Large Self-Supervised Models Bridge the Gap in Domain Adaptive Object Detection

Supplementary Material

A. Algorithm and Implementation Details

Our proposed algorithm trains the student model in multiple phases for both target domain pseudo-label training and alignment. We present all the training phases here and summarize our method in Algorithm 1. For pseudo-label training, we have two phases:

1. If $iter < n^{initPL}$, train only on source images with ground truth labels.
2. If $iter \geq n^{initPL}$, train on source images with ground truth labels and on target images with DINO labeller labels.

In all of the presented runs, we use $n^{initPL} = 20{,}000$. For alignment, we have 2 phases:

1. If $iter < n^{initSim}$, align student and DINO features only from source images.
2. If $iter \geq n^{initSim}$, align student and DINO features from both source and target images.

In all of the presented runs, we use $n^{initSim} = 5000$. With weak-strong augmentation [3], two versions of each source and target image are generated. We do labelled detection training and DINO alignment with both weakly and strongly augmented source images and target images. We present Algorithm 1 assuming that $n^{initSim} \leq n^{initPL}$, which is true for all the reported runs.

Algorithm 1 Pseudocode for DINO Teacher

while $iter < n^{max}$ **do**
  Input: Source images and labels (X_S, Y_S, B_S)
  Input: Target images X_T
  Augment source images $\mathbf{X}_S = \left[X_S^{weak}, X_S^{strong}\right]$
  Compute source similarity loss $\mathcal{L}_S^{sim}\left(\mathbf{X}_S\right)$
  Compute source detection loss $\mathcal{L}_S^{det}\left(\mathbf{X}_S, Y_S, B_S\right)$
  if $iter \geq n^{initSim}$ **then**
    Augment target images $\mathbf{X}_T = \left[X_T^{weak}, X_T^{strong}\right]$
    Compute target similarity loss $\mathcal{L}_T^{sim}\left(\mathbf{X}_T\right)$
  end if
  if $iter \geq n^{initPL}$ **then**
    Get DINO labels $\left(\tilde{Y}_T, \tilde{B}_T\right)$
    Compute target detect. loss $\mathcal{L}_T^{det}\left(\mathbf{X}_T, \tilde{Y}_T, \tilde{B}_T\right)$
  end if
  Sum all losses and update student model
end while

Method	n^{initPL}	mAP
$\mathcal{L}^{dis}$	28.5	46.8
$\mathcal{L}^{sim}\mathbf{S}$	31.9	47.4
$\mathcal{L}^{sim}\mathbf{B}$	33.0	47.8
$\mathcal{L}^{sim}\mathbf{L}$	32.3	48.3

Table 7. **Ablation of the size of DINOv2 backbone model for the alignment teacher.** We consider ViT-S, B and L for the alignment loss ($\mathcal{L}^{sim}\mathbf{S}$, $\mathcal{L}^{sim}\mathbf{B}$, $\mathcal{L}^{sim}\mathbf{L}$).

B. Additional Ablations and Results

B.1. Alignment with Larger ViT-L

In our ablations in the main paper in Tab. 5b, we present results for alignment with a ViT-S and ViT-B DINOv2 backbone and discuss in Sec. 3.3 that using larger models during online training was too time consuming. We add the ViT-L result in Tab. 7 and show the continuing trend of larger alignment targets leading to improvements in final performance. This shows the potential of aligning with larger models if the training cost could be reduced. One possible solution could be to precompute and store the DINO features for each image in the target dataset, thus only requiring a single forward pass on the dataset for the alignment target. This single forward pass could even be combined with the forward pass of the DINO labeller used to generate the target pseudo-labels. However, this would have significant data storage requirements for the larger datasets.

B.2. Ablation on the Choice of Student Backbone

Our main results in Sec. 4.3 use VGG16 as the backbone for domain adaptation to Foggy Cityscapes and to BDD100k, following previous works. However, we chose ResNet-50 for our experiments on ACDC as this is a more common architecture. Similarly, the small ViT-S architecture is of similar size to VGG and is generally a stronger baseline, particularly when using better initializations such as the distilled DINOv2 weights.

