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.

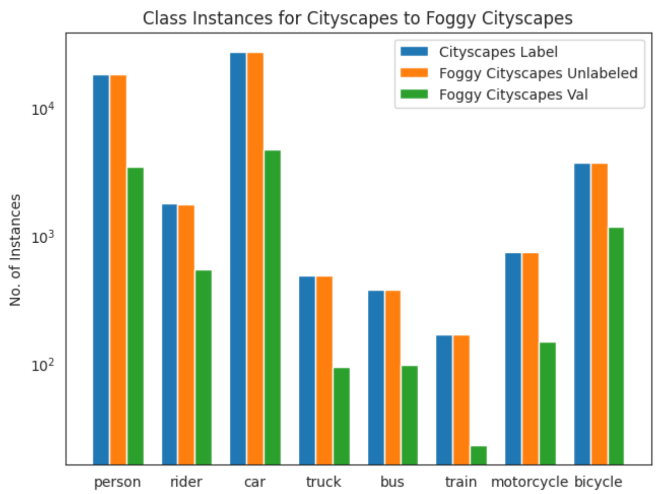

Figure 4. Class distribution of datasets used for the Cityscapes → Foggy Cityscapes task. We can see that person and car classes form the majority of all classes. The distribution of classes for the labeled dataset and validation set is similar which makes for an simpler task.

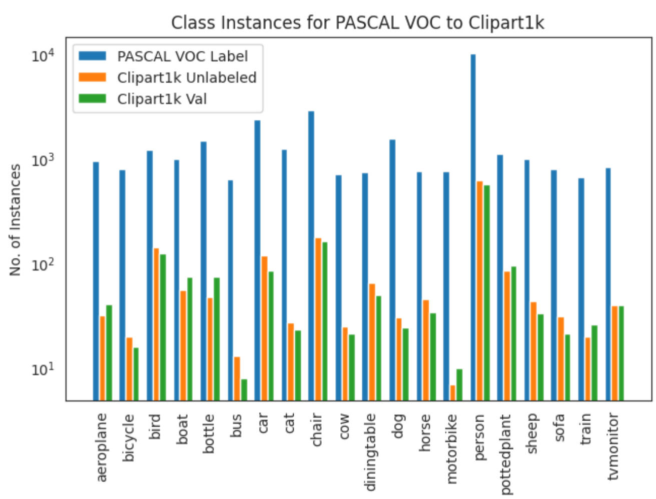

Figure 5. Class distribution of datasets used for the PASCAL VOC → Clipart1k task. Person is a majority class for all datasets, however other classes for PASCAL VOC have a similar number of instance. The imbalance is stronger in the Clipart1K dataset with classes such as motorbike and bus being a minority.

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 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.

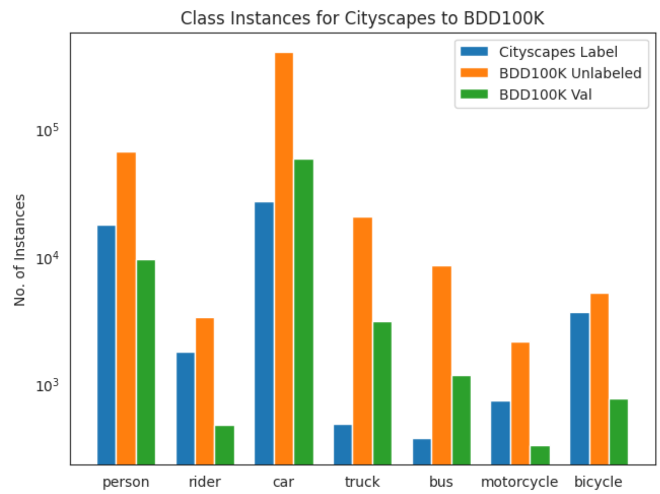

Figure 6. Class distribution of datasets used for the Cityscapes → BDD100K (Daytime) task. We can see that the car class form the majority of all classes, especially for the BDD100K dataset. Note that the class distribution of labeled and validation set differs, especially for minority classes which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

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.