Expert-Teacher-Student Collaborative Learning for Domain Adaptive Object Detection

Supplementary Material

A. Alternative Expert Models in ETS Method

We point that ETS provides a conceptual paradigm for how vision foundation models (VFMs) can better guide teacher-student framework learning. Therefore, the selection of expert models is not limited to DINOv3. We also validate Grounding DINO [16] as an alternative expert model.

Zero-Shot Performance of Grounding DINO

Since the official Grounding DINO implementation only provides inference code and releases checkpoints for both Swin-T and Swin-B backbones, (as illustrated in Tab. S1) We directly evaluate their zero-shot performance on the target domain in Tab. S2.

Table S1. Official Grounding DINO released checkpoints.

Name	Backbone	Pre-Training Data	AP on COCO
GDINO-T	Swin-T	O365, Gold, Cap4M	48.4
GDINO-B	Swin-B	COCO, O365, Gold, Cap4M, OpenImage, ODinW-35, RefCOCO	56.7

Table S2. Zero-shot performance of Grounding DINO.

Target Domain	GDINO-T	GDINO-B
Foggy Cityscapes (0.02)	32.8	41.5
BDD100k	32.9	42.9

However, the zero-shot performance of the pre-trained Grounding DINO in the target domain proves inadequate to serve as the expert model, as it underperforms mainstream DAOD methods. We further visualize Grounding DINO's detection results in Fig. S1, revealing three key limitations of its zero-shot application in the target domain without fine-tuning:

- Limited Scenario Adaptability: While demonstrating generalizability across most scenes, it exhibits poor performance in challenging conditions (*e.g.*, missed detections in dense fog in Fig. S1 (a) and (c), false positives in Fig. S1 (d)).
- Tends to Over-detect: The model tends to over-identify fine-grained objects, such as detecting vehicle occupants (Fig. S1 (a) and (d)) or incidental human figures within advertisements (Fig. S1 (b)).

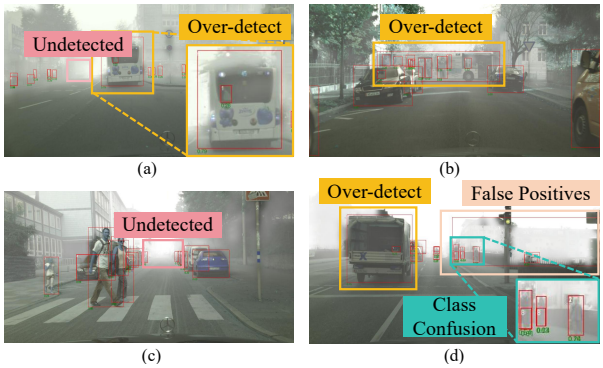

Figure S1. Visualization of Grounding DINO (Swin-T)'s zero-shot detection results reveals three critical limitations: (1) significantly degraded performance in domain-specific regions (particularly dense fog areas), (2) over-detection of irrelevant details (*e.g.*, vehicle drivers and advertisement figures). These observations demonstrate its inadequacy as a teacher model, and (3) fundamental confusion exists between semantically similar categories.

- Semantic Class Confusion: Fundamental confusion exists between semantically similar categories (*e.g.*, classifying both "person" and "rider" classes exclusively as "person").

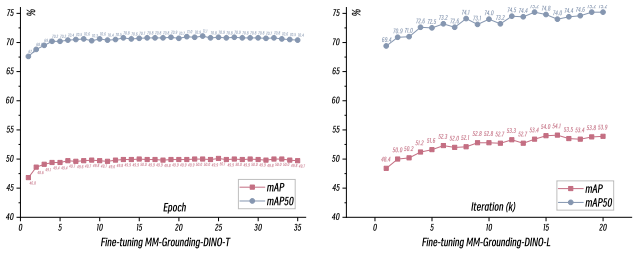

Figure S2. We fine-tune the pre-trained MM Grounding DINO with both Swin-T and Swin-L backbones on the labeled Cityscapes dataset.

Fine-tuning MM Grounding DINO

Fortunately, MM Grounding DINO [21] is proposed to address the lack of technical details in Grounding DINO. As an open-source, comprehensive, and user-friendly baseline built upon the MMDetection toolbox [2]. MM Grounding DINO incorporates abundant vision datasets for pre-training along with various detection and grounding datasets for fine-tuning. The method provides thorough analyses of all reported results and detailed configuration settings to ensure reproducibility.

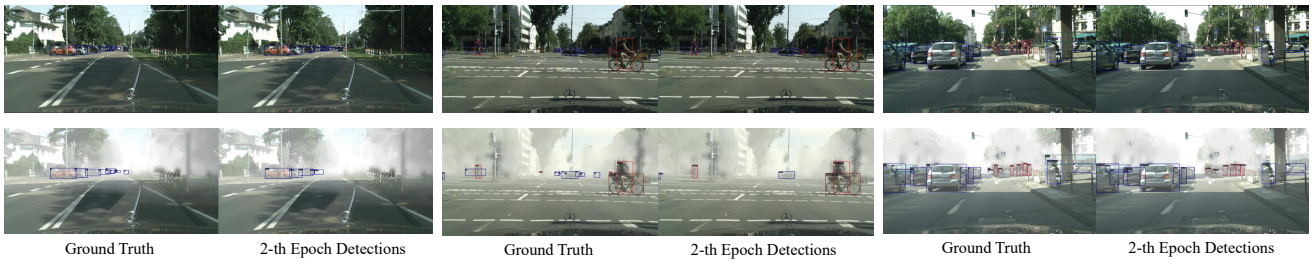

Figure S3. We only exhibit the second training epoch of MM Grounding DINO (Swin-T), significant performance improvements are already observable in the detection results.

