Appendix

A. Notations

The notations and their corresponding descriptions used in the main paper are presented in Tab. 5.

B. Discussion of Unsupervised Calibration

It is extremely challenging for unsupervised domain adaptation (UDA) models to provide calibrated results due to the lack of labels in the target domain and semantics shift between the source and target domains. Despite its significance, miscalibrated UDA remains largely under-explored. PseudoCal [18] provides a post-hoc solution for miscalibrated UDA through inference-stage mixup synthesis, which aim to turn unsupervised UDA into supervised one. Specifically, it first generates a set of pseudo labeled target set by taking convex combinations of multiple pairs of real target samples and their pseudo labels. Then they perform supervised calibration such as temperature scaling [12] based on the pseudo-labeled target set.

However, this pseudo calibration method requires access to unlabeled target dataset, making it not directly applicable to FTTA where only a single test image is available. Instead, we propose a unsupervised calibration method for FTTA by utilizing the predictive probability vector. Without requiring access to labeled target dataset, we design a instance-specific weight based on the divergence between the largest logit and the second largest logit to calibrate the pre-trained source model. Despite the limited resources, we have achieved an improvement of ECE on DWD dataset of 6.3%.

C. Experimental Details

Implementation details. The default prompt fine-tuning step is set to 10. We increase tuning steps on Night Clear to 15 and decrease the steps on Day Foggy and Dusk Rainy to 5, base on the data domain similarity and difficulty. In bounding box prediction, we remove box of low maximum confidence, i.e. last 30%. We only average the logit prediction of those box with larger than 65% IoU. The training is conducted with 2 A100.

Art Image Dataset The Art Image dataset is proposed by [19], it includes three types of art images, 1k clipart with 20 classes, 2k watercolor with 6 classes, and 2k comic with 6 classes. We only evaluate the common classes in the three datasets, that is, bike, bird, car, cat, dog, and person.

Comparison baselines Our method is compared with the following methods: (1) without adaptation (w/o adpt): Faster RCNN [34] (FR), GDINO [27]; (2) Unsupervised domain adaptation (UDA): Universal Adaptation Network (UAN) [60] and Calibrated Multiple Uncertainties (CMU) [10], Domain Adaptive Faster RCNN (DAF) [4], Multi-adversarial Faster RCNN (MAF) [14], Asymmetric Triway Faster Rcnn (ATF) [15], Hierarchical Transferability Calibration Network (HTCN) [3], Strong Weak Domain Adaptation (SWDA) [36], Augmented Feature Alignment Network (AFAN) [50], Channel-wise Alignment for Adaptive Object Detection (CAD) [54], Partial Alignment Asymmetric Tri-way Faster RCNN (PAATF) [16], Paradigm Teacher Multi-Adversarial Faster RCNN (PTMAF) [17], Implicit Domain-invariant Faster-RCNN (IDF) [23], Sequence Feature Alignment (SFA) [51]. Universal Scale-Aware Domain Adaptive Faster RCNN (USDAF) [40], Confused and Disentangled Extraction (CODE) [41]. (3) Fully test-time adaptation (FTTA) method:Tent [49], TPT [42], DART [28], and IOUFilter [35]; (4)Calibrated test-time prompt-tuning methods: C-TPT[59], ZS-Norm[30], Penalty[30], SaLS[30], and O-TPT[39].

D. Additional Experiment Results

In this section, we present some additional experimental results and details including some ablation studies in Sec. D.1, some qualitative result for TGIA and its corresponding detection results in Sec. D.2, and the class-specific analysis for the DWD and Art Image datasets in Sec. D.3.

D.1. Ablation Studies

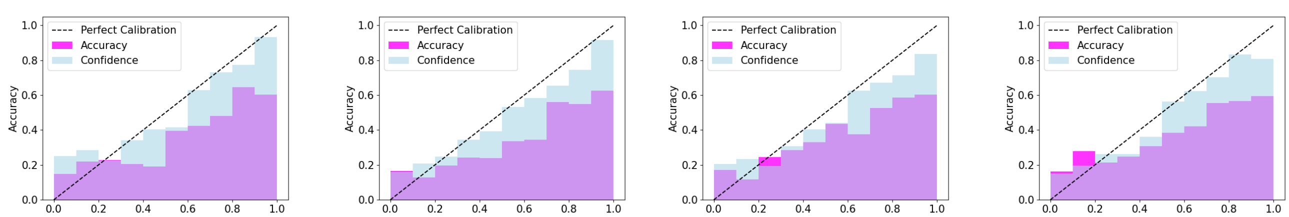
Impact of Hyperparameter α. For Dusk Rainy and Night Rainy data sets, because of the difficulty due to the combination the rainy effect and less lighting and the inborn complexity of unsupervised calibration, the ECE improvement is marginal. For relative simpler Day Foggy dataset where we have better lighting conditions, we achieve prominent improvement through calibration balance hyperparameter fine-tuning. As shown in Figure 6, we tested different α values ranging from 0.1 to 1, and the D-ECE is minimized when $\alpha = 1$.

