this by generating a confusion matrix at each training batch that cross-references the ground-truth labels with the model's predictions. This matrix is normalised with respect to the ground truth to estimate the bias between classes.

Subsequently, we employ EMA to iteratively update a global matrix, which represents a more stable and comprehensive approximation of the model's class biases. The EMA's utility extends beyond smoothing; it removes the need for each class's presence in every batch, simplifying the training process. The process for constructing this matrix is outlined in Algorithm 1. The ICRm is formulated separately for both source images, referencing actual ground-truth labels, and target images, utilising pseudo-labels to mirror the ground-truth.

4.2. Class-Relation Augmentation

In classification task, oversampling is a common technique to counter class imbalance by increasing the presence of minority images. However, this approach presents challenges in object detection, where images often contain a mix of multiple object classes. Our analysis on the Cityscapes [7] dataset indicates that most images include at least one instance of a majority class, rendering image-level resampling ineffective. This complexity demands more nuanced augmentation strategies for object detection.

Similar to Zhang et al. [46], we employ instance-level oversampling. Instances are cropped from their images using bounding box annotations and are then strategically inserted into other images.

Images are randomly selected from each batch and utilising the ICRm, we differentiate classes as majority or minority based on their likelihood of correct classification. We then derive the mean probability :

$$\mathrm{ICRm}_{\mathrm{avg}} = \frac{1}{C}\sum_{c=0}^{C}\mathrm{ICRm}(c,c), \tag{3}$$

where C is the number of classes. Classes with a probability, $\mathrm{ICRm}(c,c)$, above and below the mean probability are designated as majority and minority classes, respectively.

Instead of random overlaying, which has been adopted by other previous approaches, our method matches highly related minority and majority instances and uses MixUp [47] to blend them, allowing the model to have better generalisation towards minority classes.

We achieve this by pairing each base instance in an image with a sampled instance chosen through weighted random sampling, using the ICRm class probabilities as weights. For majority base instances, the corresponding column in ICRm, namely, $\mathrm{ICRm}(:,c)$ is used, discounting the class's own probability by setting $\mathrm{ICRm}(c,c)$ to zero to avoid self-augmentation. This allows us to select classes that are frequently misclassified as the majority class. Conversely, for minority base instances, we use the corresponding row from ICRm as weights without zeroing out $\mathrm{ICRm}(c,c)$, allowing for the possibility of self-augmentation, which can be beneficial for minority classes. This is demonstrated in Figure 2.b.

Sampled instances are resized to match the base instance dimensions for bounding box consistency. The MixUp augmentation is then applied as per the following formulation, where a beta distribution determines the mixing ratio:

$$\begin{aligned}\hat{I} &= \beta \cdot I_{\mathrm{base}} + (1-\beta)\cdot I_{\mathrm{mix}},\\ \hat{c} &= \beta \cdot c_{\mathrm{base}} + (1-\beta)\cdot c_{\mathrm{mix}},\end{aligned} \tag{4}$$

where I_{base} and I_{mix} represent the cropped images of the base and mixed instances, respectively, with $\hat{I}$ denoting the resulting augmented image. Similarly, c_{base} and c_{mix} refer to the classes of the base and mixed instances, while $\hat{c}$ indicates the class vector of the augmented instance.

For source domain images, we incorporate instances from both domains to leverage accurate source labels and to aid domain adaptation with target domain samples. Whereas for target domain images, we prioritise target instances, using source instances only when no target instances are available for a specific class. This ensures that

Algorithm 1 Inter-Class Relation module (ICRm)

Require: Global class-relation matrix $ICRm$ with shape (C, C), where each element equals 0.
  while $training$ **do**
    Create batch matrix ICRm_l with shape (C, C), where each element equals 0.
    for ground-truth, c_i, and prediction,x_i, in Y, X **do**
      $\mathrm{ICRm}_l[c_i, x_i] \leftarrow \mathrm{ICRm}_l[c_i, x_i] + 1$
    end for
    for each class c in C **do**
      Normalise class matrix $\mathrm{ICRm}_l[c]$
      if global class matrix $\mathrm{ICRm}[c]$ is empty **then**
        //Copy batch class matrix to global
        $\mathrm{ICRm} \leftarrow \mathrm{ICRm}_l$
      else
        //Update global class matrix with EMA
        $\mathrm{ICRm} \leftarrow \beta * \mathrm{ICRm} + 1(1 - \beta) * \mathrm{ICRm}_l$
      end if
    end for
  end while

more focus is put onto the target domain for stronger domain adaptation. Additionally, we do not apply augmentations to minority base instances in the target domain to preserve their integrity. This ensures that the model is able to focus on the target domain and does not drift to an intermediate domain.

To implement Class-Relation Augmentation, we store class-specific instance crops from each batch, which we term a Cropbank [46]. These crops are extracted from bounding box annotations of labelled source images and pseudo-labelled target images. Separate Cropbanks are maintained for both the source and target datasets, allowing for more targeted augmentation. To ensure a diverse range of samples, we update the class instances on a first-in-first-out basis. This is particularly beneficial for the target Cropbank, where earlier samples may less accurate due to the early pseudo-labels' robustness.