We present additional results on the Cityscapes $\rightarrow$ BDD100k test case with a ResNet-50 backbone and a ViT-S backbone for 3 settings: source only (SO), self-generated pseudo labels with Mean Teacher (MT) and our DINO Teacher (DT) in Tab. 8. We use the same training protocol for ResNet-50 as our tests on ACDC described in Sec. 4.2 and use the same ViT-G generated labels on BDD100k as in our original test. We consider cases where the ViT-S backbone is frozen and unfrozen. When unfrozen, the ViT-

Backbone	State	Method	mAP
VGG16	Unfr.	SO	29.1
		MT	30.1
		DT	47.8
ResNet-50	Unfr.	SO	37.9
		MT	42.7
		DT	52.5
ViT-S	Frozen	SO	37.3
		MT	38.2
		DT	43.2
ViT-S	Unfr.	SO	43.9
		MT	47.1
		DT	53.9

Table 8. **Ablation of the choice of student backbone model.** We maintain significant improvements on ResNet-50 and ViT-S.

S backbone learning rate is scaled down by a factor of 0.01 compared to the detector head learning rate.

For all settings, the better architectures lead to improved performance compared to the VGG results. Moreover, our DINO Teacher remains significantly better ($+5\%$) compared to Mean Teacher for all tested architectures. Even with the stronger student models, we see large improvements when using labels from the external labeller instead of self-labelling.

B.3. Ablation on Choice of Detector

We use single-scale (SS) Faster R-CNN in all our results, but some works used multi-scale Faster R-CNN (FR) with a feature pyramid (FPN) or single-stage methods like FCOS. We provide results for both in Tab. 9, comparing source only (SO) and the self-generated pseudo-labels from Mean Teacher (MT) to our DINO Teacher (DT). Again, we see that our DINO Teacher approach gives substantial improvements for all tested detectors. Moreover, there is a larger improvement ($+1.0\%$ to $+2.8\%$) when going from single-scale to multi-scale Faster R-CNN for Mean Teacher compared to source only, highlighting that the performance of Mean Teacher self-labelling approaches is sensitive to the source-only model performance from which they are initialized. We observe similar trends when considering the changes in backbones in Tab. 8. We are unable to generate good Mean Teacher pseudo-labels for the single-stage FCOS approach, and find that the box confidence scores are generally much lower than those of the two-stage Faster R-CNN methods. We test multiple bounding box confidence thresholds from 0.3 to 0.8 for pseudo-label selection but all led to rapid performance drop on both source and target datasets, meaning best performance occurs prior to using any pseudo-labels.

Detector	Features	Method	mAP
FR	SS	SO	29.1
		MT	30.1
		DT	47.8
FR	FPN	SO	32.7
		MT	35.5
		DT	51.2
FCOS	FPN	SO	33.1
		MT	33.1*
		DT	51.8

Table 9. **Ablation of the choice of student model detector.** We maintain significant improvements across detectors. *We do not obtain good self-generated pseudo-labels with MT on FCOS, and performance collapses rapidly when using the generated labels.

Backbone	State	Labeller	Student
ViT-L	Frozen	45.7	46.9
	Unfr.	50.0	47.2
ViT-G	Frozen	51.1	47.8

Table 10. **Ablation on unfreezing the labeller backbone.** Performance of the labeller and the student improve when unfrozen, but the transfer is less effective compared to using a larger frozen backbone.

B.4. Ablation on Unfreezing the Labeller Backbone

In the main paper, we consider the simplest setting in which the labeller ViT backbone is kept frozen, maintaining the pretrained DINOv2 weights, and present ablations on the effects of labeller backbone size and performance in Tab. 5a. Here, we consider the case where the labeller backbone is unfrozen and is trained with the detector on source images. We generally follow the training regiment from Sec. 4.2 when training, but downscale the learning rate of the backbone by a factor of 0.01. We present the performance of the labeller trained on Cityscapes images on the unseen target BDD and the performance of the VGG student trained with the labeller pseudo-labels on BDD in Tab. 10. We see that while unfreezing the backbone leads to significant improvements on the labeller ($+4.3\%$), these are smaller for the student ($+0.3\%$). We find that the improvements in labeller performance due to unfreezing do not transfer to the student as well as those from using a larger model.

