Sony AI

SONY

ICCV OCT 19-23, 2025 HONOLULU HAWAII

DuET

Dual Incremental Object Detection via Exemplar-Free Task Arithmetic

Munish Monga[1,2] Vishal Chudasama[1] Pankaj Wasnik[1] Biplab Banerjee[2]

[1]Sony Research India [2]Indian Institute of Technology, Bombay, India

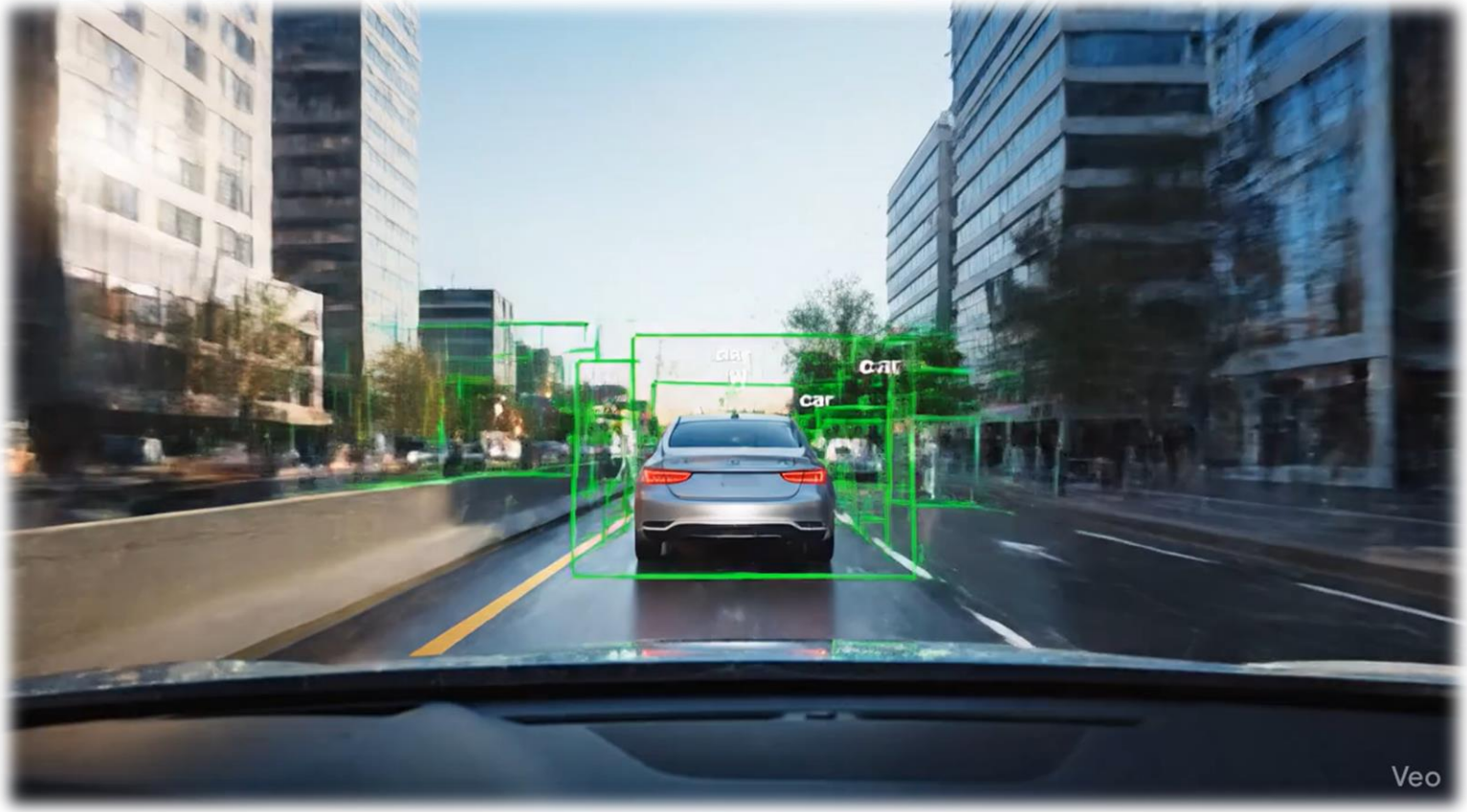
Problem Motivation

Dynamic Environments Demand Incremental Learning:

- ❑ Autonomous vehicles must recognize **new road signs, vehicles, and objects** under varying weather, lighting, and city conditions.
- ❑ Similarly, surveillance systems must detect **new objects, threats, or anomalies** in changing environments.

Challenges in Traditional Object Detection:

- ❑ Trained once on a fixed dataset → **Fails in new environments *(domain shift)***
- ❑ New objects emerge over time → **Cannot recognize unseen classes *(class shift)***
- ❑ Retraining from scratch is **computationally expensive** and **impractical**.