4.3. Inter-Class Loss

To further mitigate class bias, we introduce a weighted parameter to the classification loss, informed by the Inter-Class Relation module (ICRm) for foreground classes. This weighting prioritises classes that are frequently misclassified as majority classes, allowing the model to concentrate on refining their performance. To emphasise the focus on underperforming classes, we employ a non-linear transformation on the ICRm values:

$$w_i = \begin{cases} \sqrt{1 - \mathrm{ICRm}(c_i, x_i)}, & \text{if } c_i = x_i \\ \sqrt{\mathrm{ICRm}(c_i, x_i) / \mathrm{ICRm}(c_i, c_i)}, & \text{otherwise,} \end{cases} \quad (5)$$

where w_i is the ith weight in W, and c_i and x_i are the ith ground-truth and predicted class, respectively. We normalise on the diagonal when $c_i \neq x_i$ as our primary objective is to prioritise low performing classes. Weights for background classes are uniformly set to 1 to avoid biasing the model against them. To reconcile the disparity between foreground and background class weights, the weights of the foreground instances are normalised so that their mean equals the background class weight:

$$W_f = W_f / \mathrm{mean}(W_f), \quad (6)$$

where W_f denotes the collection of foreground instance weights. Moreover, we integrate an additional regularisation term, λ_l, across all class-relation weights to prevent extreme weight values from distorting the loss:

$$W = \frac{(W + \lambda_l)}{(1 + \lambda_l)} \quad (7)$$

This regularisation ensures a moderated, balanced impact on the classification loss, which is now defined as:

$$\mathcal{L}_{\mathrm{cls}} = \frac{1}{N} \sum_{i=0}^{N} w_i * \mathrm{CE}(c_i, x_i) \quad (8)$$

where N is the number of instances and CE is the cross-entropy loss.

The overall loss is then:

$$\mathcal{L} = \mathcal{L}_{\mathrm{sup}} + \lambda_u \mathcal{L}_{\mathrm{unsup}} + \lambda_d \mathcal{L}_{\mathrm{dis}}, \quad (9)$$

where λ_u and λ_d represent the weights for the unsupervised and discriminator losses, respectively.

5. Experiments

5.1. Datasets

We assess the performance of Class-Aware Teacher (CAT) using benchmarks in domain adaptive object detection (DAOD) following prior work [29].

Cityscapes → Foggy Cityscapes: The Cityscapes dataset [7] is a road-centric dataset with 2,975 training and 500 validation images from various urban settings under clear weather, annotated across 8 classes. Foggy Cityscapes [34] is a synthesised dataset generated on Cityscapes to simulate foggy weather, using the same base images and annotations. We conduct our experiment on the most severe fog level (0.02) where Foggy Cityscapes is used as the target domain.

PASCAL VOC → Clipart1K: We use PASCAL VOC 2012 [10], which comprises of 11,540 real-world images

Method	Detector	person	rider	car	truck	bus	train	mcycle	bicycle	mAP
Source [38]	FCOS	36.9	36.3	44.1	18.6	29.3	8.4	20.3	31.9	28.2
SIGMA [28]	FCOS	46.9	48.4	63.7	27.1	50.7	35.9	34.7	41.4	43.5
OADA [45]	FCOS	47.8	46.5	62.9	32.1	48.5	50.9	34.3	39.8	45.4
HT [9]	FCOS	52.1	55.8	67.5	32.7	55.9	49.1	40.1	50.3	50.4
Source [52]	Def DETR	37.7	39.1	44.2	17.2	26.8	5.8	21.6	35.5	28.5
AQT [19]	Def DETR	49.3	52.3	64.4	27.7	53.7	46.5	36.0	46.4	47.1
MRT [49]	Def DETR	**52.8**	51.7	**68.7**	35.9	58.1	**54.5**	41.0	47.1	51.2
Source [32]	FRCNN	22.4	26.6	28.5	9.0	16.0	4.3	15.2	25.3	18.4
Oracle	FRCNN	39.5	47.3	59.1	33.1	47.3	42.9	38.1	40.8	43.5
MeGA [39]	FRCNN	37.7	49.0	52.4	25.4	49.2	46.9	34.5	39.0	41.8
TIA [48]	FRCNN	34.8	46.3	49.7	31.1	52.1	48.6	37.7	38.1	42.3
UMT [8]	FRCNN	33.0	46.7	48.6	34.1	56.5	46.8	30.4	37.3	41.7
TDD [16]	FRCNN	39.6	47.5	55.7	33.8	47.6	42.1	37.0	41.4	43.1
PT [4]	FRCNN	40.2	48.8	59.7	30.7	51.8	30.6	35.4	44.5	42.7
AT‡ [29]	FRCNN	45.3	55.7	63.6	36.8	64.9	34.9	42.1	51.3	49.3
CMT [1]	FRCNN	45.9	55.7	63.7	39.6	**66.0**	38.8	41.4	51.2	50.3
MILA [26]	FRCNN	45.6	52.8	64.8	34.7	61.4	54.1	39.7	51.5	50.6
CAT (Ours)	FRCNN	44.6	**57.1**	63.7	**40.8**	**66.0**	49.7	**44.9**	**53.0**	**52.5**

Table 1. Object detection results on the Foggy Cityscapes test set for **Cityscapes → Foggy Cityscapes (0.02) domain adaptation**. We group methods based on their detector frameworks (FCOS/Def DETR/FRCNN) and highlight the best performing method. CAT is able to outperform the previous state-of-the-art, MRT, by 1.3 mAP and improve on AT by 3.2 mAP. The mean average precision at .50 IoU (mAP) is reported for all classes. ‡ AT performance is reproduced on Foggy Cityscapes (0.02) with publicly available code for fairness.

across 20 categories, for training. The Clipart1k dataset [20] consists of 20 corresponding clipart object categories. Following [29], we split Clipart1k into 500 training and 500 testing images.