Table S3. Source-only performance of fine-tuned MM Grounding DINO-L on Foggy Cityscapes.

Category	mAP_s	mAP_m	mAP_l	mAP50
person	44.2	81.5	97.8	68.7
rider	36.3	81.5	92.5	65.0
car	47.7	89.6	98.4	77.9
truck	1.3	35.5	82.1	47.6
bus	1.8	47.7	93.6	71.4
train	0.0	35.0	76.2	58.9
motorcycle	26.3	70.2	69.3	54.0
bicycle	39.9	76.5	79.1	64.0
Average	25.9	64.7	86.1	63.4

Table S4. Source-only performance of fine-tuned MM Grounding DINO-L on BDD100k.

Category	mAP_s	mAP_m	mAP_l	mAP50
person	38.5	84.2	96.1	63.4
rider	20.2	63.4	89.8	50.5
car	40.3	83.8	95.8	69.9
truck	15.3	42.2	72.4	48.4
bus	13.9	34.8	81.9	49.8
train	-	-	-	-
motorcycle	16.4	54.6	73.2	42.0
bicycle	13.1	52.0	78.6	37.2
Average	22.5	59.3	83.9	51.6

In Fig. S2, we employ pre-trained MM-Grounding-DINO with both Swin-T and Swin-L backbones, fine-tuning them on the labeled Cityscapes dataset. All fine-tuning experiments were conducted using only four RTX 3090 GPUs. We note that employing additional GPUs may further improve performance. In Fig. S3, we visualize the detection results of MM Grounding DINO (Swin-L) after the second fine-tuning epoch, demonstrating that many issues have been resolved: the model now accurately distinguishes between "person" and "rider" while avoiding over-detection of drivers inside vehicles and figures in advertisements. We evaluate the source-only performance of our Cityscapes-finetuned MM Grounding DINO-L on the validation sets of Foggy Cityscapes and BDD100k datasets, as quantitatively demonstrated in Tab. S3 and Tab. S4.

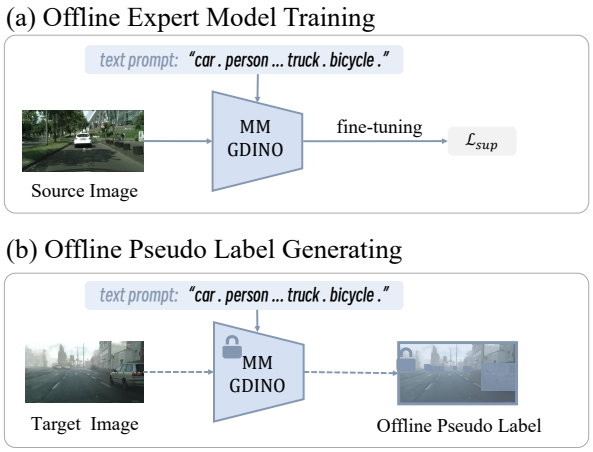

Figure S4. Schematic diagram of fine-tuning MM Grounding DINO as the expert model.

MM Grounding DINO as Expert in ETS Method

Following the same process as DINOv3 expert model implementation, it also requires three steps: (1) Fine-tuning MM Grounding DINO on labeled source domain data, (2) Generating offline pseudo labels using the fine-tuned MM Grounding DINO, and (3) Executing expert-teacher-student collaborative learning, as illustrated in Fig. S4. **Notably**, due to architectural differences between MM Grounding DINO and Faster R-CNN, we do not extract prototypes from MM Grounding DINO. The Expert-Teacher Joint Consolidation (ETJC) module consequently retains the prototypes extracted by DINOv3.

In Tab. S5 bottom, we present the performance of our ETS method using MM Grounding DINO as the expert model, compared with using DINOv2 and DINOv3 as the expert model. Since MM Grounding DINO is a vision-

Table S5. The Comparison of different expert models and they trained ETS method's performance, respectively. Results of Cityscapes to BDD100k. The average precision (AP, %) on all classes is presented.

Method	Backbone	Parameters	person	rider	car	truck	bus	mcycle	bicycle	mAP
Expert (DINOv2)	ViT-B	86M	47.6	42.8	62.9	38.9	38.6	37.3	33.6	43.1
ETS	VGG16	138M	47.8	44.3	64.5	40.8	42.6	31.1	39.4	44.4
Expert (DINOv2)	ViT-L	300M	48.9	45.3	65.5	43.9	41.6	41.1	31.1	45.3
ETS	VGG16	138M	52.5	47.6	67.2	44.5	44.5	37.7	40.2	47.7
Expert (DINOv2)	ViT-G	1,100M	54.3	52.1	67.4	47.1	45.4	47.2	40.2	50.5
ETS	VGG16	138M	53.5	49.5	67.1	47.1	46.3	40.2	43.1	49.5
Expert (DINOv3)	ViT-B	86M	41.9	40.7	60.3	37.8	42.0	35.7	29.5	41.2
ETS	VGG16	138M	46.8	41.6	62.1	38.9	37.8	37.7	32.6	42.5
Expert (DINOv3)	ViT-L	300M	47.0	48.3	64.2	45.8	49.2	50.1	37.2	48.8
ETS	VGG16	138M	51.5	49.4	65.5	46.5	48.9	43.3	37.0	48.9
Expert (DINOv3)	ViT-H+	840M	51.4	52.4	64.6	47.2	51.6	49.8	43.1	51.4
ETS	VGG16	138M	53.9	49.2	67.2	47.1	47.2	40.4	43.5	49.8
Expert (MM GDINO)	Swin-L	341M	63.4	50.5	69.9	48.4	49.8	42.0	37.2	51.6
ETS	VGG16	138M	55.7	50.4	65.0	46.7	52.5	40.3	39.9	50.0

Table S6. Comparison of different expert models and backbones in terms of model parameters, runtime, and memory consumption on Cityscapes→BDD100k.