Source: AI generated video (using Veo)

Limitations of existing approaches

Class Incremental Object Detection (CIOD)

✅ Learns new object classes ($C_1 \rightarrow C_2$)

❌ Fails in unseen domains ($D_1 \neq D_2$)

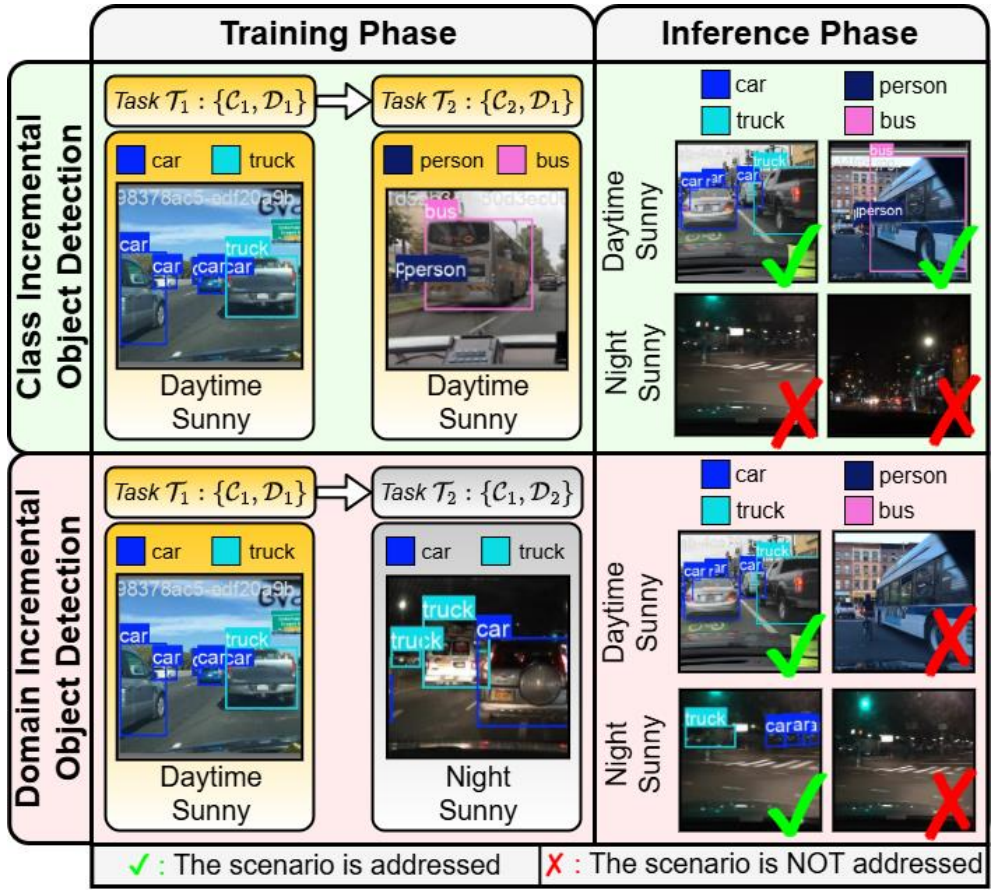
❌ Suffers from **catastrophic forgetting** in new environments

Domain Incremental Object Detection (DIOD)

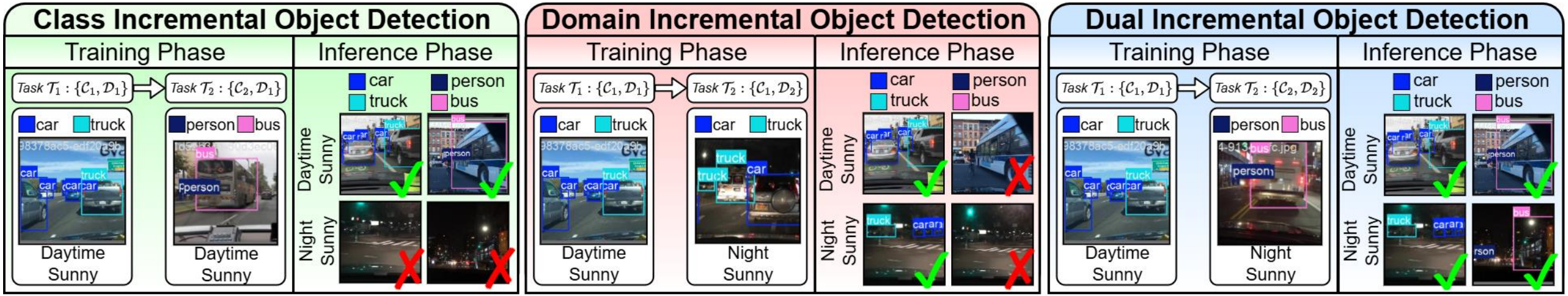
✅ Adapts to **new domains** ($D_1 \rightarrow D_2$)

❌ Cannot detect **new object classes** ($C_1 \neq C_2$)

❌ Suffers from **catastrophic forgetting** when adapting to new object categories



Dual Incremental Object Detection (DuIOD)



We present a new paradigm in object detection literature that requires object detectors to:

- ✅ Incrementally learn new object categories: $(C_1 \rightarrow C_2 \rightarrow C_3 \rightarrow \cdots C_T)$
- ✅ Simultaneously adapt to unseen domains: $(D_1 \rightarrow D_2 \rightarrow D_3 \rightarrow \cdots D_T)$
- ✅ Learn in an **Exemplar-Free manner**, without access to previously seen training data.

Problem Formulation: Defining DuIOD

Sony AI

SONY

Key Requirements to setup DuIOD:

✅ The model must detect objects from all learned classes across all encountered domains.

✅ It must prevent **catastrophic forgetting** while generalizing to new environmental conditions.

✅ **Exemplar-Free Constraint**: the model should not store previous task data for retraining.

Formal Problem Definition:

Let $T = \{T_1, T_2, \cdots T_T\}$ represent a sequence of tasks where:

❑ Each task T_t introduces a new set of object classes C_t from a novel domain D_t.