B.5. Ablation on Using EMA Teacher Labels

Our method differs from the Mean Teacher baseline by generating target pseudo-labels with our DINO labeller instead of using the EMA teacher, leading to better performance. However, in many cases (smaller labeller in Tab. 5b, ACDC fog and snow splits, larger student backbone and detectors), the teacher model derived from the student exposed to target data performed better than the labeller trained only on

Method	mAP
MT	35.3
EMA only	43.3*
EMA mixed	47.3
DT	47.8

Table 11. **Ablation of using EMA teacher pseudo-labels.** *Best performance occurred at 25,000 iterations, just before switching to using only EMA teacher pseudo-labels.

source data. Thus, we investigate whether EMA teacher labels could be useful in addition to using the DINO labeller. We assume that one of the most significant advantages of the DINO labels is that they are more accurate for the first few iterations of training on the target data, after which the EMA teacher is adapted to the target domain and could be good enough to generate useful labels.

We propose to add a third phase to pseudo-label training (see Appendix A above): after $n^{initEMA}$ iterations, we start using EMA teacher pseudo-labels following the regular Mean Teacher strategy. We propose two variations: use only EMA teacher labels after $n^{initEMA}$ (EMA only), or alternate between DINO ViT-G and EMA labels in alternating iterations (EMA mixed). The second variation assumes that there might be some advantage in using the EMA labels but tries to avoid any potential issue of drift from biased EMA teacher labels by still using the DINO labels.

We present results in Tab. 11 of these variations on transfer to BDD with a VGG backbone and compare to the nominal DINO Teacher (DT) and Mean Teacher with DINO alignment but no labels (MT, corresponds to case 1 in Tab. 4). We follow the training protocol from Sec. 4.2 and use $n^{initEMA} = 25{,}000$. For both variations, the best performance was worse than the nominal DT but better than MT. Notably, the EMA only case led to a monotonic reduction in performance from the initial state at $n^{initEMA} = 25{,}000$ iterations, meaning the EMA pseudo-labels made performance worse after the initial training phase with DINO labels. This highlights the potential risk of using pseudo-labels and the importance of more robust generation.

B.6. Issues with Self-Labelling Approaches

When looking at our results comparing self-generated labels with using an external source of labels, we generally find that performance is not only worse when self-labeling but it can be unstable: after some initial improvements compared to source-only training, performance on the target data can drop. This is most striking in our ablations in Tab. 11, where even when initialized with the stronger DINO Teacher labels, using only Mean Teacher labels after $n^{initEMA}$ leads to a reduction in performance over time. We believe this can be explained by class confusion in the pseudo-labels, particularly for rare classes, which lead to degraded class representation over time.

To explore this, Fig. 4 presents the ratio of pseudo-labels with class confidence values above the threshold of $\delta = 0.8$ (and thus kept as labels) compared to the number of real instances of a given class on the target BDD dataset for three labellers: our ViT-G DINO labeller, source-only VGG student at $n^{initPL} = 20{,}000$ iterations (when Mean Teacher pseudo-label generation begins), and the Mean Teacher EMA teacher at $40{,}000$ iterations. We observe that the source-only student is a poor labeller for rare classes like Truck, Bus and Motorcycle, and this poor initial performance leads to even worse labels as training continues. We see an increase in confident boxes of common classes like car that are matched to rare class boxes (orange lines for truck and bus) or that do not match any ground truth instance (green bar on car).

B.7. Results on BDD Daytime-Sunny to BDD Night-Sunny

We followed existing domain adaptation works in considering domain adaptation from Cityscapes to BDD100k Daytime in our main results in Tab. 1. However, the BDD100k [6] dataset contains many images that are not daytime scenes, including many night scenes. A recent work in domain generalization by Wu and Deng [5] proposes a new dataset of multiple weather conditions composed mainly of BDD100k images. It contains five splits: daytime-sunny (day and without significant weather) from BDD100k, night-sunny also from BDD100k, the smaller dusk-rainy and night-rainy rendered from BDD100k images, and daytime-foggy composed of Foggy Cityscapes [4] and Adverse-Weather [1] images. The proposed daytime-sunny split differs from the Daytime split we present in Sec. 4.1.