Calibration metrics. We have tested the calibration performance with other calibration metrics including ACE, and SCE for both the comparative baselines and our method. The results, as shown in evaluated on Day Foggy, show that our method outperforms the baselines across all metrics.

Interpretability. To demonstrate interpretability, we present T-SNE results comparing calibrated test-time prompt tuning (CTPT) and the uncalibrated version in Figure 8. In the left figure, Truck, Car, and Person are blended, whereas in the right figure, they are clearly separated, highlighting the effectiveness of CTPT.

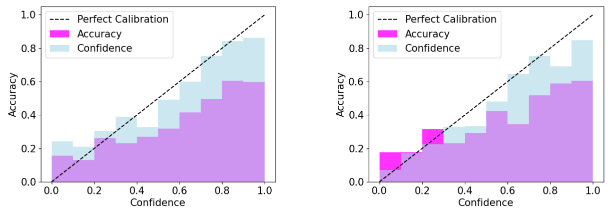
Table 5. Notations

Notation	Description
$\mathcal{L}_{cls}$	classification loss of object detection
$\mathcal{L}_{reg}$	regression loss of object detection
$\mathbf{x}$	image
y	ground truth class label
$\mathbf{b}$	ground truth bounding box coordinates
s_{conf}	confidence score
$\hat{y}$	predicted class label
$\hat{\mathbf{b}}$	predicted bounding box coordinates
$\hat{s}_{\text{conf}}$	predicted confidence score
γ	IoU threshold
ECE	expected calibration error
D-ECE	detection expected calibration error
$I(m)$	the set of all samples in bin m
$\text{prec}(m)$	average precision in bin m
$\text{conf}(m)$	average confidence in bin m
$\mathcal{D}_s$	s-th source dataset
$\mathbf{x}_{\text{test}}$	test image from the target dataset
ENC_I^s	pre-trained s-th source image encoder
ENC_T	pre-trained s-th text encoder
S	number of source domains
$\mathbf{p}$	learnable test-time prompt
θ	weight parameters for TGIA
$\mathcal{A}_\theta(\cdot)$	TGIA augmentation
$\mathcal{A}(\cdot)$	TGIA augmentation
src_{sty}	source style text
tgt_{sty}	target style text
$\mathbf{z}$	image patch embedding
ΔI	difference between augmented image embedding and original image embedding
ΔT	difference between source style and target style
λ	hyperparameter
$SIM(\cdot)$	cosine similarity
N	number of augmentations for each view
ρ	cutoff percentile of augmented views
y^{1st}	class with highest prediction normalized logit
y^{2nd}	class with second highest prediction normalized logit
p^{1st}	highest prediction normalized logit
p^{2nd}	second highest prediction normalized logit



(a) α=0.1 (D-ECE: 14.54%) (b) α=0.5 (D-ECE: 14.07%) (c) α=0.8 (D-ECE: 10.64%) (d) α=1.0 (D-ECE: 12.80%)

Figure 6. Comparison of TPT fine-tuned performance on Day Foggy from the DWD dataset [53] with different calibration balance factors α in training.



(a) Day Foggy w/o calibration. (D-ECE: 15.32%)

(b) Day Foggy w/ calibration. (D-ECE: 12.80%)

Figure 7. Comparison of TPT fine-tuned on Day Foggy from the DWD dataset [53] w/ and w/o calibration loss in training.

Table 6. Calibration performance by ACE and SCE.

Methods	**ACE** (%)	**SCE** (%)
SED	26.2	25.5
MSMT	27.5	26.8
MixUp	27.9	27.2
HCL	27.8	27.1
IRG	30.5	29.7
GDINO	22.0	21.2
Tent	23.1	22.2
TPT	23.2	22.0
C-TPT	22.9	21.6
ZS-Norm	23.0	21.6
Penalty	23.1	21.7
SaLS	23.2	21.8
O-TPT	23.4	21.9
DART	20.9	19.8
IOUFilter	34.3	33.5
InsCal(Ours)	**18.5**	**17.8**