❑ The cumulative set of all classes and domains encountered up to task T_t is:

$$C_{1:t} = \bigcup_{j=1}^{t} C_j\,, \qquad D_{1:t} = \bigcup_{j=1}^{t} D_j$$

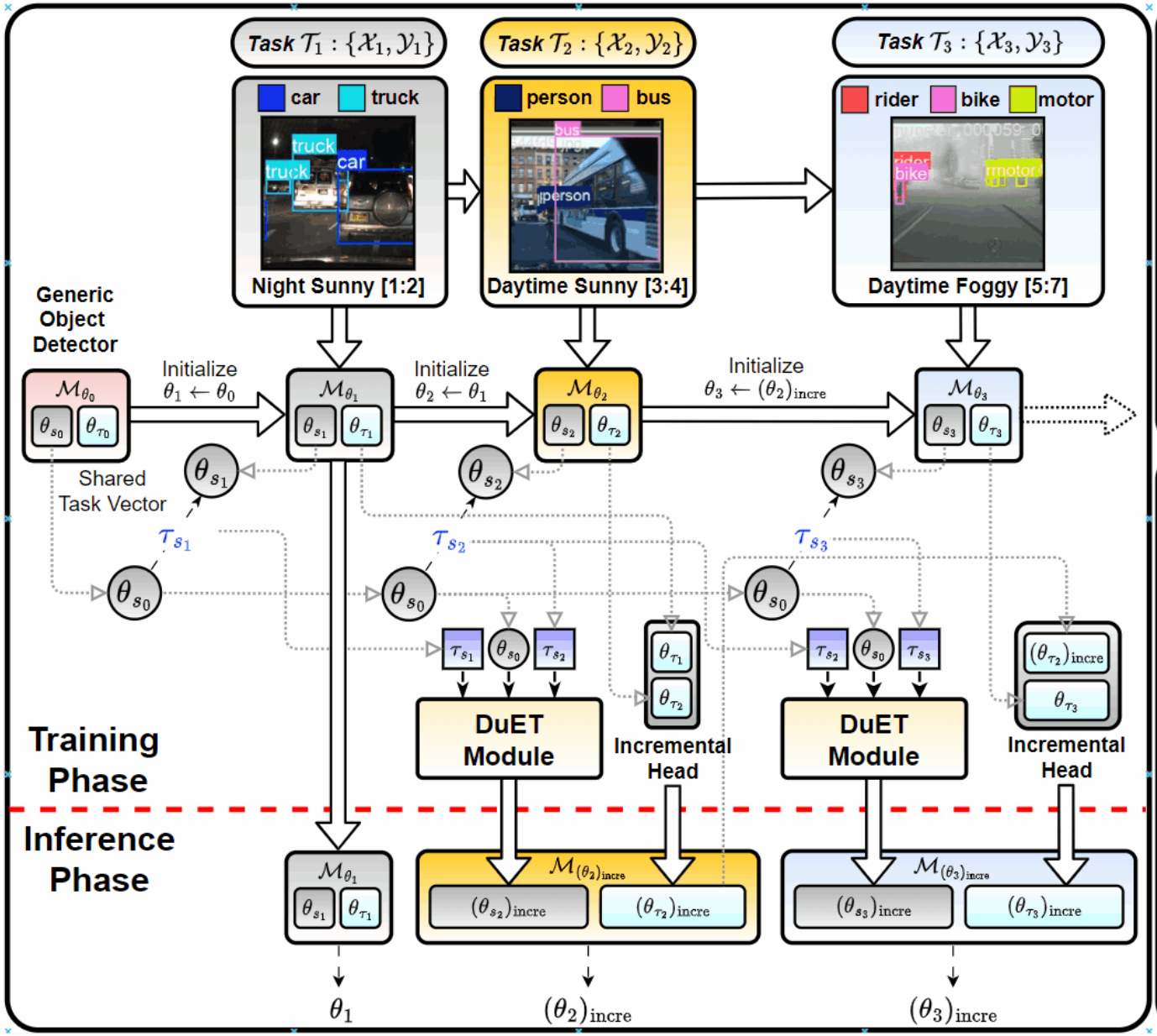
❑ **Our objective** is to incrementally update the object detection model M_{θ_t}, parameterized by θ_t, using only data from D_t, ensuring that M_{θ_t} can accurately detect objects from all learned classes $C_{1:t}$ across all encountered domains $D_{1:t}$.