We consider the daytime-sunny and night-sunny splits, which have a similar number of images, with 19,395 training and 8313 testing images for daytime-sunny and 18,310 training and 7848 testing images for night-sunny. This test case is similar to Cityscapes $\rightarrow$ Foggy Cityscapes except for larger datasets and a non-synthetic domain gap. We use the training protocol described in Sec. 4.2 for a VGG16 backbone. Because we use a new source dataset, we train a new ViT-L DINO labeller on the daytime-sunny split. We present results for the Adaptive Teacher baseline (AT) and our DINO Teacher (DT) in Tab. 12. Similar to the domain adaptation from Cityscapes to BDD100k Daytime, we find that DINO teacher leads to a substantial improvement of $+9.2\%$ on the target dataset compared to Adaptive Teacher.

B.8. Complete ACDC results

We present the full per-class AP and mAP for the ACDC runs that are averaged to obtain the results of Tab. 3 in the paper. Tab. 13 presents results from our reimplementation

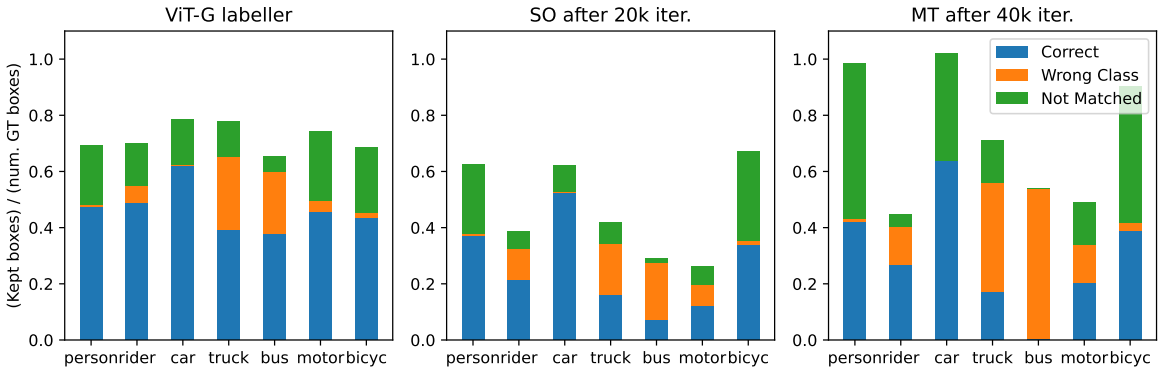

Figure 4. **Quality of generated pseudo-labels.** Ratio of number of high-confidence pseudo-labels compared to the total number of instances per class. The student model (SO and MT) is much weaker for the rare classes, and as training progresses Mean Teacher pseudo-labels, the label quality becomes worse.

Method	mAP
AT	34.9
DT	44.1

Table 12. **Results for domain adaptive object detection from BDD daytime-sunny to BDD night-sunny.** We maintain significant improvements on this task.

of Adaptive Teacher, and Tab. 14 presents our DINO teacher results. We show that our DINO Teacher consistently improves on the Adaptive Teacher baseline. However, both the baseline and our proposed method struggle with certain rare classes in the hardest night and rain splits, specifically Rider and Bicycle. This is not seen on the easier fog and snow splits or when adapting to BDD100k Daytime (Tab. 1) or to Foggy Cityscapes (Tab. 2). This could be because of limited labels, overlapping boxes between Rider and Bicycle instances, or Rider instances being incorrectly pseudo-labelled as Person, all of which could cause training issues.

C. Qualitative Results

We present qualitative results in Fig. 5 that compare our approach with the baseline Adaptive Teacher for the transfer to BDD100k, Foggy Cityscapes and ACDC Night. In general, our method performs better for rare classes like trucks and trains and can be better in complex scenarios with overlapping objects. We also generate fewer false positives from wrong classes.