Aggregation of multiple source models. In this ablation, we study the aggregation of multiple pre-trained source models. Results show that directly aggregate information from multiple source models is not always useful, sometimes may even have a negative impact on the adaptation to the target domain. As shown in Figure 9, we use different combinations of source models, and adapt to the unseen target models on DWD dataset. The results on Night Clear show that using Night Rainy achieve best performance, while incorporating other domains such as Dusk Rainy or Day Foggy harm the performance. Similar results can be observed from Day Foggy and Dusk Rainy domains. For Night Rainy, incorporating sources such as Dusk Rainy, Day Foggy improve the performance. The reason is that, Night Rainy show similar semantic with Night Clear. For example, they are both night images. Using Night Rainy as source will lead to a good performance on Night Clear dataset. However, there is a huge domain gap between other domains and Night Clear, including Dusk Rainy and Day Foggy. Incorporation of these domain as sources will induce negative transfer, further degrade the transferring from Night Rainy to Night Clear.

Necessity of multi-source TPT (3TZe, WnvT). We clarify that no source model is fine-tuned; only lightweight prompt parameters are updated at test time under the FTTA setting. Large VLM detectors are pre-trained on heterogeneous, implicit domain mixtures, and treating them as a single source conflates incompatible priors, which exacerbates miscalibration under domain shift. Multi-source TPT preserves source-specific priors and adapts them via prompt tuning, allowing complementary knowledge to be selectively exploited. Multi-source TPT is not equivalent to model ensembling: ensembling aggregates predictions post hoc without adapting to the target distribution, whereas our method performs joint test-time prompt adaptation across multiple sources using unlabeled target data, enabling source-aware, instance-level adaptation. This setting is practically motivated in FTTA, where adaptation resources are limited but multiple pre-trained models are often available. Ablations in Tab. 7 presents the performance of adapting a single source model, it show that adding sources consistently improves both accuracy and calibration, confirming the gains are not due to simple aggregation.

Table 7. mAP (%) and D-ECE (%) results on DayFoggy.

Source Domain	**mAP**	**D-ECE**
Day Clear	36.5	12.1
Dusk Rainy	35.9	12.5
Night Rainy	35.3	12.7
Night Clear	34.8	12.8
Multi-Source	**37.1**	**10.6**

Computational Complexity Analysis. Tab. 8 presents the time and memory analysis of our method compared to baseline approaches. InsCal incurs a slight increase in inference time and parameter size compared to TPT while outperforming other FTTA methods.

Table 8. Time and Memory Complexity

	Tent	TPT	CTPT	O-TPT	DART	IOUFilter	InsCal
Inference Time (FPS)	**0.18**	0.25	0.25	0.24	0.29	0.28	0.24
Parameter Size (M)	**75**	84	85	85	92	93	85

D.2. Qualitative Results of TGIA

We present some augmented images using TGIA on the art images dataset in Figure 10. In Tab. 9, we presented the text-based style description for the corresponding source domain. We formulate the source domain style description by simply prepending the template “A photo of" to the dataset name, which requires no extra information and no access to the source domain images. As shown in Figure 10, TGIA manages to transfer the source domain styles to the target images while preserving the details of the original content.

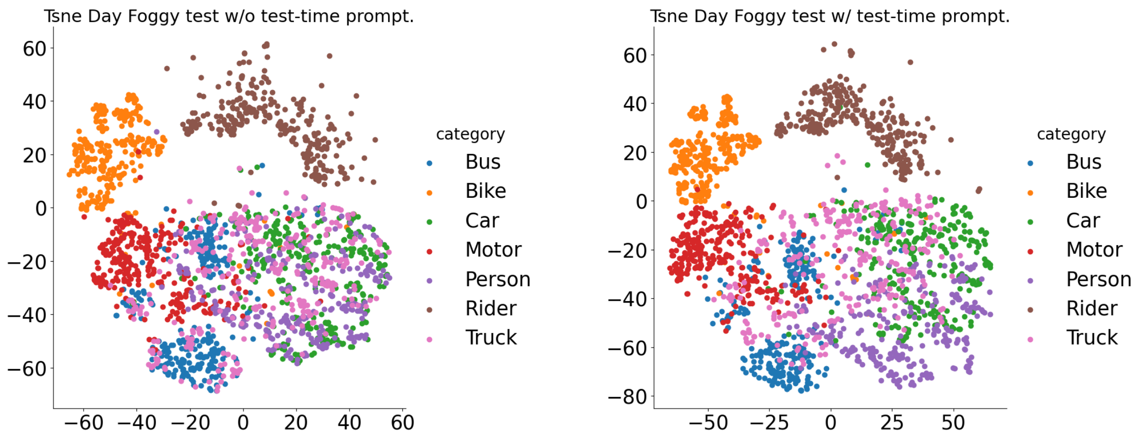
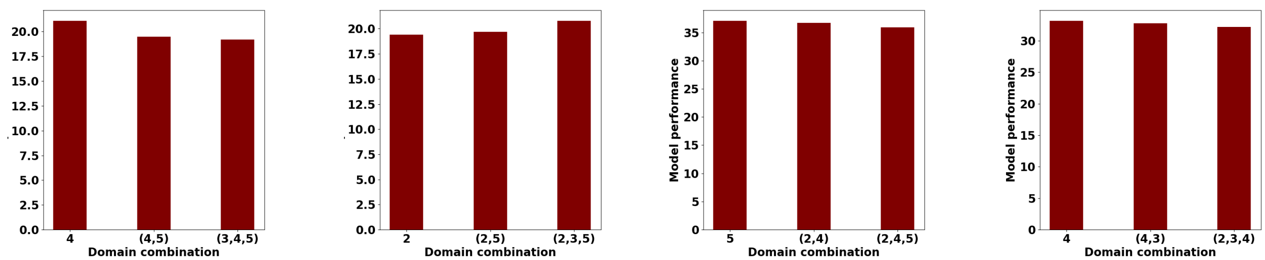