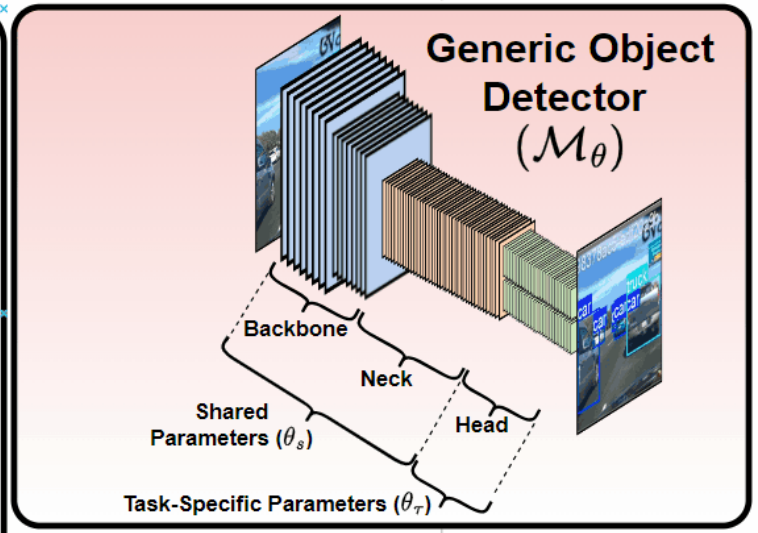
ICCV OCT 19-23, 2025 HONOLULU HAWAII

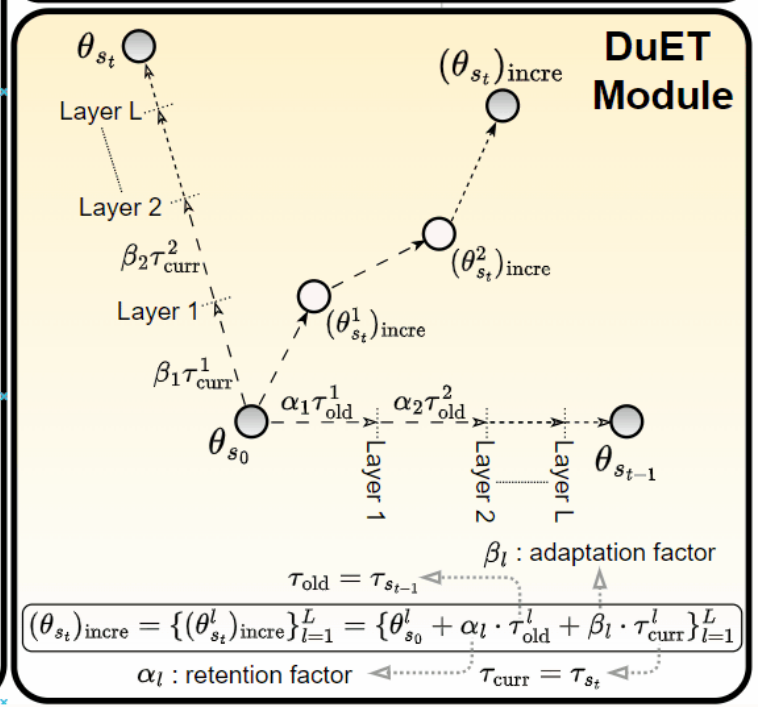
A quick look at the DuET framework!