Expert	Backbone	Parameters	Throughput	GPU Memory	Expert mAP	Ours mAP
DINOv2	ViT-S	21M	7.77 it/s	3.82 GB	35.5	–
DINOv2	ViT-B	86M	4.01 it/s	6.31 GB	43.1	44.4
DINOv2	ViT-L	300M	1.57 it/s	11.40 GB	45.3	47.7
DINOv2	ViT-G	1,100M	0.58 it/s	21.80 GB	50.5	49.5
DINOv3	ViT-S+	29M	8.36 it/s	3.37 GB	36.9	–
DINOv3	ViT-B	86M	4.75it/s	5.26 GB	41.2	42.5
DINOv3	ViT-L	300M	1.99 it/s	8.29 GB	48.8	48.9
DINOv3	ViT-H+	840M	1.02 it/s	15.62 GB	51.4	49.8
MM GDINO	Swin-L	341M	0.72 it/s	10.36 GB	51.6	50.0

language-based vision foundation model specifically designed for object detection tasks, it has fewer parameters than DINOv2 and DINOv3, yet achieves better performance. MM Grounding DINO slightly outperforms the DINOv3-based expert model on most classes, enabling the corresponding ETS method to achieve the optimal performance of 50.0% mAP. This demonstrate that our method is not limited to any specific VFMs. Any advancements in open-source VFMs can promote the progress of our method. When better-performing VFMs become available as expert models, our ETS method can further improve detection performance. **Note that**, following previous baselines and using their codebases [1, 9, 14, 15], our teacher-student architecture is employed Faster RCNN and remains unchanged.

B. Ablation on Different ViT Backbones

As illustrated in Tab. S5, we conduct an ablation study on different ViT backbones within the expert model and compare their parameter counts. The results demonstrate that expert models with larger backbones achieve better performance, thereby transferring more knowledge to the student model in our ETS method. We also find that the performance of DINOv2 ViT-B is 1.9% higher than that of DINOv3 ViT-B under the same parameter count. We find that the expert model is not the performance ceiling of the ETS method. When the backbone of the expert model employs ViT-L or ViT-B, the performance of our student model surpasses that of the expert model. It demonstrates that when the performance of the expert model is insufficient to sup-

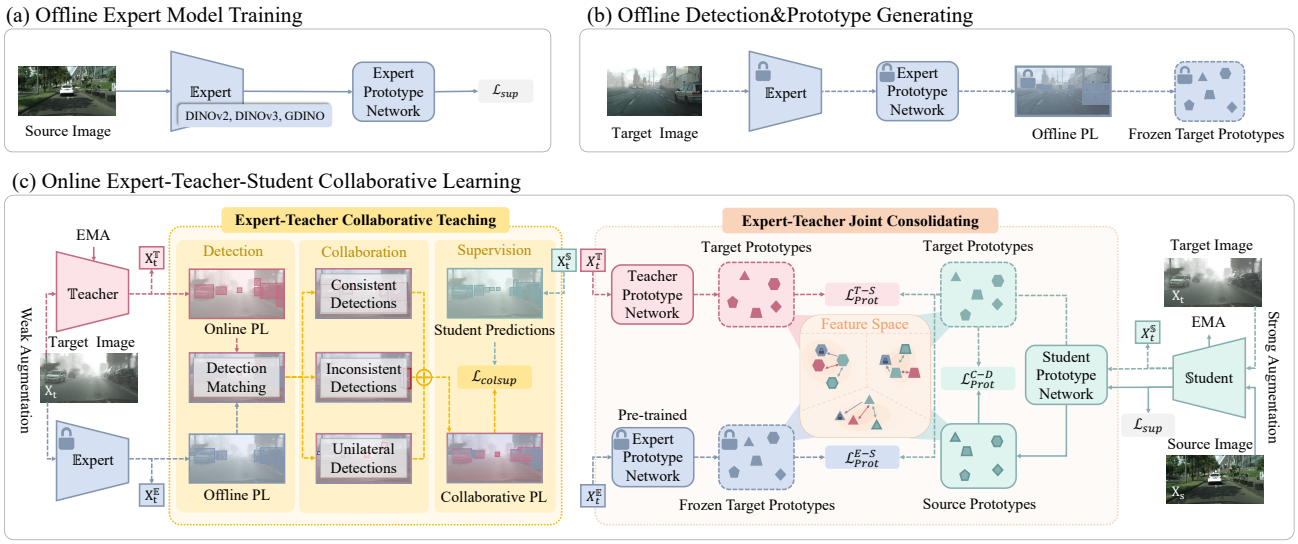

Figure S5. Details of our proposed ETS, which follows a three-step workflow: (a) Offline Expert Model Training, where we first train the expert model using labeled source domain data; (b) Offline Pseudo Label & Prototype Generating, where the trained expert model processes target domain data to generate both pseudo labels and class prototypes through its prototype network; (c) Online Expert-Teacher-Student Collaborative Learning, implementing each-then-consolidate paradigm via two key modules: the Expert-Teacher Collaborative Teaching (ETCT) that produces collaborative pseudo labels by detection matching from both expert and teacher models to supervise the student model's learning, followed by the Expert-Teacher Joint Consolidating (ETJC) module that consolidates the student model's representations by enforcing prototype alignment between all three models. This comprehensive approach ensures effective knowledge transfer while maintaining robust domain adaptation capabilities.

port the learning of the student model for the target domain, the teacher model corrects the negative impact of the expert model and guides the student model to learn the correct representation of the target domain.