Figure 8. T-SNE results w/ and w/o calibrated test-time prompt.



(a) Night Clear (b) Night Rainy (c) Day Foggy (d) Dusk Rainy

Figure 9. mAP results on DWD domains 9a-9d. In each domain we use different combination of source models as multi-source models, each number stands for a specific domain: 1-Day Clear, 2-Night Clear, 3-Day Foggy, 4-Night Rainy, and 5-Dusk Rainy.

The corresponding detection results are shown in Figure 11, where we present the detection performance using entropy minimization (EM), TPT, multiple source (MS) and our method InsCal. The detection results show that EM and TPT both fail to detect multiple people in the figure. And utilizing multiple source and InsCal improve the detection performance.

Table 9. Source domains and its corresponding descriptions.

Source	Description
Clipart	A photo of clipart
Comic	A photo of comic
Watercolor	A photo of watercolor
Day Foggy	A photo of foggy day

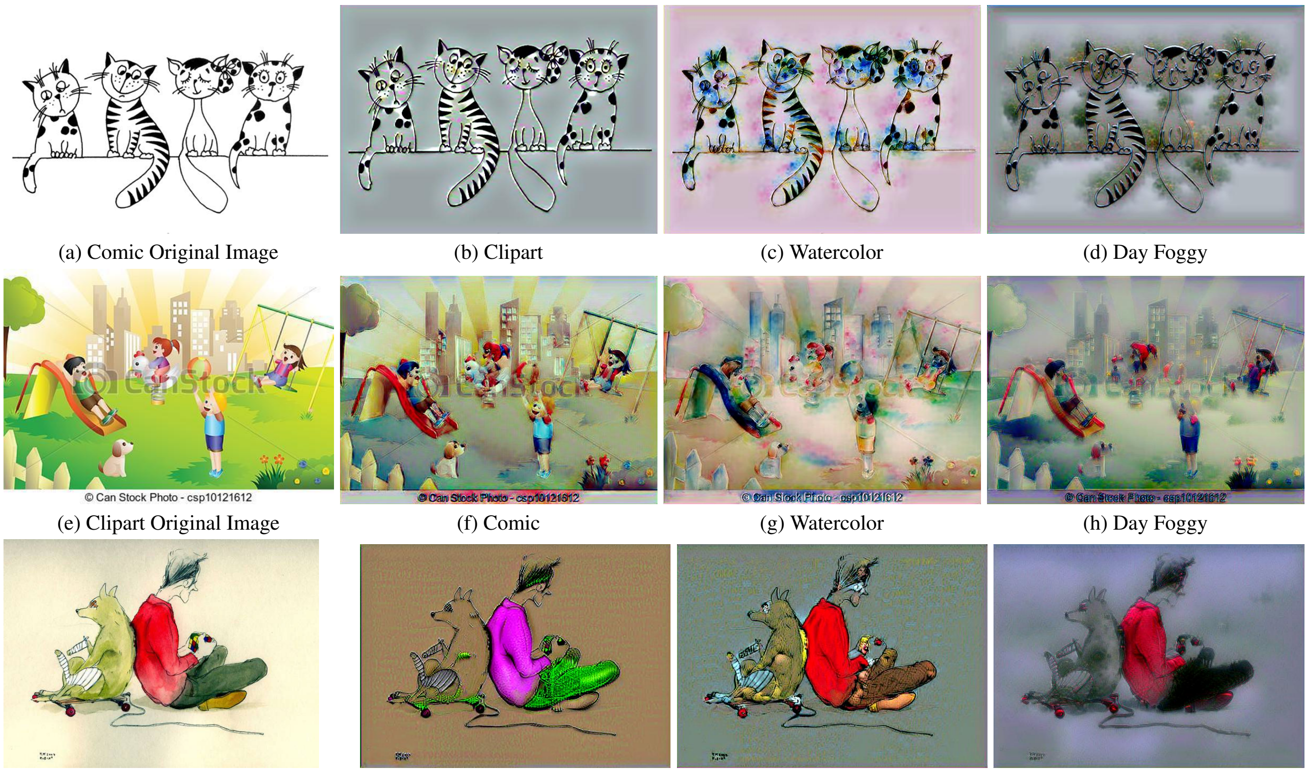
D.3. Class-specific Results Analysis

In this section, we provide the class-specific results analysis for each domain in DWD and the Art Image datasets. Since some UDA and SFDA methods mentioned in Tab. 2 do not provide the details of class-specific results, we remove them from the per-class analysis in this section.

Day Foggy. In Tab. 10, we observe that InsCal has outperform all the FTTA methods in every category. Some categories such as Motor, Person, and Truck achieve the highest performance across UDA, SFDA and FTTA.

Dusk Rainy. As shown in Tab. 11, in UDA, CLIPAug performs better than other methods in every category. For SFDA, MixUp performans consistently better than other methods. In FTTA, our method outperforms other methods in every category. In addition, our methods performs better than CLIPAug and MixUp in most of the categories, showing a better overall mAP across UDA, SFDA and FTTA.