Sony AI

SONY







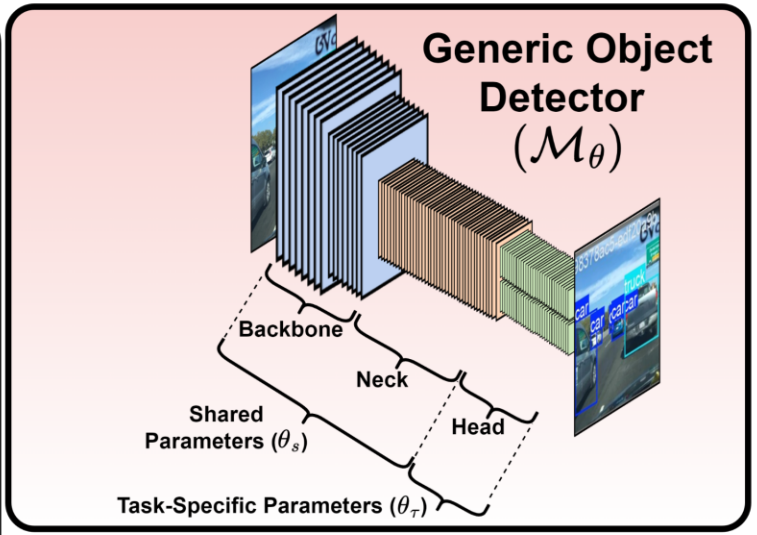
ICCV OCT 19-23, 2025 HONOLULU HAWAII

Building up the DuET Framework

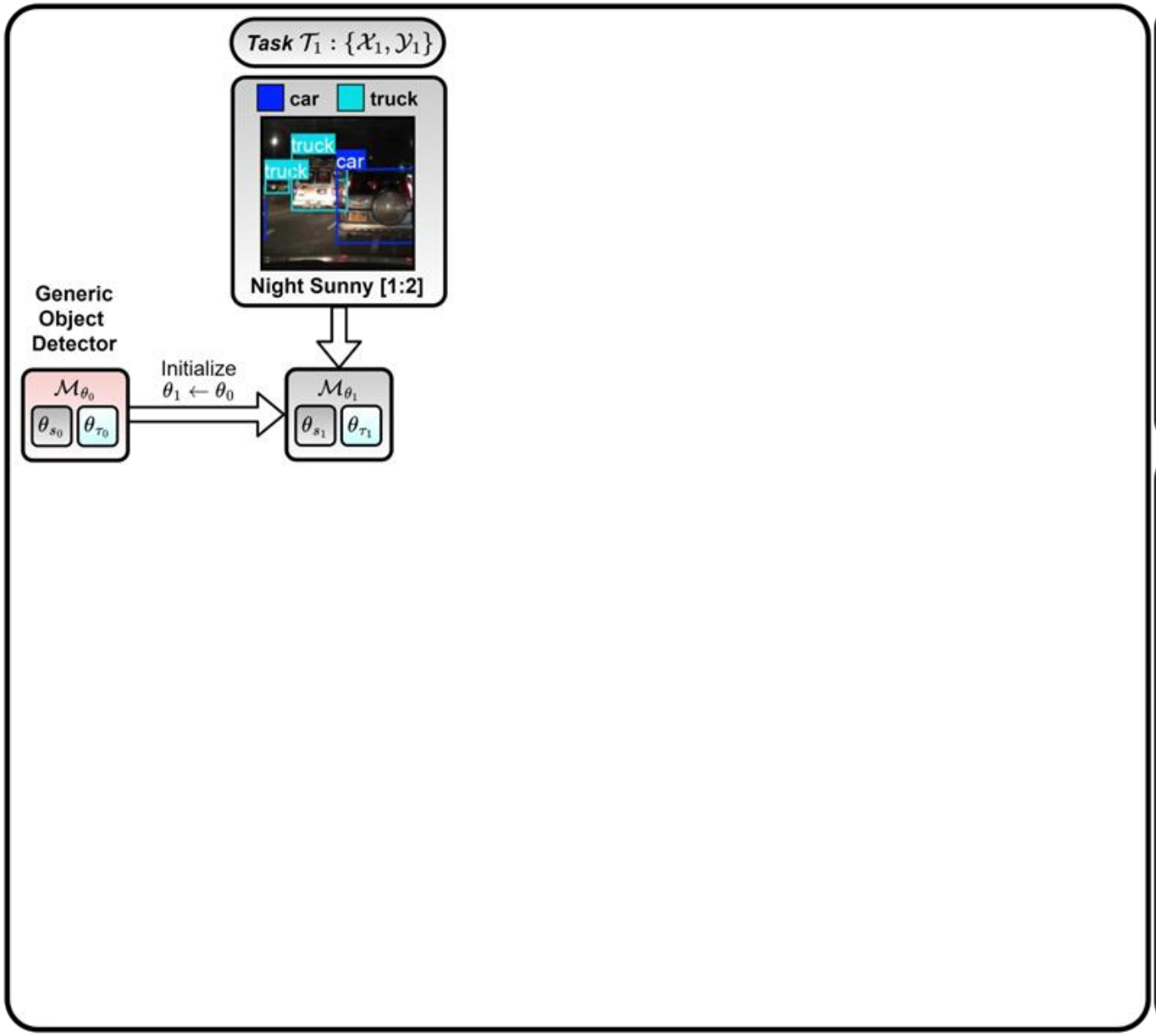