C. Details of our proposed ETS

In Fig. 3 of the main paper, we only illustrated the expert-teacher-student collaborative learning process. Here, we present the complete workflow of the ETS method. As shown in Fig. S5, our proposed ETS method follows a three-stage workflow:

1) Offline Expert Model Training: The expert model with a prototype network is pre-trained on labeled source data using supervised learning to learn the fundamental discrimination ability.

2) Offline Pseudo Label&Prototype Generating: The expert generates pseudo labels and class prototypes for the target domain via confidence-weighted and momentum-based prototype update.

3) Online Expert-Teacher-Student Collaborative Learning: which contains ETCT and ETJC two modules, enabling the student model to inherit advantageous knowledge from both teacher and expert models through a progressive strategy of teach-then-consolidate.

D. Implementation Details

Due to the varying scales of different datasets, the implementation details of some parameters may differ. Therefore, we provide detailed hyper-parameters in Tab. S7.

E. More Quantitative Results on BDD100k

Due to space constraints, we could not include more qualitative results in the main text. Here, as shown in Fig. S8, Fig. S9, and Fig. S10, we present more detailed qualitative results on the Cityscapes→BDD100k cross-domain benchmark, comparing with the state-of-the-art method DT [14]. The comparison will be conducted from three key aspects:

Detection Recall Capacity

In Fig. S8, we present a comparative analysis of the recall capacity between our TES and DT. DT demonstrates significantly inferior recall rates when detecting rare categories, particularly for buses (*rows 1, 2, 5*) and trucks (*rows 3-5, 7*). This object omission phenomenon stems from incomplete class-knowledge inheritance without fine-grained prototype alignment, it becomes challenging to transfer the expert model's rich categorical semantic features, especially when confronting substantial domain gaps. Notably, DT also underperforms in detecting common categories. As shown in *rows 3, 6,* and *7* in Fig. S8, it frequently misses "person" instances that our ETS method can reliably identify.

Table S7. Detailed hyper-parameters of ETS for each benchmark

Hyperparameter	Description	C→F	P→Cl	C→B
δ	Standard confidence threshold	0.8	0.8	0.8
δ'	Strict confidence threshold	1.0	1.0	1.0
ϵ	Discrepancy threshold	0.15	0.15	0.15
τ	IoU matching threshold	0.5	0.5	0.5
α	EMA update factor	0.9996	0.9996	0.9996
α'	Prototype EMA update factor	0.999	0.999	0.999
d	Prototype dimension	128	128	128
K	Number of shared categories	8	20	7
λ_1	Supervise loss weight	1.0	1.0	1.0
λ_2	Collaborative teaching loss weight	1.0	0.5	1.0
λ_3	E-S prototype weight	0.1	0.1	0.1
λ_4	T-S prototype weight	1.0	1.0	1.0
λ_5	C-D prototype weight	1.0	1.0	1.0
T_{expert}	Expert pretraining iterations	40k	30k	40k
T_{expert_pro}	Extract expert prototype iterations	20k	10k	20k
T_{burn}	Burn-in iterations	20k	20k	20k
T_{align}	Alignment start iter	25k	20k	25k
T_{max}	Total iterations	100k	30k	100k
Backbone	Teacher/Student	VGG16	ResNet101	VGG16
Expert Backbone	Frozen	DINOv3-H+	DINOv3-H+	DINOv3-H+
Input size (px)	Teacher/Student	600	600	600
Input size (px)	Expert	592	592	592
lr	Learning rate	0.04	0.01	0.04
b	Batch size (source + target)	8+8	12+12	8+8
GPUs	Training devices RTX3090	4	4	4

This comprehensive superiority highlights our method's robust capability in both rare and common category detection through an effective teach-then-consolidate strategy.

Localization&Classification Capacity

As shown in Fig. S9, we present a comprehensive comparison of classification and localization accuracy between our ETS method and DT. DT exhibits several critical misclassifications: (1) mistaking a bench for a "bicycle" (*row 1*), (2) identifying a traffic barrier as a "person" (*row 2*), (3) classifying a billboard as a truck (*row 3*), and (4) detecting reflected buildings as pedestrians (*row 4*). These misjudgments indicate that DT fails to fully utilize the fundamental knowledge contained in the expert model. Merely aligning at the image-level cannot learn fine-grained category semantic knowledge. Meanwhile, the lack of target-domain distillation knowledge from the teacher model results in insufficient pseudo-labels, which cannot provide high-quality supervision, thus leading to these misjudgments. In contrast, our ETS method effectively addresses these issues through multi-model prototype alignment and collaborative teaching optimization, achieving significantly more robust cross-domain detection performance.

Long-distance Detection Capacity

Detecting small targets at long-distance is a challenging task in object detection, especially in the Domain Adaptive Object Detection (DAOD) setting with significant domain gaps, as small targets are more susceptible to domain interference. As shown in Fig. S10 *rows 1-3*, our ETS method performs well in recognizing small and distant cars, while the DT misses most of the distant objects. In the blurry scenario under rainy conditions in *row 4*, DT misses most of the cars, while our method is the least affected.

F. Complementary Advantages of Expert and Teacher Models

In the main text, we have already mentioned that "*VFMs excel at capturing domain-invariant cues, whereas teacher models specialize in domain-specific regions*", but no more observational evidence was provided. Now, we provide additional qualitative analysis of their detection outputs (Fig. S11). As shown in *row 1*, the expert model leverages

Figure S6. We present nighttime scenes from the TDND dataset, which cover real low-light environments such as dusk and night, including various severe weather like heavy rain and snow.

its real-world generalized knowledge to distinguish mirrored reflections from physical objects, a critical safety advantage in autonomous driving scenarios where the teacher model fails by misclassifying reflections. *Row 2* demonstrates the teacher's domain-specific superiority: it reliably detects distant cars obscured by dense fog. The expert model's advantage is that it can identify cars through residual rear features (as shown in the 3-th image of the right side).