References

[1] Mahmoud Hassaballah, Mourad A Kenk, Khan Muhammad, and Shervin Minaee. Vehicle detection and tracking in adverse weather using a deep learning framework. *IEEE transactions on intelligent transportation systems*, 22(7):4230–4242, 2021. 3

[2] 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 *CVPR*, pages 7581–7590, 2022. 5

[3] 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 *ICLR*, 2021. 1

[4] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *IJCV*, 126: 973–992, 2018. 3

[5] Aming Wu and Cheng Deng. Single-domain generalized object detection in urban scene via cyclic-disentangled self-distillation. In *CVPR*, pages 847–856, 2022. 3

[6] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. BDD100k: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, pages 2636–2645, 2020. 3

Run		Person	Rider	Car	Truck	Bus	Train	Motor	Bicycle	mAP	Avg	Stdev
Fog	1	62.6	67.2	83.0	26.1	70.3	66.3	53.3	65.4	61.8		
	2	67.4	62.3	86.2	42.3	60.3	66.3	47.0	66.2	62.3	62.2	0.4
	3	64.9	64.1	85.4	35.2	68.6	66.3	50.7	65.6	62.6		
Night	1	35.6	28.0	57.5	18.5	-†	34.0	19.5	21.8	30.7		
	2	36.2	17.0	56.8	32.3	-†	30.6	16.9	15.2	29.3	29.5	1.1
	3	34.9	20.3	59.5	38.7	-†	11.6	15.2	20.2	28.6		
Rain	1	45.5	4.0	76.6	58.4	37.6	21.0	50.2	4.3	37.2		
	2	45.6	3.8	76.0	45.1	36.5	44.9	43.6	5.9	37.7	37.0	0.9
	3	40.5	19.7	76.5	37.5	37.7	14.2	61.4	0.4	36.0		
Snow	1	51.8	52.3	77.8	56.3	21.3	61.9	71.0	49.9	55.3		
	2	53.3	64.6	78.9	56.2	20.2	61.1	70.8	44.4	56.2	55.2	1.0
	3	50.5	56.0	78.6	49.5	27.2	54.5	71.8	45.8	54.2		

Table 13. **Full results, Cityscapes to ACDC splits, Adaptive Teacher [2] baseline.** †There are no labels for the Bus class in the night validation split.

Run		Person	Rider	Car	Truck	Bus	Train	Motor	Bicycle	mAP	Avg	Stdev
Fog	1	73.4	70.2	85.6	36.8	76.2	77.6	45.8	66.0	66.4		
	2	66.4	75.1	85.2	40.9	89.2	80.8	56.4	72.3	70.8	68.6	2.2
	3	67.1	67.9	85.4	41.7	100.0	80.8	44.4	62.3	68.7		
Night	1	38.3	32.0	58.2	34.9	-†	40.6	32.8	20.1	36.7		
	2	41.4	27.5	58.8	37.6	-†	50.6	23.2	16.8	36.5	36.4	0.4
	3	38.9	24.4	57.8	35.3	-†	45.0	32.5	17.8	36.0		
Rain	1	49.3	6.4	79.9	49.8	40.0	26.2	56.3	6.5	39.3		
	2	51.3	7.9	78.1	41.0	37.8	30.9	58.6	0.7	38.3	39.0	0.6
	3	52.3	7.7	79.2	48.5	38.1	25.9	57.6	6.0	39.4		
Snow	1	50.9	70.3	78.4	55.0	20.1	62.0	60.3	60.4	57.2		
	2	50.5	70.3	77.4	53.0	25.2	57.2	59.0	60.6	56.7	56.9	0.3
	3	55.4	64.6	76.6	55.9	20.2	66.5	63.9	50.5	56.7		

Table 14. **Full results, Cityscapes to ACDC splits, proposed DINO Teacher.** †There are no labels for the Bus class in the night validation split.

Adaptive Teacher DINO Teacher

Figure 5. **Qualitative results on target domain.** We compare Adaptive Teacher (left) to our DINO Teacher (left) on BDD (rows 1 and 2), Foggy Cityscapes (rows 3 and 4) and ACDC Night (rows 5 and 6). **Green**, **Yellow**, **Orange** and **Red** indicate true positive, low-confidence positives, false positive, and false negatives respectively. We use a threshold of 0.7 for true positives and false positives.