The benchmarks of Sim10K [22] → Cityscapes and KITTI [12] → Cityscapes are excluded from our evaluation. Despite their popularity in DAOD research, they focus solely on the 'Car' category, which does not align with our class imbalance setting.

5.2. Experimental Setup

Following previous works in DAOD, we adopt the Faster R-CNN object detector with VGG-16 [35] and ResNet-101 [15] as backbones for our detection model. Our hyperparameters: EMA decay rate $\alpha = 0.9996$, beta-distribution hyper parameters [0.5,0.5], adversarial loss weight, λ_d 0.1, and unsupervised loss weight, λ_u 1.0, regularisation term, λ_l 1.0. We employ a hard threshold τ of 0.8 for pseudo-labelling. Weak-strong augmentation [30] is applied to both source and target images. We train our student model for 20,000 iterations on the labelled source data. The parameters of the student is copied to the teacher which is then updated via EMA of the student at each iteration. We continue training for 60,000 iterations with both labelled source and unlabelled target data. Our framework is developed on top of the publicly available Detectron2 [41]. Experiments are performed using a batch size of 8 source and 8 target images, distributed across 4 NVIDIA RTX3090 GPUs. Additional details regarding our experimental setup are provided in the supplementary materials.

5.3. Comparison with SOTA methods

We compare our method with the state-of-the-art in DAOD as well as reporting a source only FCOS/Def DETR/Faster RCNN for a baseline comparison. We additionally include an oracle upper bound, which is trained on only the target domain and its ground truth annotations.

Foggy Weather Adaptation When object detectors are deployed in real-world scenarios, the performance could drop significantly under sub-optimal conditions, e.g. adverse weather. This is because that the samples in adverse weather do not present in the training of the model resulting in a domain shift. The domain adaptation task is designed to overcome this gap between normal and adverse conditions. To demonstrate this, we conduct an experiment on the commonly used Cityscapes → Foggy Cityscapes benchmark.

Our results are shown in Table 1. We can observe that methods utilising the student-teacher framework (HT, UMT, TDD, PT, AT, CMT, MILA, MRT) outperform non-student-teacher frameworks by a large margin. CAT, which

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mtr	prsn	plant	shp	sofa	train	tv	mAP
Source [13]	23.0	39.6	20.1	23.6	25.7	42.6	25.2	0.9	41.2	25.6	23.7	11.2	28.2	49.5	45.2	46.9	9.1	22.3	38.9	31.5	28.8
Oracle	33.3	47.6	43.1	38.0	24.5	82.0	57.4	22.9	48.4	49.2	37.9	46.4	41.1	54.0	73.7	39.5	36.7	19.1	53.2	52.9	45.0
I3Net [3]	30.0	**67.0**	32.5	21.8	29.2	62.5	41.3	11.6	37.1	39.4	27.4	19.3	25.0	67.4	55.2	42.9	19.5	36.2	50.7	39.3	37.8
ICR-CCR [42]	28.7	55.3	31.8	26.0	40.1	**63.6**	36.6	9.4	38.7	49.3	17.6	14.1	33.3	74.3	61.3	46.3	22.3	24.3	49.1	44.3	38.3
HTCN [2]	33.6	58.9	34.0	23.4	45.6	57.0	39.8	12.0	39.7	51.3	20.1	20.1	39.1	72.8	61.3	43.1	19.3	30.1	50.2	51.8	40.3
DM [25]	25.8	63.2	24.5	42.4	47.9	43.1	37.5	9.1	47.0	46.7	26.8	24.9	48.1	78.7	63.0	45.0	21.3	36.1	52.3	**53.4**	41.8
UMT [8]	39.6	59.1	32.4	35.0	45.1	61.9	48.4	7.5	46.0	**67.6**	21.4	**29.5**	48.2	75.9	70.5	56.7	25.9	28.9	39.4	43.6	44.1
TIA [48]	42.2	66.0	36.9	37.3	43.7	71.8	49.7	18.2	44.9	58.9	**18.2**	29.1	40.7	87.8	67.4	49.7	**27.4**	27.8	**57.1**	50.6	46.3
AT‡ [29]	33.1	66.1	35.3	44.9	57.5	**44.9**	51.0	5.8	59.5	54.9	34.6	23.5	64.3	84.0	75.4	51.5	17.1	30.3	43.3	37.2	45.7
CMT [1]	39.8	56.3	38.7	39.7	60.4	35.0	**56.0**	7.1	**60.1**	60.4	35.8	28.1	**67.8**	84.5	**80.1**	55.5	20.3	32.8	42.3	38.2	47.0
CAT (Ours)	**40.5**	64.1	**38.8**	41.0	**60.7**	55.5	55.6	14.3	54.7	59.6	**46.2**	20.3	58.7	**92.9**	62.6	**57.5**	22.4	**40.9**	49.5	46.0	**49.1**

Table 2. Object detection results on the Clipart1k test set for **PASCAL VOC → Clipart1k domain adaptation**. CAT improves on the previous state-of-the-art, CMT, by 2.1 mAP, achieving the new best of 49.1 mAP. The mean average precision at .50 IoU (mAP) is reported for all classes.‡ AT performance is reproduced following [1].