The teacher's domain-specific detection capabilities are further evidenced in *rows 3-4* and *rows 6-7* for fog-occluded objects, with these additional detections incorporated as unilateral detections to enrich the collaborative pseudo labels for comprehensive student supervision. Conversely, the expert model shows consistent domain-invariant performance in low-fog regions (*rows 3-5*). Particularly in *row 5*, where domain-specific attributes (fog) are absent, the expert detects severely occluded and small-scale "person" instances (while they missed by the teacher model) with GT-comparable accuracy. This generalizability bridges the source-target domain gap, constituting the core advantage of our ETS method.

G. Pseudo-code

As illustrated in Algorithm 1, we present a pseudo-code pipeline of our ETS. In the diagram, $[\cdot\cdot\cdot\cdot\cdot]$ denotes the offline operations about the expert model of steps (a) and (b), which ensure that no additional computational cost from the expert models is introduced during the online training phase. $[\cdot\cdot\cdot\cdot\cdot]$ represents the steps associated with the student model, whereas $[\cdot\cdot\cdot\cdot\cdot]$ corresponds to those of the teacher model. $[\cdot\cdot\cdot\cdot\cdot]$ and $[\cdot\cdot\cdot\cdot\cdot]$ indicate the procedures related to the ETCT and ETJC modules, respectively.

H. Day-to-Night DAOD Benchmark

Recently, the safety of autonomous driving at night has attracted growing attention from researchers, with domain adaptation emerging as one of the most effective solutions. To further evaluate the domain adaptation capability of our ETS method, we test it under the day-to-night adaptation scenario. It is worth noting that our approach is designed for general cross-domain adaptation rather than being specifically tailored for day-to-night adaptation. This scenario is particularly challenging due to significant variations in lighting conditions, such as low illumination, high dynamic range, and increased noise, which severely impact perception performance. Therefore, it serves as a rigorous benchmark to assess the robustness and effectiveness of our ETS.

Dataset

Following [5], we evaluate our method on the TDND [17] dataset, a large-scale benchmark specifically designed for object detection in challenging nighttime driving scenarios as illustrated in Fig. S6. These images encompass a variety of difficult visual conditions characteristic of nighttime driving, including complex illumination changes, glare from headlights, light refraction, and motion blur. Additionally, the dataset covers severe weather conditions such as heavy rain and snow, which further increases the perception difficulty. In our experiments, we use 1,916 daytime images as the source domain and 7,663 nighttime images as the target domain, with 2,523 nighttime images reserved for validation.

Evaluation

As illustrated in Tab. S8, our ETS demonstrates state-of-the-art performance, achieving a remarkable mAP of 50.9% mAP. This represents a significant improvement of 4.2% over the previous best method, DeT [5] (46.7%), despite that our approach is designed for general cross-domain adaptation rather than being specifically optimized for day-

Table S8. Results of TDND dataset (daytime → nighttime). Where "D→N" denotes the method specifically designed for day-to-night. Note that we demonstrate the performance of four different expert model variants, but we only leverage DINOv3 ViT-H+ as the expert model's backbone in our ETS.

Method	Venues	D→N	car	person	bus	minibus	truck	t-sign	mAP
SADA [3]	IJCV'21	×	72.4	35.7	36.9	14.7	14.9	30.1	34.1
MIC [10]	CVPR'23	×	79.6	28.0	42.4	17.2	22.9	27.5	36.3
2PCNet [13]	CVPR'23	✓	77.4	39.3	52.5	8.4	10.5	25.9	35.6
ISP-Teacher [20]	AAAI'24	✓	72.3	44.5	37.5	15.4	27.0	26.2	37.2
SOCCER [4]	MM'24	×	73.9	17.4	51.4	16.2	38.7	26.5	37.4
CoS [12]	ICME'24	✓	74.8	49.3	53.1	17.6	27.2	31.2	42.2
DeT [5]	ICCV'25	✓	79.8	48.3	57.6	20.9	40.8	32.8	46.7
ETS	-	×	82.3	45.6	67.1	26.1	47.0	37.6	50.9
Expert|DINOv3$_{\text{ViT-S+}}$	-	×	72.1	40.6	47.4	15.2	30.6	22.7	38.1
Expert|DINOv3$_{\text{ViT-B}}$	-	×	75.2	45.3	44.8	22.4	36.2	27.5	41.9
Expert|DINOv3$_{\text{ViT-L}}$	-	×	80.0	50.4	67.9	26.3	46.1	33.8	50.7
Expert|DINOv3$_{\text{ViT-H+}}$	-	×	81.3	45.3	66.8	33.7	53.2	32.2	52.2

Table S9. Results of FLIR dataset (RGB → Thermal). Where "R→T" denotes the method specifically designed for RGB-to-Thermal adaptation. Note that we demonstrate the performance of four different expert model variants, but we only leverage DINOv3 ViT-H+ as the expert model's backbone in our ETS. Following [7], we report all results with two decimal places for a precise comparison.