Generic Object Detector

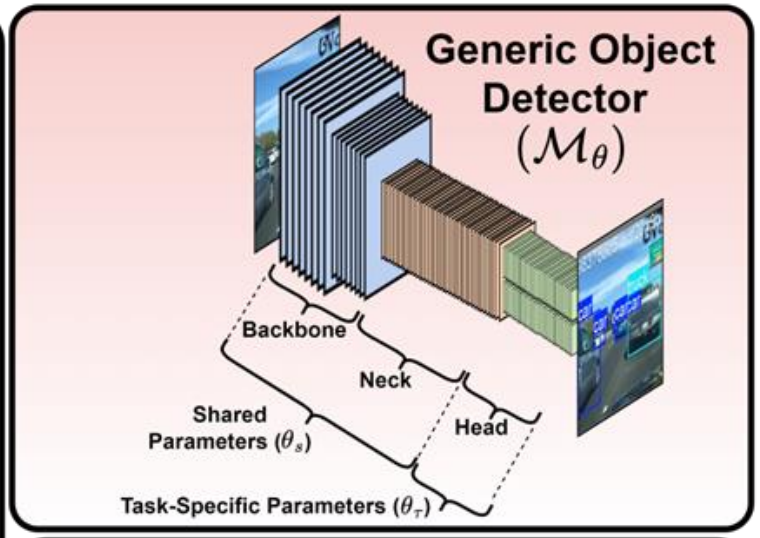
$\mathcal{M}_{\theta_0}$

θ_{s_0} θ_{τ_0}

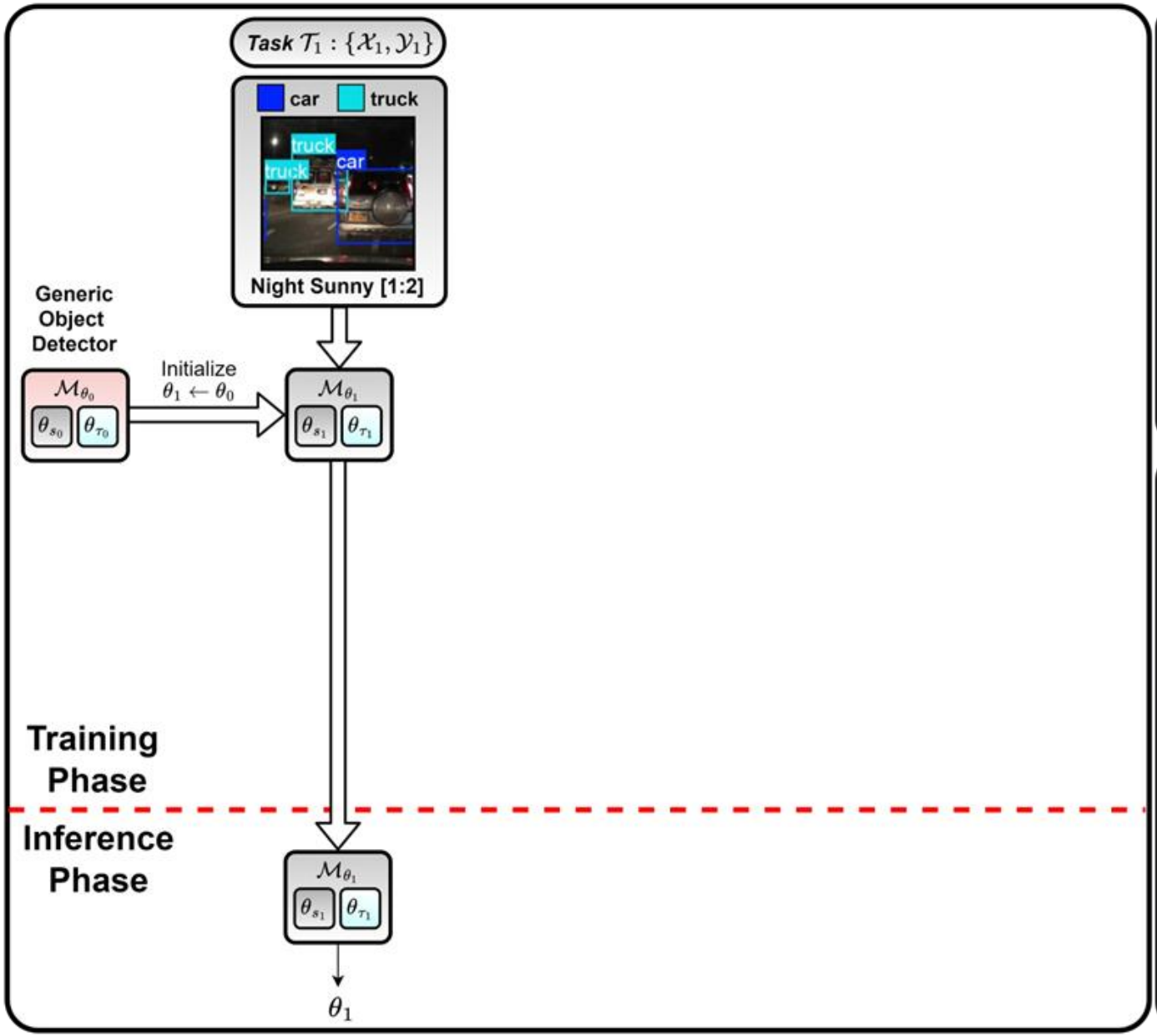


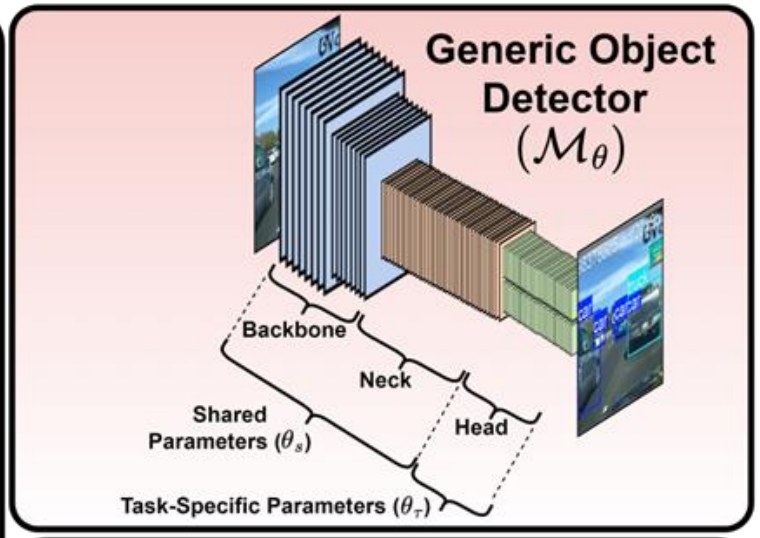