Method	ICRm	CRA	ICL	mAP	$\sigma \downarrow$
Base (AT [29])				49.3	10.8
CAT (Ours)	✓	✓		51.0	8.8
	✓		✓	51.6	10.3
	✓	✓	✓	52.5	8.6

Table 3. **Ablation studies on CAT components.** ICRm is included in all studies as it forms the basis of CRA and ICL. We report the mean average precision at .50 IoU (mAP) and the standard deviation of class-mAP (σ). Our contributions are not included in the base framework (AT).

is built on existing SOTA mean teacher frameworks, significantly improves the performance at 52.5 mAP. Additionally, our method is able to improve minority classes while not impacting majority classes.

Real to Artistic Adaptation Adapting object detection from real to artistic domains is particularly challenging due to the significant differences between these domains. In our experiment, detailed in Table 2, we observe CAT achieves a mAP of 49.1, outperforming the previous best, by 2.1 mAP and AT by 3.4 mAP.

Notably, CAT shows substantial improvements in minority classes, such as 'motorbike' which in the Clipart1k training set consists of only 7 images. The results of this experiment demonstrate CAT's effectiveness in addressing class imbalances even in dissimilar domains.

5.4. Ablation Studies

To verify the significance of our contributions, we conducted an ablation study. All experiments within this study were performed on the Cityscapes → Foggy Cityscapes benchmark using the VGG16 backbone.

Selection Method	mAP	$\sigma \downarrow$
Random (0.5)	51.8	8.5
CRA (1.0)	50.2	9.2
CRA (0.5)	52.5	8.6

Table 4. Comparison of selecting class instances randomly and via CRA. Values in brackets refer to the likelihood of an instance being augmented.

Quantitative Ablation Table 3 quantitatively showcases the efficacy of each contribution within our framework. The base framework, prior to integrating our modules, corresponds to the AT model as described in [29]. Since our Inter-Class Relation module (ICRm) is pivotal for both the class-relation augmentation and loss, it is a constant across all experimental variations. To highlight our method's capability in addressing class imbalance, we introduce σ. This value represents the standard deviation of the mAP across different classes and serves as an indicator of performance equity among classes.

Inclusion of our Inter-Class Loss (ICL) yields a notable 2.3 mAP gain over the base AT model and a decreased σ, indicating a more balanced performance across classes. The Class Relation Augmentation (CRA) also benefits AT, though to a lesser extent than ICL in terms of mAP. Notably, CRA significantly narrows the performance gap between minority and majority classes, as reflected by a reduced σ of 8.8. Employing both ICL and CRA not only enhances overall performance but also achieves a lower σ compared to the base model, reinforcing our method's effectiveness in managing class imbalance.

Impact of Augmentation We demonstrate the impact of augmentation in terms of ratio and selection criteria. Augmenting images is a key part of our approach, enriching the dataset with additional representations of minority classes.

Figure 3. **Qualitative results of CAT.** We show the results of AT and CAT on the top and bottom, respectively. CAT is able to address misclassification (col 1,2,4), false negatives (col 1,3), and false positives (col 1,3,4). Box colour represents: Green → true positives, Blue → misclassified, Red → false negatives, Pink → false positives.

Yet, if images are augmented excessively, the model may fail to learn an accurate representation of the classes. To strike a balance, we selectively augment a random subset of the images. Additionally, how class instances are paired is key to improve the quality of augmentation. Compared to randomly selecting class pairs, CRA is able to prioritise pairing highly related minority and majority class instances. This ensures that the Mixup output is more meaningful for minority class performance.

The experimental results of this approach are given in Table 4. We compare randomly selecting class instances for MixUp with our Class-Relation Augmentation at different values. These values represent the likelihood of a instance in the base image being augmented. Randomly applying MixUp improves the overall performance by +0.2 mAP. However, by using CRA we can further increase performance by +0.9 mAP. This shows that pairing highly-related classes for MixUp strengthens the performance of the minority class while minimally affecting the majority class. In addition, our experiments show that too much augmentation can have a negative effect on the performance of the model.