Method	Venues	R→T	person	bicycle	car	mAP
DANN [8]	JMLR'16	×	32.02	30.52	48.88	37.14
SWDA [18]	CVPR'19	×	30.91	36.03	47.94	38.29
EPM [11]	ECCV'20	×	40.97	38.95	53.83	44.60
HT [6]	CVPR'23	×	70.87	48.11	78.45	65.81
D3T [7]	CVPR'24	✓	70.77	57.44	79.68	69.30
ETS	-	×	65.81	61.47	82.55	69.94
Expert|DINOv3$_{\text{ViT-S+}}$	-	×	40.35	48.95	64.15	51.15
Expert|DINOv3$_{\text{ViT-B}}$	-	×	66.71	55.98	78.61	67.10
Expert|DINOv3$_{\text{ViT-L}}$	-	×	68.18	60.12	83.12	70.47
Expert|DINOv3$_{\text{ViT-H+}}$	-	×	68.17	62.10	82.60	70.95

to-night scenarios. ETS consistently outperforms all competing methods across most object classes, with particularly notable gains in challenging classes such as "bus", "truck", and "traffic signs". This proves that our method effectively utilizes the generalization knowledge derived from VFMs. Notably, ETS achieves performance comparable to the expert model employing DINOv3 ViT-L backbone and approaches the capability of the largest expert model (*i.e.*, DINOv3 ViT-H+) which attains 52.2% mAP. The progressive performance improvement observed across different expert model variants further validates the critical role of foundation model capacity in addressing challenging domain adaptation scenarios. These results collectively demonstrate that ETS effectively bridges the domain gap between daytime and nighttime driving conditions, establishing new state-of-the-art performance on this benchmark.

I. RGB-to-Thermal DAOD Benchmark

Thermal imaging provides crucial capabilities in low-visibility conditions where RGB sensors fail, but transitioning from RGB to thermal domains presents significant challenges. The two modalities exhibit substantial domain gaps due to their different imaging principles—RGB captures reflected light while thermal detects heat signatures. This cross-modal shift, combined with limited annotated thermal data, makes RGB-to-thermal adaptation an ideal benchmark for evaluating domain adaptation methods.

Dataset

Following [7] We use the updated FLIR dataset [19] containing 5,142 aligned RGB-thermal pairs,as shown is Fig. S7. We select 2,064 RGB images as the source domain and 2,064 different thermal images as the target do-

Figure S7. Sample images from the FLIR dataset illustrating the domain adaptation from RGB to thermal imaging.

main, ensuring no aligned pairs are used during training. This prevents overfitting and provides a realistic evaluation of cross-modal adaptation. We evaluate on three consistently annotated categories: person, car, and bicycle, with 1,013 images reserved for testing.

Evaluation

In Tab. S9, our ETS achieves competitive performance on the challenging RGB-to-Thermal adaptation benchmark, attaining 69.94% mAP without specific design for this task. ETS demonstrates particular strength in detecting "bicycles" and "cars", highlighting its effectiveness in handling the substantial domain shift between visible and thermal spectra. Additionally, ETS achieves performance comparable to the largest expert model (DINOv3 ViT-H+), validating the guidance capabilities of VFMs in the thermal domain are limited.

Algorithm 1 Expert-Teacher-Student Collaborative Learning (ETS)

Input: Source data $\mathcal{D}_s = \{X_s, B_s, C_s\}$, Target data $\mathcal{D}_t = \{X_t\}$

Output: Trained student detector for target domain

Step (a): Offline Expert Model Training

Train expert model (frozen backbone) on source domain using $\mathcal{L}_{sup}(X_s, B_s, C_s)$

Step (b): Offline Detection&Prototype Generation

Generate expert pseudo labels $\tilde{Y}(\tilde{B}, \tilde{C})$ with confidence threshold δ and prototypes p^{exp} on target domain

Step (c): Online Expert-Teacher-Student Collaborative Learning

while $iter < T_{max\_iterations}$ **do**
  if $iter < T_{burn_in}$ **then**
    Burn-in Stage (Supervised Learning on Source)
    Train student model on source domain using $\mathcal{L}_{sup}(X_s, B_s, C_s)$
  else if $T_{burn\_in} < iter < T_{align}$ **then**
    Mutual-learning Stage (Supervised Learning on Source&Target)
    Keep training student model on source domain using $\mathcal{L}_{sup}(X_s, B_s, C_s)$
    Initialize teacher model from copy student weights
    Generate Teacher Pseudo Labels
    Generate teacher pseudo labels $\hat{Y}$ from teacher model on target images
    Label-level: Expert-Teacher Collaborative Teaching (ETCT)
    Match $\tilde{Y}$ and $\hat{Y}$ to generate collaborative pseudo labels Y_{col}
    Train student model with $\mathcal{L}_{colsup}(X_t, B_{col}, C_{col})$
  else
    Prototype-level: Expert-Teacher Joint Consolidating (ETJC)
    Generate prototypes $p^{exp}, p^{tea}, p^{stu}$ FROM each prototype network
    Compute prototype alignment losses: $\mathcal{L}_{Prot}^{E\text{-}S}$, $\mathcal{L}_{Prot}^{T\text{-}S}$, $\mathcal{L}_{Prot}^{C\text{-}D}$
  end if
  Optimization and EMA Update
  Optimize total loss:
  $\mathcal{L} = \lambda_1 \mathcal{L}_{colsup} + \lambda_2 \mathcal{L}_{colsup} + \lambda_3 \mathcal{L}_{Prot}^{E\text{-}S} + \lambda_4 \mathcal{L}_{Prot}^{T\text{-}S} + \lambda_5 \mathcal{L}_{Prot}^{C\text{-}D}$
  Update teacher model via EMA: $\theta_t \leftarrow \alpha\theta_t + (1-\alpha)\theta_s$
end while
return Final student model

Figure S8. Quantitative analysis of detection recall capacity on Cityscapes→BDD100k.