Night Rainy. As presented in Tab. 12, CLIPAug provides consistently better performance than other methods, resulting a 18.7 mAP, outperforming the rest of the UDA methods. The lack of lighting (Night) and the raining effect makes the overall detection on night rainy domain very difficult. In SFDA, MSMT performs consistently better than the others. The utilization of multiple source models and the unlabeled target images lead to the stable performance of 29.1 mAP. In FTTA, our methods outperform the rest of the methods in every category, leading to the highest mAP, which is better

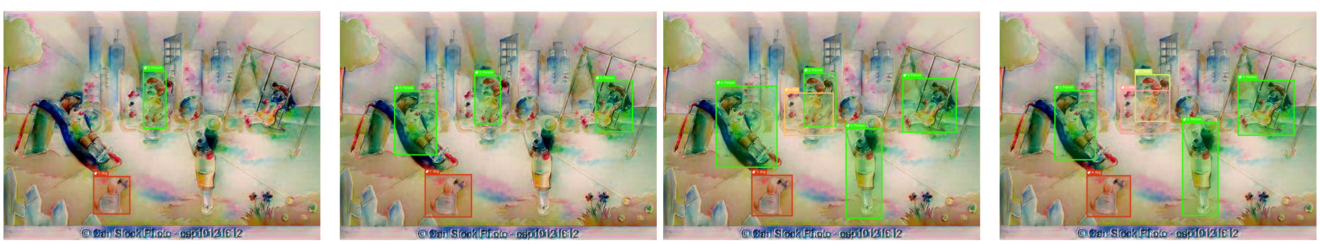


(a) Comic Original Image (b) Clipart (c) Watercolor (d) Day Foggy

(e) Clipart Original Image (f) Comic (g) Watercolor (h) Day Foggy

(i) Watercolor Original Image (j) Clipart (k) Comic (l) Day Foggy

Figure 10. Images from Comic (a), Clipart (e) and Watercolor (i) are transfer to clipart, comic, watercolor and day foggy style with source description in Tab. 9, respectively.



(a) EM (b) TPT (c) MS (d) InsCal

Figure 11. Ablative model components cross-domain detection results on Art Watercolor datasets using different comparative baselines.

than UDA but inferior to SFDA due to the lack of utilization of the unlabeled target images.

Night Clear. In Tab. 13, we present the class-wise results for Night Clear. In UDA, SDGOD and CLIPAug have the best performance in different categories, and their mAP are very close. In SFDA, MSMT performs consistently better than the others. In FTTA, our method performs better than others across all the categories. However, the base model GDINO has difficulty adapting to the night domain, making the overall performance consistently worse than UDA and SFDA.