Weighing Strategy for Class Loss We introduce a weighted classification loss to improve the performance of minority class performance in Eq. 5. Class-level loss is a common strategy that has been adopted to address the class imbalance in a dataset [18, 40]. We compare our Inter-Class Loss (ICL) with a variant of previous class-level losses where only the diagonal of the Inter-Class Relation module is used. The diagonal corresponds to the ground-truth class likelihood of accurate classification. In contrast, ICL uses the likelihood of the ground-truth being classified as the predicted class to influence its loss. We show in Table 5 that by using this inter-class relationship, we are able to improve the performance by +0.5 mAP. However, there may be cases where ICL overly penalises well-performing classes. This is addressed by applying a regularisation term as seen in Eq. 7. We can see that if this regularisation term is removed, performance drops significantly as the weight of certain classes gets too small during training.

Class Weight	mAP	$\sigma \downarrow$
Class-Level	52.0	9.0
ICL w/o Reg.	51.3	9.5
ICL	52.5	8.6

Table 5. Class Loss Weighing Strategies. Class-level only uses the diagonal values in our ICRm along with regularisation, ICL refers to our Inter-Class Loss.

Qualitative Results Figure 3 illustrates the qualitative results of our method, with the top and bottom row displaying predictions from AT and CAT, respectively.

CAT is able to correct misclassifications, represented by blue boxes. Additionally, CAT can bring a reduction in both false positives and false negatives, represented by pink and red boxes, showcasing improved detection accuracy across various scales and classes.

6. Conclusion

In this paper, we propose Class-Aware Teacher (CAT) for Domain Adaptive Object Detection. We demonstrate that CAT, by leveraging our Inter-Class Relation module, effectively approximates and mitigates class biases, leading to more equitable performance across classes. Furthermore, Class-Relation Augmentation and Inter-Class Loss were shown to be effective in enhancing minority class representation. Our experimental results on Cityscapes → Foggy Cityscapes and PASCAL VOC → Clipart1K have demonstrated the effectiveness of our method achieving SOTA performance at 52.5 mAP and 49.1 mAP, respectively. Based on our findings, we believe that further investigation into inter-class dynamics is a promising direction for advancing class imbalance in the DAOD setting.

References

[1] Shengcao Cao, Dhiraj Joshi, Liang-Yan Gui, and Yu-Xiong Wang. Contrastive mean teacher for domain adaptive object detectors. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 23839–23848, 2023. 1, 2, 6, 7

[2] Chaoqi Chen, Zebiao Zheng, Xinghao Ding, Yue Huang, and Qi Dou. Harmonizing transferability and discriminability for adapting object detectors. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. 7

[3] Chaoqi Chen, Zebiao Zheng, Yue Huang, Xinghao Ding, and Yizhou Yu. I3net: Implicit instance-invariant network for adapting one-stage object detectors. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12576–12585, 2021. 1, 2, 7

[4] 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. In International Conference on Machine Learning, pages 3040–3055, 2022. 6, 3

[5] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3339–3348, 2018. 1, 2

[6] H. Chou, S. Chang, J. Pan, W. Wei, and D. Juan. Remix: rebalanced mixup. Computer Vision – ECCV 2020 Workshops, pages 95–110, 2020. 3

[7] 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3213–3223, 2016. 1, 2, 4, 5

[8] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4089–4099, 2021. 1, 2, 6, 7

[9] Jinhong Deng, Dongli Xu, Wen Li, and Lixin Duan. Harmonious teacher for cross-domain object detection. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 23829–23838, 2023. 6

[10] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. International Journal of Computer Vision, pages 303–308, 2009. 5

[11] A. Galdrán, G. Carneiro, and M. Á. G. Ballester. Balanced-mixup for highly imbalanced medical image classification. Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, pages 323–333, 2021. 3

[12] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. International Journal of Robotics Research, 2013. 6

[13] Ross Girshick. Fast r-cnn. In IEEE/CVF International Conference on Computer Vision, pages 1440–1448, 2015. 7

[14] Haibo He and Edwardo A Garcia. Learning from imbalanced data. IEEE Transactions on knowledge and data engineering, 21(9):1263–1284, 2009. 2

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 770–778, 2016. 6, 1

[16] Mengzhe He, Yali Wang, Jiaxi Wu, Yiru Wang, Hanqing Li, Bo Li, Weihao Gan, Wei Wu, and Yu Qiao. Cross domain object detection by target-perceived dual branch distillation. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9560–9570, 2022. 1, 2, 6, 3

[17] Tzu Ming Harry Hsu, Wei Yu Chen, Cheng-An Hou, Yao-Hung Hubert Tsai, Yi-Ren Yeh, and Yu-Chiang Frank Wang. Unsupervised Domain Adaptation with Imbalanced Cross-Domain Data. In IEEE/CVF International Conference on Computer Vision, pages 4121–4129, 2015. 4

[18] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. 8

[19] Wei-Jie Huang, Yu-Lin Lu, Shih-Yao Lin, Yusheng Xie, and Yen-Yu Lin. Aqt: Adversarial query transformers for domain adaptive object detection. In International Joint Conference on Artificial Intelligence (IJCAI), 2022. 6

[20] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018. 6

[21] Xiang Jiang, Qicheng Lao, Stan Matwin, and Mohammad Havaei. Implicit Class-Conditioned Domain Alignment for Unsupervised Domain Adaptation. In International Conference on Machine Learning, pages 4816–4827, 2020. 4