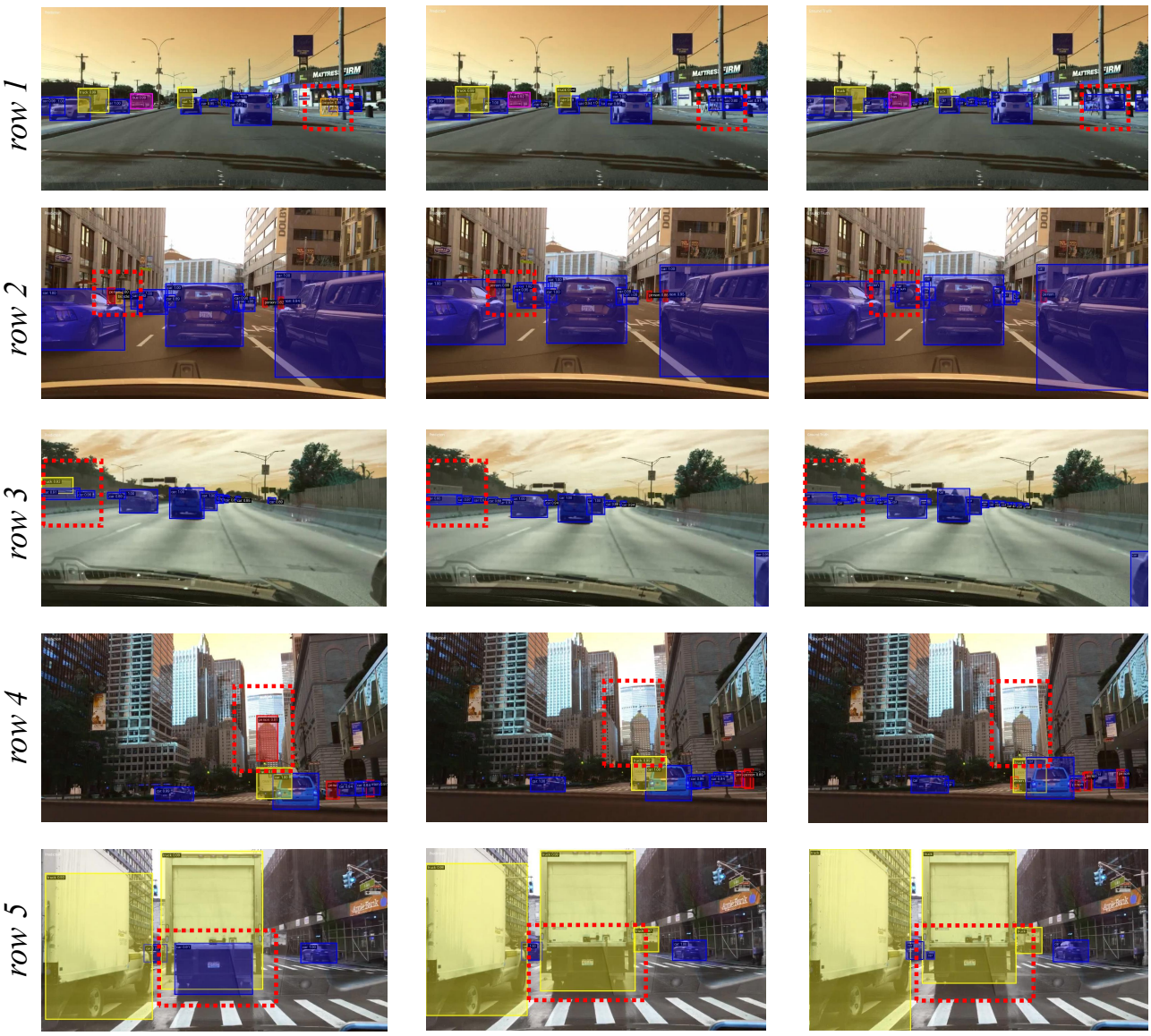

Figure S9. Quantitative analysis of localization&classification capacity on Cityscapes→BDD100k.

Figure S10. Quantitative analysis of Long-distance detection capacity on Cityscapes→BDD100k.