Class-specific results on Comic In Tab. 14, we present the class-wise AP for Comic. As observed, InsCal performs consistently better across different classes, resulting the highest mAP. GDINO performs better then FR across all categories, thanks to the larger pre-trained datasets. UDA methods perform better on some classes such as bike, car or dog, but InsCal has a better mAP due to the advantages over the rest of the classes. By utilizing multiple source models and calibration, InsCal achieve comparable performance with UDA methods.

Class-specific results on Watercolor In Tab. 15, we present the class-wise results for Watercolor dataset. Similar results can be observed from Tab. 15 for the Watercolor dataset. Showing the consist trend of our method and the

Table 10. Class-wise mAP on Day Foggy domain of DWD dataset.

Class Names	Method	Bus	Bike	Car	Motor	Person	Rider	Truck	mAP
UDA	FR	28.1	29.7	49.7	26.3	33.2	35.5	21.5	32.0
	SDGOD	32.9	28.0	48.8	29.8	32.5	38.2	24.1	33.5
	CLIPAug	**36.1**	**34.3**	**58.0**	**33.1**	**39.0**	**43.9**	**25.1**	**38.5**
SFDA	SED	28.4	29.1	28.5	24.1	33.9	30.4	32.7	29.4
	MSMT	**35.4**	**37.9**	**40.2**	**39.2**	31.5	33.4	32.9	**36.8**
	MixUp	33.2	32.4	33.5	26.8	29.1	35.5	33.2	31.5
	HCL	32.5	31.3	32.1	25.9	28.0	34.2	31.8	30.2
	IRG	33.8	33.9	34.2	36.8	**37.5**	**38.9**	**34.8**	35.2
FTTA	GDINO	33.2	33.4	33.8	35.7	36.9	37.5	33.5	34.1
	Tent	31.0	31.3	31.9	33.7	34.9	35.7	31.6	32.4
	TPT	34.4	33.3	34.2	36.7	37.9	38.8	34.7	34.9
	C-TPT	35.1	33.6	35.5	38.0	39.2	39.1	33.1	35.4
	ZS-Norm	35.7	36.1	38.8	40.3	39.9	40.3	33.9	36.0
	Penalty	36.0	36.4	38.8	40.6	40.3	40.5	33.8	36.2
	SaLS	36.1	36.3	38.6	40.7	40.4	40.7	33.7	36.3
	O-TPT	36.2	36.5	38.9	40.7	40.5	**40.9**	**34.0**	36.5
	DART	28.8	27.2	28.9	31.4	32.6	32.9	29.2	30.1
	IOUFilter	35.9	24.8	25.6	28.7	30.9	30.5	27.5	28.6
	InsCal	**36.5**	**36.8**	**38.8**	**40.7**	**42.4**	39.7	33.7	**37.1**

Table 11. Class-wise mAP on Dusk Rainy domain of DWD dataset.

Type	Method	Bus	Bike	Car	Motor	Person	Rider	Truck	mAP
UDA	FR	28.5	20.3	58.2	6.5	23.4	11.3	33.9	26.0
	SDGOD	37.1	19.6	50.9	13.4	19.7	16.3	40.7	28.2
	CLIPAug	**27.8**	**28.8**	**28.7**	**27.2**	**27.9**	**28.3**	**28.4**	**28.2**
SFDA	SED	13.5	16.3	16.2	16.5	14.4	15.3	14.9	15.4
	MSMT	**34.1**	**33.6**	**31.9**	**31.4**	**31.7**	**32.3**	32.3	**32.0**
	MixUp	29.2	28.9	28.8	29.8	31.0	31.2	**32.9**	30.8
	HCL	26.7	27.2	27.4	27.2	25.9	25.3	26.5	26.9
	IRG	31.2	28.4	28.8	30.9	31.5	28.3	29.7	30.5
FTTA	GDINO	27.1	25.8	29.2	29.4	30.1	31.5	28.5	29.0
	Tent	26.8	25.9	29.5	29.2	27.3	30.4	28.7	28.9
	TPT	28.8	27.6	30.8	31.2	31.5	32.9	30.1	30.5
	C-TPT	29.2	27.8	30.9	31.4	31.6	33.4	30.6	30.8
	ZS-Norm	29.8	28.0	31.2	31.6	31.9	33.7	31.2	31.2
	Penalty	30.0	28.2	31.6	32.0	32.4	34.1	31.4	31.5
	SaLS	29.9	28.3	31.8	31.6	32.5	33.8	31.2	31.4
	O-TPT	30.4	28.8	32.3	31.9	32.7	34.3	31.5	31.8
	DART	25.4	24.7	28.3	28.1	26.4	29.0	27.1	27.4
	IOUFilter	23.5	22.9	26.5	26.4	24.8	27.2	24.9	25.5
	InsCal	**32.9**	**31.7**	**32.5**	**34.3**	**35.6**	**36.5**	**32.8**	**33.2**

comparison between InsCal and other methods. TPT performs well in cat category, but InsCal still performs better on the rest classes, resulting a higher mAP.