[22] Matthew Johnson-Roberson, Charles Barto, Rounak Mehta, Sharath Nittur Sridhar, Karl Rosaen, and Ram Vasudevan. Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? In International Conference on Robotics and Automation, pages 746–753. IEEE, 2017. 6

[23] Purbayan Kar, Vishal Chudasama, Naoyuki Onoe, and Pankaj Wasnik. Revisiting class imbalance for end-to-end semi-supervised object detection. In IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pages 4570–4579, 2023. 2

[24] Mikhail Kennerley, Jian-Gang Wang, Bharadwaj Veeravalli, and Robby T. Tan. 2pcnet: Two-phase consistency training for day-to-night unsupervised domain adaptive object detection. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11484–11493, 2023. 1, 2

[25] Taekyung Kim, Minki Jeong, Seunghyeon Kim, Seokeon Choi, and Changick Kim. Diversify and match: A domain adaptive representation learning paradigm for object detection. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019. 7

[26] Onkar Krishna, Hiroki Ohashi, and Saptarshi Sinha. Mila: Memory-based instance-level adaptation for cross-domain

object detection. British Machine Vision Conference, (BMVC), 2023. 1, 2, 6

[27] Hengduo Li, Zuxuan Wu, Abhinav Shrivastava, and Larry S. Davis. Rethinking pseudo labels for semi-supervised object detection. In AAAI, 2021. 2

[28] Wuyang Li, Xinyu Liu, and Yixuan Yuan. Sigma: Semantic-complete graph matching for domain adaptive object detection. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 6, 3

[29] 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7571–7580, 2022. 1, 2, 3, 5, 6, 7

[30] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. In International Conference on Learning Representations, 2021. 6

[31] Kemal Oksuz, Baris Can Cam, Sinan Kalkan, and Emre Akbas. Imbalance Problems in Object Detection: A Review. IEEE Transactions on Pattern Analysis and Machine Intelligence, pages 1–1, 2020. 2

[32] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in Neural Information Processing Systems, page 91–99, 2015. 3, 6

[33] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019. 1, 2

[34] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. International Journal of Computer Vision, 126(9):973–992, 2018. 5

[35] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. pages 1–14. Computational and Biological Learning Society, 2015. 6, 1

[36] Korawat Tanwisuth, Xinjie Fan, Huangjie Zheng, Shujian Zhang, Hao Zhang, Bo Chen, and Mingyuan Zhou. A Prototype-Oriented Framework for Unsupervised Domain Adaptation. In Advances in Neural Information Processing Systems, pages 17194–17208, 2021. 4

[37] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In Advances in Neural Information Processing Systems, page 1195–1204, 2017. 1, 2

[38] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In IEEE/CVF International Conference on Computer Vision, 2019. 6

[39] Vibashan Vs, Vikram Gupta, Poojan Oza, Vishwanath A Sindagi, and Vishal M Patel. Mega-cda: Memory guided attention for category-aware unsupervised domain adaptive object detection. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4516–4526, 2021. 6

[40] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. In Advances in Neural Information Processing Systems. Curran Associates, Inc., 2017. 8

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

[42] Chang-Dong Xu, Xingjie Zhao, Xin Jin, and Xiu-Shen Wei. Exploring categorical regularization for domain adaptive object detection. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11721–11730, 2020. 7

[43] Minghao Xu, Hang Wang, Bingbing Ni, Qi Tian, and Wenjun Zhang. Cross-domain detection via graph-induced prototype alignment. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. 1, 2

[44] Yanchao Yang and Stefano Soatto. FDA: Fourier domain adaptation for semantic segmentation. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4084–4094, 2020. 2

[45] Jayeon Yoo, Inseop Chung, and Nojun Kwak. Unsupervised domain adaptation for one-stage object detector using offsets to bounding box. In European Conference on Computer Vision, pages 691–708. Springer, 2022. 6

[46] Fangyuan Zhang, Tianxiang Pan, and Bin Wang. Semi-supervised object detection with adaptive class-rebalancing self-training. AAAI, 36(3):3252–3261, 2022. 2, 4, 5

[47] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. International Conference on Learning Representations, 2018. 3, 4, 1

[48] Liang Zhao and Limin Wang. Task-specific inconsistency alignment for domain adaptive object detection. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 1, 2, 6, 7

[49] 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 (ICCV), pages 19039–19049, 2023. 1, 6

[50] Ziqiang Zheng, Yang Wu, Xinran Nicole Han, and Jianbo Shi. Forkgan: Seeing into the rainy night. In European Conference on Computer Vision, 2020. 1, 2

[51] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In IEEE/CVF International Conference on Computer Vision, pages 2242–2251, 2017. 1, 2

[52] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable {detr}: Deformable transformers for end-to-end object detection. In International Conference on Learning Representations, 2021. 6

CAT: Exploiting Inter-Class Dynamics for Domain Adaptive Object Detection

Supplementary Material

7. Additional Details on Methods

7.1. Additional Details on Class-Relation Augmentation

We further describe the details of our Class-Relation Augmentation (CRA) approach below. CRA augments random images in a batch based on the source and target augmentation ratio. For each selected image, we identify class instances using labels or pseudo-labels for source and target images, respectively, termed 'base instances.'