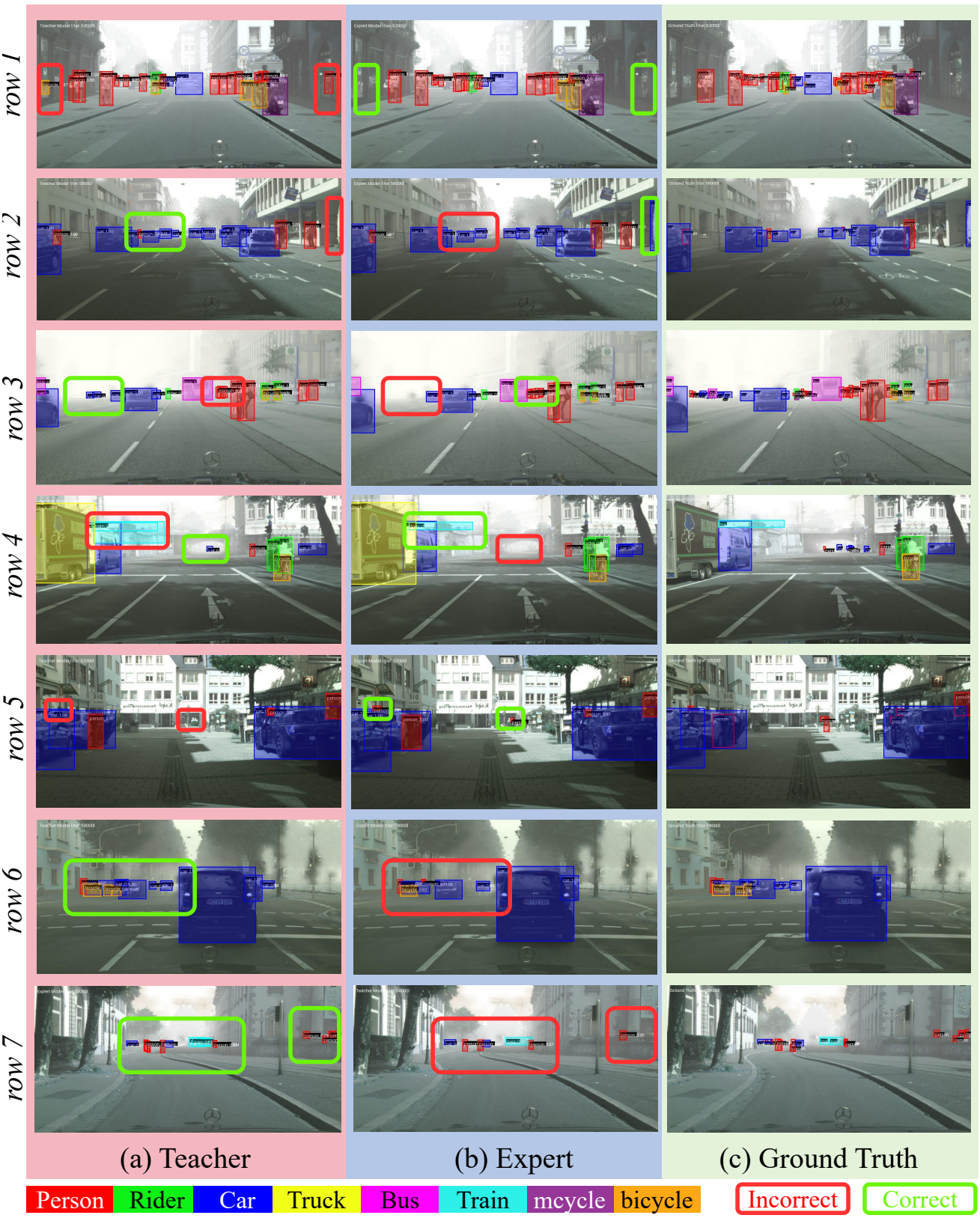

Figure S11. Quantitative analysis of complementary advantages of expert and teacher models on Cityscapes→Foggy Cityscapes.

References

[1] Yunfei Bai, Yiqiang Wu, Bin Zhu, and Xiaomao Li. Contrastive-domain mean teacher for domain adaptive object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025. 3

[2] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 1

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

[4] Yiming Cui, Liang Li, Jiehua Zhang, Chenggang Yan, Hongkui Wang, Shuai Wang, Heng Jin, and Li Wu. Stochastic context consistency reasoning for domain adaptive object detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1331–1340, 2024. 7

[5] Yiming Cui, Liang Li, Haibing Yin, Yuhan Gao, Yaoqi Sun, and Chenggang Yan. Debiased teacher for day-to-night domain adaptive object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2577–2587, 2025. 6, 7

[6] Jinhong Deng, Dongli Xu, Wen Li, and Lixin Duan. Harmonious teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23829–23838, 2023. 7

[7] Dinh Phat Do, Taehoon Kim, Jaemin Na, Jiwon Kim, Keonho Lee, Kyunghwan Cho, and Wonjun Hwang. D3t: Distinctive dual-domain teacher zigzagging across rgb-thermal gap for domain-adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23313–23322, 2024. 7

[8] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016. 7

[9] Liqiang He, Wei Wang, Albert Chen, Min Sun, Cheng-Hao Kuo, and Sinisa Todorovic. Bidirectional alignment for domain adaptive detection with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18775–18785, 2023. 3

[10] Lukas Hoyer, Dengxin Dai, Haoran Wang, and Luc Van Gool. Mic: Masked image consistency for context-enhanced domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11721–11732, 2023. 7

[11] Cheng-Chun Hsu, Yi-Hsuan Tsai, Yen-Yu Lin, and Ming-Hsuan Yang. Every pixel matters: Center-aware feature alignment for domain adaptive object detector. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 733–748. Springer, 2020. 7

[12] Yuan Jicheng, Le-Tuan Anh, Hauswirth Manfred, and Le-Phuoc Danh. Cooperative students: Navigating unsupervised domain adaptation in nighttime object detection. *2024 IEEE International Conference on Multimedia and Expo (ICME)*, 2024. 7

[13] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11484–11493, 2023. 7

[14] Marc-Antoine Lavoie, Anas Mahmoud, and Steven L Waslander. Large self-supervised models bridge the gap in domain adaptive object detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4692–4702, 2025. 3, 4

[15] Yu-Jhe Li, Xiaoliang Dai, Chih-Yao Ma, Yen-Cheng Liu, Kan Chen, Bichen Wu, Zijian He, Kris Kitani, and Peter Vajda. Cross-domain adaptive teacher for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7581–7590, 2022. 3

[16] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024. 1

[17] Chang Nie, Muhammad Ali Qadar, Shaodong Zhou, Hui Zhang, Yang Shi, Jinwu Gao, and Zhifeng Sun. Transnational image object detection datasets from nighttime driving. *Signal, Image and Video Processing*, 17(4):1123–1131, 2023. 6

[18] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6956–6965, 2019. 7

[19] Heng Zhang, Elisa Fromont, Sébastien Lefevre, and Bruno Avignon. Multispectral fusion for object detection with cyclic fuse-and-refine blocks. In *2020 IEEE International conference on image processing (ICIP)*, pages 276–280. IEEE, 2020. 7

[20] Yin Zhang, Yongqiang Zhang, Zian Zhang, Man Zhang, Rui Tian, and Mingli Ding. Isp-teacher: Image signal process with disentanglement regularization for unsupervised domain adaptive dark object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7387–7395, 2024. 7

[21] Xiangyu Zhao, Yicheng Chen, Shilin Xu, Xiangtai Li, Xinjiang Wang, Yining Li, and Haian Huang. An open and comprehensive pipeline for unified object grounding and detection. *arXiv preprint arXiv:2401.02361*, 2024. 1