Class-specific results on Clipart In Tab. 16, we can draw similar conclusions of the per-class analysis for Clipart dataset as the previous Comic and Watercolor datasets, where InsCal provides consistent and stable performance.

E. Source Code

For source code, please refer to https://github.com/wdr123/InsCal.

Table 12. Class-wise mAP on Night Rainy domain of DWD dataset.

Type	Method	Bus	Bike	Car	Motor	Person	Rider	Truck	mAP
UDA	FR	16.8	6.9	26.3	0.6	11.6	9.4	15.4	12.4
	SDGOD	24.4	11.6	29.5	**9.8**	10.5	**11.4**	19.2	16.6
	CLIPAug	**28.6**	**12.1**	**36.1**	9.2	**12.3**	9.6	**22.9**	**18.7**
SFDA	SED	15.8	14.5	14.2	18.6	6.9	16.5	18.8	15.1
	MSMT	**16.9**	**16.8**	**16.4**	**16.6**	**16.2**	**16.8**	**16.7**	**16.5**
	MixUp	15.6	15.2	15.4	15.8	15.7	15.2	15.3	15.5
	HCL	15.2	14.8	15.0	15.5	15.6	15.7	15.4	15.3
	IRG	15.5	15.5	15.6	16.1	16.3	16.5	15.7	15.8
FTTA	GDINO	12.5	12.3	13.9	14.2	14.5	14.8	13.2	13.6
	Tent	13.7	13.4	14.5	16.3	17.1	17.2	15.0	15.8
	TPT	14.5	14.1	15.8	17.2	18.0	17.8	15.8	16.5
	C-TPT	14.6	14.4	15.9	17.0	17.9	17.9	16.0	16.6
	ZS-Norm	14.8	14.3	16.0	17.2	17.8	17.8	16.1	16.6
	Penalty	14.9	14.6	16.2	17.3	17.9	18.1	16.4	16.8
	SaLS	14.8	14.4	16.3	17.4	17.8	18.0	16.2	16.7
	O-TPT	14.7	14.3	16.2	17.7	17.9	18.2	16.5	16.9
	DART	11.2	11.0	12.5	14.8	15.9	13.7	12.9	13.4
	IOUFilter	10.8	10.5	12.0	14.2	15.5	13.1	12.4	12.7
	InsCal	**21.8**	**22.2**	**21.8**	**22.7**	**25.8**	**23.5**	**20.8**	**20.8**

Table 13. Class-wise mAP on Night Clear domain of DWD dataset.

Type	Method	Bus	Bike	Car	Motor	Person	Rider	Truck	mAP
UDA	FR	34.7	32	56.6	13.6	36.8	27.6	38.6	34.4
	SDGOD	**40.6**	**35.1**	45.7	19.7	34.7	**32.1**	**43.4**	36.6
	CLIPAug	37.7	34.3	**48.0**	**29.2**	**37.6**	28.5	42.9	**36.9**
SFDA	SED	31.9	34.5	33.8	31.2	32.5	34.9	33.7	33.4
	MSMT	**38.2**	**35.8**	**39.2**	**39.0**	**43.2**	**38.1**	**37.0**	**37.7**
	MixUp	35.8	34.2	34.5	34.6	36.0	36.2	34.5	35.0
	HCL	29.4	31.9	31.2	29.5	29.9	32.4	31.2	30.8
	IRG	37.3	35.8	36.4	36.5	37.8	37.9	36.1	36.7
FTTA	GDINO	27.6	26.5	28.8	29.9	30.5	30.4	28.5	29.2
	Tent	29.6	28.5	30.8	31.9	32.5	32.4	30.5	32.2
	TPT	33.6	32.8	32.8	34.4	34.8	34.5	32.8	33.7
	C-TPT	33.7	33.0	33.1	34.9	35.4	34.8	33.3	34.1
	ZS-Norm	34.7	34.1	34.3	36.2	36.3	35.7	34.5	35.2
	Penalty	34.9	34.3	34.8	36.4	36.4	36.1	34.8	35.5
	SaLS	34.8	34.1	34.5	36.2	36.0	36.0	34.5	35.3
	O-TPT	35.6	35.5	35.4	37.3	37.4	37.2	35.6	37.5
	DART	31.4	30.7	32.5	34.5	34.9	34.8	33.0	33.5
	IOUFilter	29.2	28.8	30.3	32.4	33.0	32.6	31.1	31.4
	InsCal	**36.3**	**37.1**	**37.7**	**38.8**	**39.5**	**40.8**	**37.9**	**38.5**

Table 14. Class-specific AP on Comic Dataset.