Following the methodology outlined in Section 4.2, we select 'mix instances' that exhibit a strong relationship with the base instances, determined by our Inter-Class Relation module (ICRm). A 'mixed instance' is then randomly chosen from a predefined crop bank. To mitigate the effects of upsampling degradation, we ensure the mixed instances is at least 0.25 of the base instance's size.

We resize the mixed instance to the base instance's dimensions, allowing the aspect ratio of the mixed instance to be adjusted. This resizing allows us to use a single bounding box to represent both the base and mixed instance after augmentation. Experimental results, presented in Table 1, demonstrate that this resizing strategy not only maintains but enhances model performance compared to maintaining the mixed instance aspect ratio. This is because the ambiguity of labelling when two bounding boxes are used is complex, especially when employing mixup.

Once the mixed instance has been resized, mixup [47] is then applied to combine the two instances and their labels. Given the distinct class representations, we employ one-hot encoding to support multi-class labelling. This process is repeated across all objects in the selected image.

Maintain Aspect Ratio	mAP
✗	52.5
✓	51.1

Table 6. Performance of Class-Aware Teacher (CAT) with and without maintaining the aspect ratio during CRA. We can see that disregarding the aspect ratio during resizing improves performance while being a simpler resizing strategy.

8. Experiments

8.1. Additional Details on Experimental Setup

In this section, we provide additional details on the experimental setup. Consistent with prior research in the domain of adaptive object detection, our experiments are conducted using the Faster R-CNN detection framework. VGG-16 [35] and ResNet-101 [15] are used as the backbones for our detection model depending on the benchmark used. PASCAL VOC → Cliapart1K utilises the ResNet-101 backbone. Both Cityscapes → Foggy Cityscapes and Cityscapes → BDD100K utilises the VGG-16 backbone.

Across all experiments, we maintain consistent hyperparameter settings, which are detailed in Table 2.

8.2. Additional Details on Dataset Class Distributions

The distribution of classes in our datasets plays an important role during training. Minority classes tend to under perform, especially when there is a distribution shift between training and validation datasets. To validate the effectiveness of our method, we show the class distributions of the evaluation datasets and how our method is able to address minority class performance.

Figure 4 shows the class distribution for the Cityscapes → Foggy Cityscapes task. Car and person forms the majority in all the datasets used for this task and truck, bus, and train form the minority. This is to be expected as the datasets are from the same source and would share similar distributions. This forms a simpler task as we do not need to account for a distribution shift during testing. Our method matches or outperforms SOTA for the truck and bus class, as well as strongly outperforming our base method [29] for all three minority classes.

The class distribution of the PASCAL VOC → Clipart1k task is shown in Figure 5. The PASCAL VOC dataset is fairly balanced with the only outlier being the person class. This ensures that the initial training has less bias towards specific classes, however, Clipart1k exhibits stronger class imbalance. This results in a distribution shift during unsupervised training and evaluation which may result in suboptimal performance. CAT is able to have strong performance on the motorbike minority class and is able to outperform its base on the bus class.

The Cityscapes → BDD100K (Daytime) task contains two road-centric datasets taken in different locations which would result in both imbalanced data as well as a distribution shift as seen in Figure 6 . This would be a harder task as a minority class in one dataset may not be the same minority the other. For example, truck and bus are the minority for Cityscapes but motorcycle and bicycle are the minority for BDD100K. CAT is able to outperform SOTA for truck, bus, and bicycle and is only 0.1 mAP lower for the motorcycle minority class.

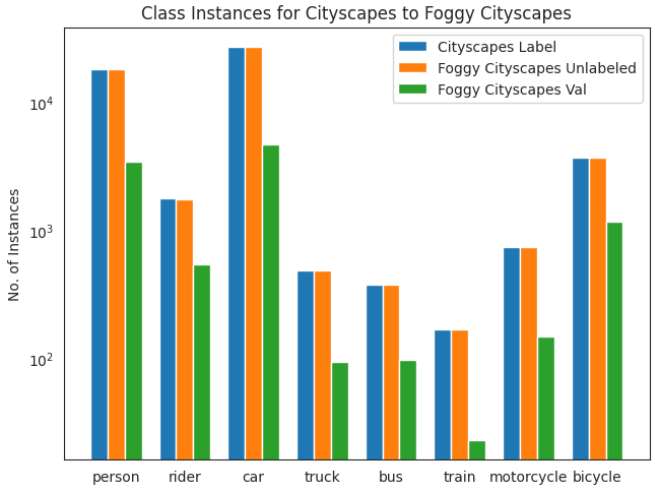

Figure 4. Class distribution of datasets used for the Cityscapes → Foggy Cityscapes task. We can see that person and car classes form the majority of all classes. The distribution of classes for the labeled dataset and validation set is similar which makes for an simpler task.