Building up the DuET Framework



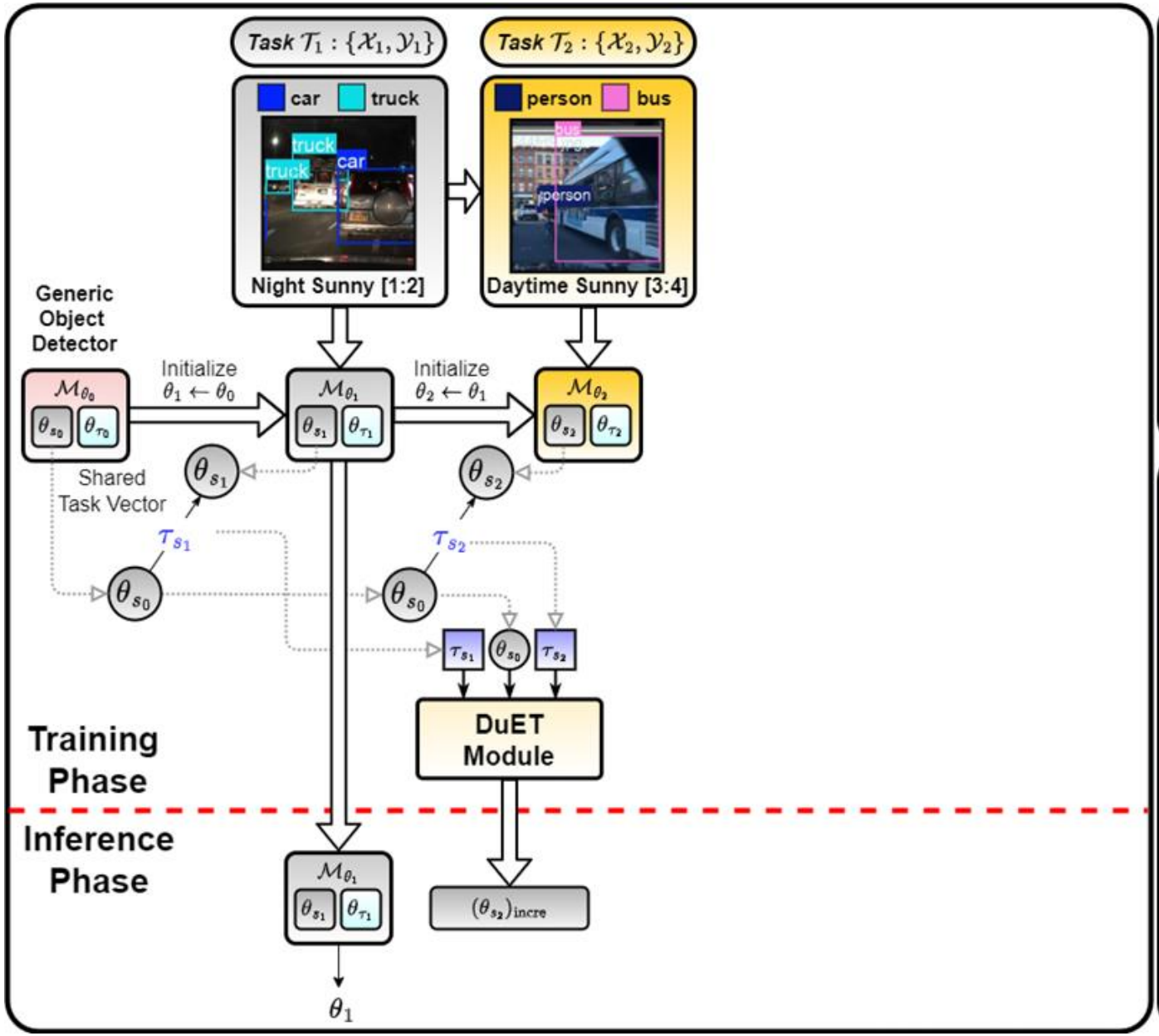


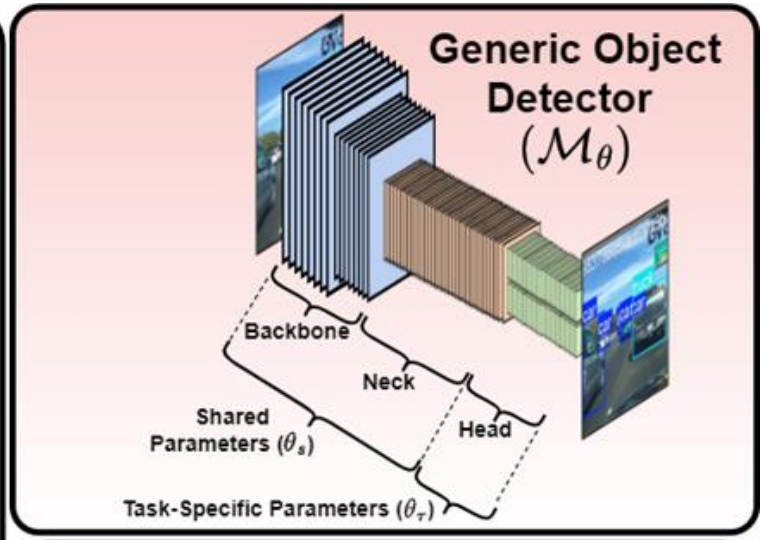
Building up the DuET Framework

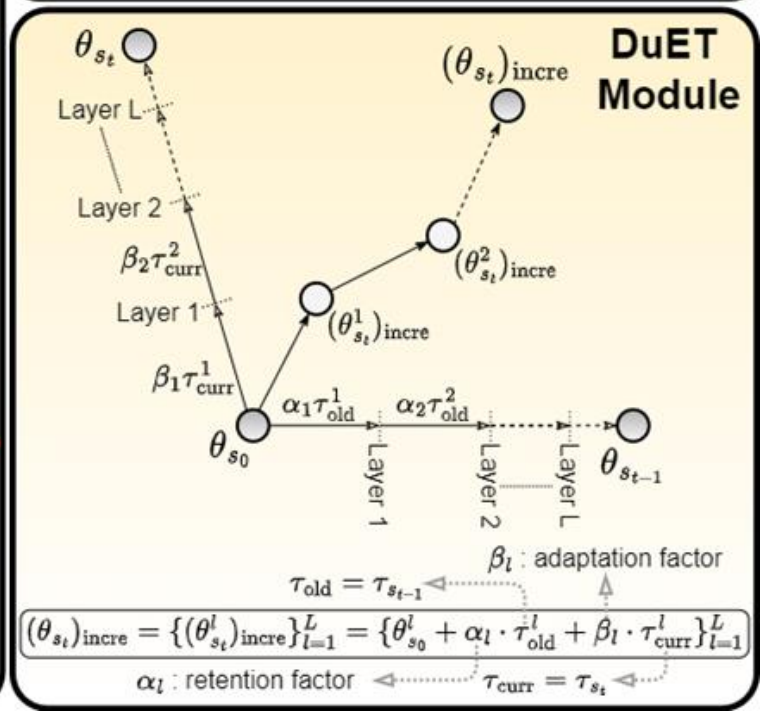