Category	Methods	bike	bird	car	cat	dog	person	mAP
w/o adpt	FR	39.6	11.3	30.4	12.9	15.4	40.3	25.0
	GDINO	**40.7**	**12.4**	**31.4**	**13.8**	**16.2**	**50.2**	**25.9**
UDA	UAN	41.0	16.0	29.1	8.6	14.4	43.8	25.5
	CMU	36.8	17.8	24.5	18.3	28.9	54.5	30.1
	DAF	32.1	21.3	26.4	12.5	**31.1**	46.2	28.3
	MAF	**43.1**	17.5	24.2	19.4	22.4	49.1	29.3
	HTCN	30.0	13.9	27.7	7.5	26.1	38.4	24.0
	CAD	39.1	24.8	25.8	11.0	22.0	49.9	28.8
	IDF	19.9	20.5	25.8	15.0	22.8	44.6	24.8
	USDAF	39.8	15.9	**38.6**	18.1	26.6	56.5	32.6
	CODE	40.2	**26.9**	29.7	**19.5**	26.6	**59.8**	**33.8**
FTTA	Tent	-	-	-	-	-	-	25.5
	TPT	40.8	12.6	31.5	13.7	16.4	50.1	25.9
	IOUFilter	-	-	-	-	-	-	20.2
	InsCal	**40.7**	**27.4**	**30.2**	**20.1**	**27.3**	**60.3**	**34.3**

Table 15. Class-specific AP on Watercolor

Category	Methods	bike	bird	car	cat	dog	person	mAP
w/o adpt	FR	82.4	51.7	48.4	39.9	30.7	59.2	52.0
	GDINO	**83.1**	**52.5**	**49.4**	**40.9**	**31.7**	**59.9**	**52.8**
UDA	UAN	78.0	53.6	50.4	36.4	35.8	65.6	53.3
	CMU	82.0	53.9	48.6	39.6	33.1	66.0	53.9
	DAF	73.4	51.9	43.1	35.6	28.8	63.1	49.3
	MAF	70.4	50.3	44.3	36.7	30.6	62.9	49.2
	HTCN	74.1	49.8	**51.9**	35.3	**35.3**	66.0	52.1
	CAD	82.3	52.3	49.3	38.1	32.0	62.6	52.8
	IDF	81.4	54.9	46.7	36.6	29.1	66.0	52.5
	USDAF	86.5	54.1	50.0	**43.0**	34.0	63.2	55.2
	CODE	**87.9**	**55.3**	50.7	38.9	34.7	**67.5**	**55.8**
FTTA	Tent	-	-	-	-	-	-	52.5
	TPT	83.2	52.6	49.5	**41.1**	31.9	60.2	53.0
	IOUFilter	-	-	-	-	-	-	35.8
	InsCal	**88.38**	**55.7**	**51.3**	39.45	**35.3**	**68.1**	**56.3**

Table 16. Class-specific AP on Clipart

Category	Methods	bike	bird	car	cat	dog	person	mAP
w/o adpt	FR	-	-	34.7	5.1	8.3	49.6	29.8
	GDINO	56.9	18.6	**34.8**	**5.2**	**8.4**	**50.5**	**30.5**
UDA	UAN	-	-	31.5	8.6	2.4	42.8	30.3
	CMU	-	-	34.7	9.2	7.6	55.7	32.1
	DAF	-	-	35.9	2.3	4.2	59.4	31.3
	MAF	-	-	32.3	11.0	6.7	52.7	32.2
	HTCN	-	-	32.8	11.3	**10.5**	57.9	34.7
	CAD	-	-	35.9	9.8	4.7	56.1	34.2
	IDF	-	-	37.3	16.7	3.7	52.6	32.7
	USDAF	-	-	36.4	17.7	10.3	**62.5**	38.4
	CODE	-	-	**37.7**	**18.4**	8.4	61.7	**39.4**
FTTA	Tent	-	-	-	-	-	-	30.3
	TPT	60.1	18.3	35.5	10.1	8.4	50.7	30.6
	IOUFilter	-	-	-	-	-	-	29.6
	InsCal	**69.4**	**28.2**	**38.1**	**18.9**	**10.0**	**62.4**	**39.9**