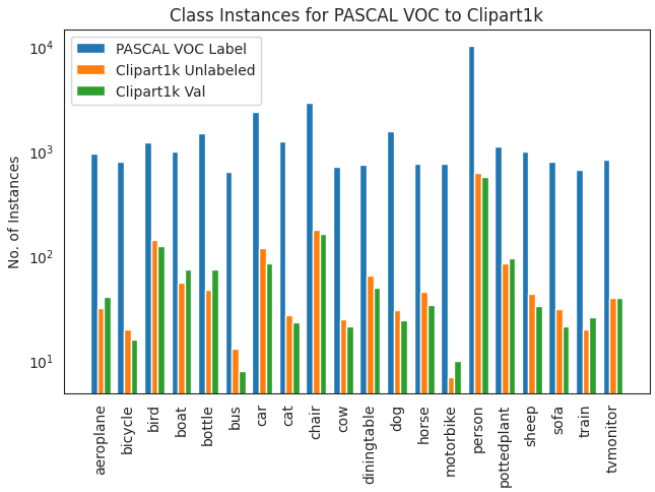

Figure 5. Class distribution of datasets used for the PASCAL VOC → Clipart1k task. Person is a majority class for all datasets, however other classes for PASCAL VOC have a similar number of instance. The imbalance is stronger in the Clipart1K dataset with classes such as motorbike and bus being a minority.

8.3. Cityscapes → BDD100K

In addition to experiments performed in Section 5.3 of the main paper, we include the Cityscapes → BDD100K-Daytime benchmark.

The BDD100K [**?**] dataset is a large-scale dataset containing 100,000 images. For this experiment, we use the day-time split which contains 36,728 training and 5,258 testing images. We remove the train, traffic light and traffic

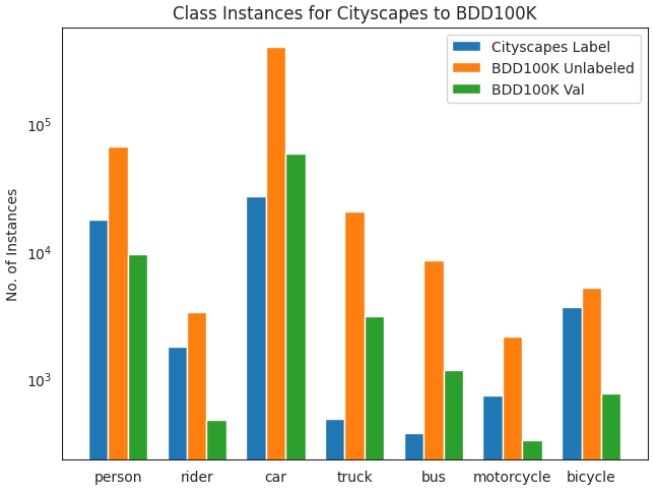

Figure 6. Class distribution of datasets used for the Cityscapes → BDD100K (Daytime) task. We can see that the car class form the majority of all classes, especially for the BDD100K dataset. Note that the class distribution of labeled and validation set differs, especially for minority classes which can make the task more difficult.

sign categories following previous work. The Cityscapes → BDD100K benchmark covers scene adaptation as well as small-to-large dataset adaptation.

Table 7 shows the results of our experiment. We can observe that CAT has stronger performance compared to the previous SOTA at 38.5 mAP. Minority classes such as rider, truck, bus, and bicycle also show a significant improvement. This shows that our strategy to address inter-class dynamics provides a viable solution to address class imbalance for domain adaptive object detection.

Method	person	rider	car	truck	bus	mcycle	bicycle	mAP
Faster RCNN [32]	28.8	25.4	44.1	17.9	16.1	13.9	22.4	24.1
SIGMA [28]	46.9	29.6	64.1	20.2	23.6	17.9	26.3	32.7
TDD [16]	39.6	38.9	53.9	24.1	25.5	24.5	28.8	33.6
PT [4]	40.5	39.9	52.7	25.8	33.8	23.0	28.8	34.9
CAT (Ours)	44.6	**41.5**	61.2	**31.4**	**34.6**	24.4	**31.7**	**38.5**

Table 7. Object detection results on the BDD100k-Daytime test set for **Cityscapes → BDD100k-Daytime domain adaptation**. The mean average precision at .50 IoU (mAP) is reported for all classes.

Hyperparameter	Description	C→F	PV→CA	C→B
-	Detector	FRCNN	FRCNN	FRCNN
-	Backbone	VGG	ResNet-101	VGG
-	BatchNorm	True	True	False
α	Decay rate for student-teacher EMA	0.9996	0.9996	0.9996
β	Beta-distribution parameters for mixup	[0.5,0.5]	[0.5,0.5]	[0.5,0.5]
λ_d	Weight for Adverserial Loss	0.1	0.1	0.1
λ_u	Weight for Unsupervised Loss	1.0	1.0	1.0
τ	Threshold value for pseudo-label confidence	0.8	0.8	0.8
λ_l	Regularization term for Inter-Class Loss	1.0	1.0	1.0
-	Source Augmentation Ratio	0.5	0.5	0.5
-	Target Augmentation Ratio	0.5	0.5	0.5
-	Burn-Up Step Iterations	20000	20000	20000
-	Total Training Iterations	80000	80000	80000
-	Learning Rate	0.2	0.2	0.2

Table 8. Model Hyperparameters for Experiments.From left to right, Cityscapes → Foggy Cityscapes, PASCAL VOC → Clipart1K, and Cityscapes → BDD100K-Day.