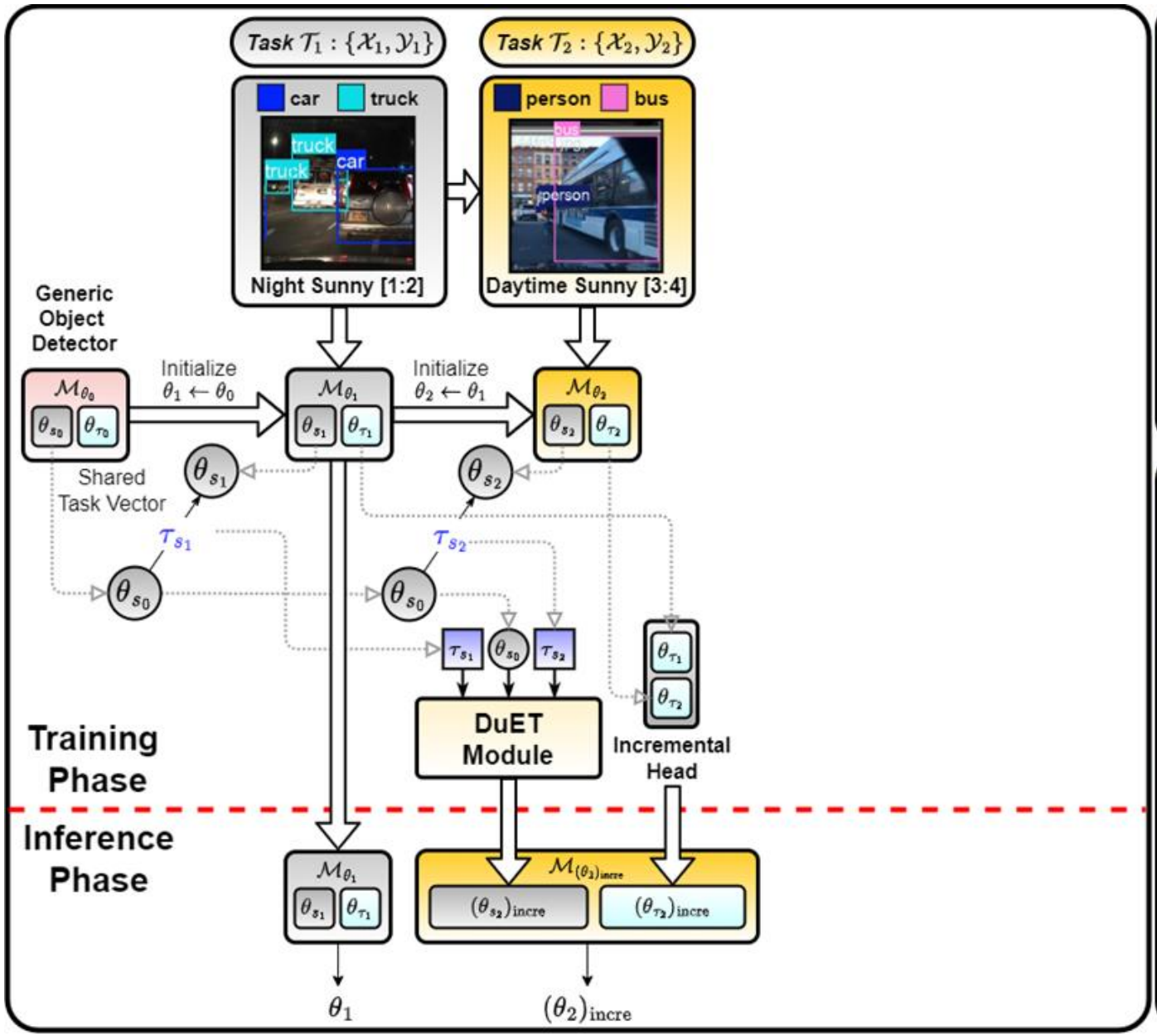
Building up the DuET Framework

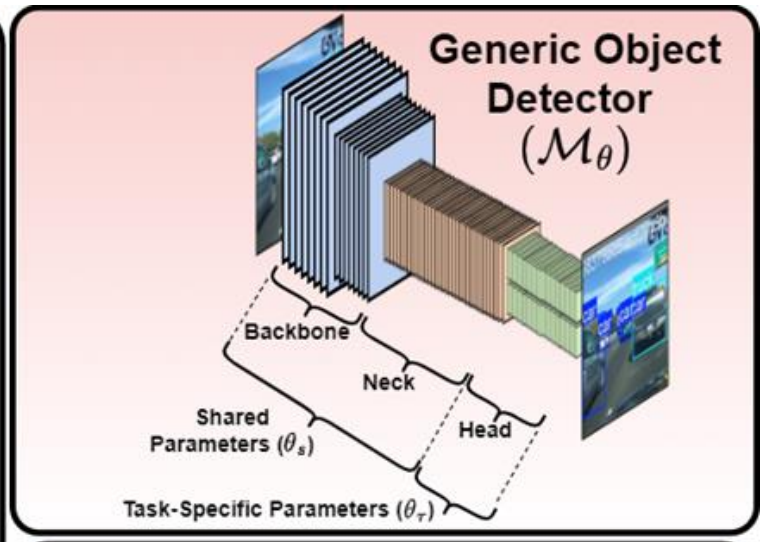


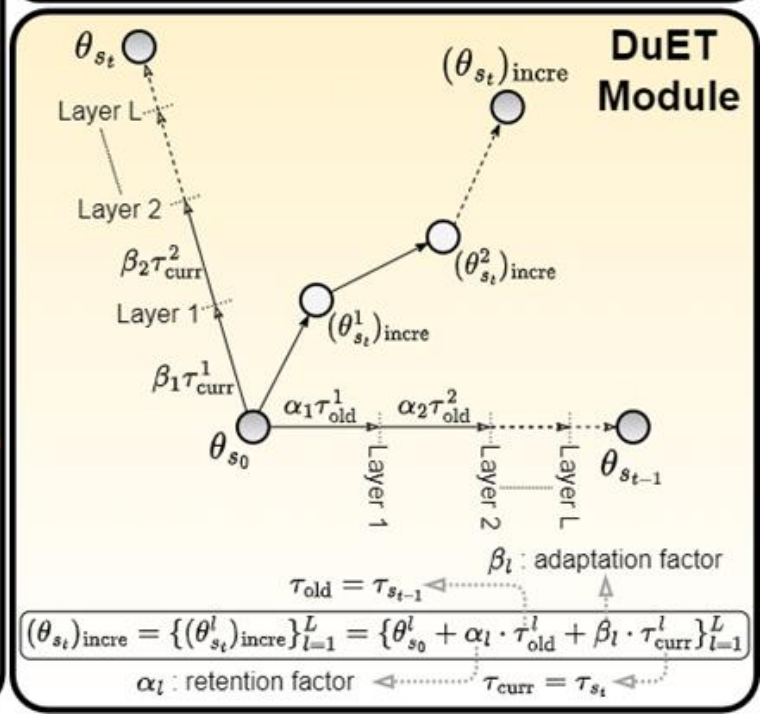




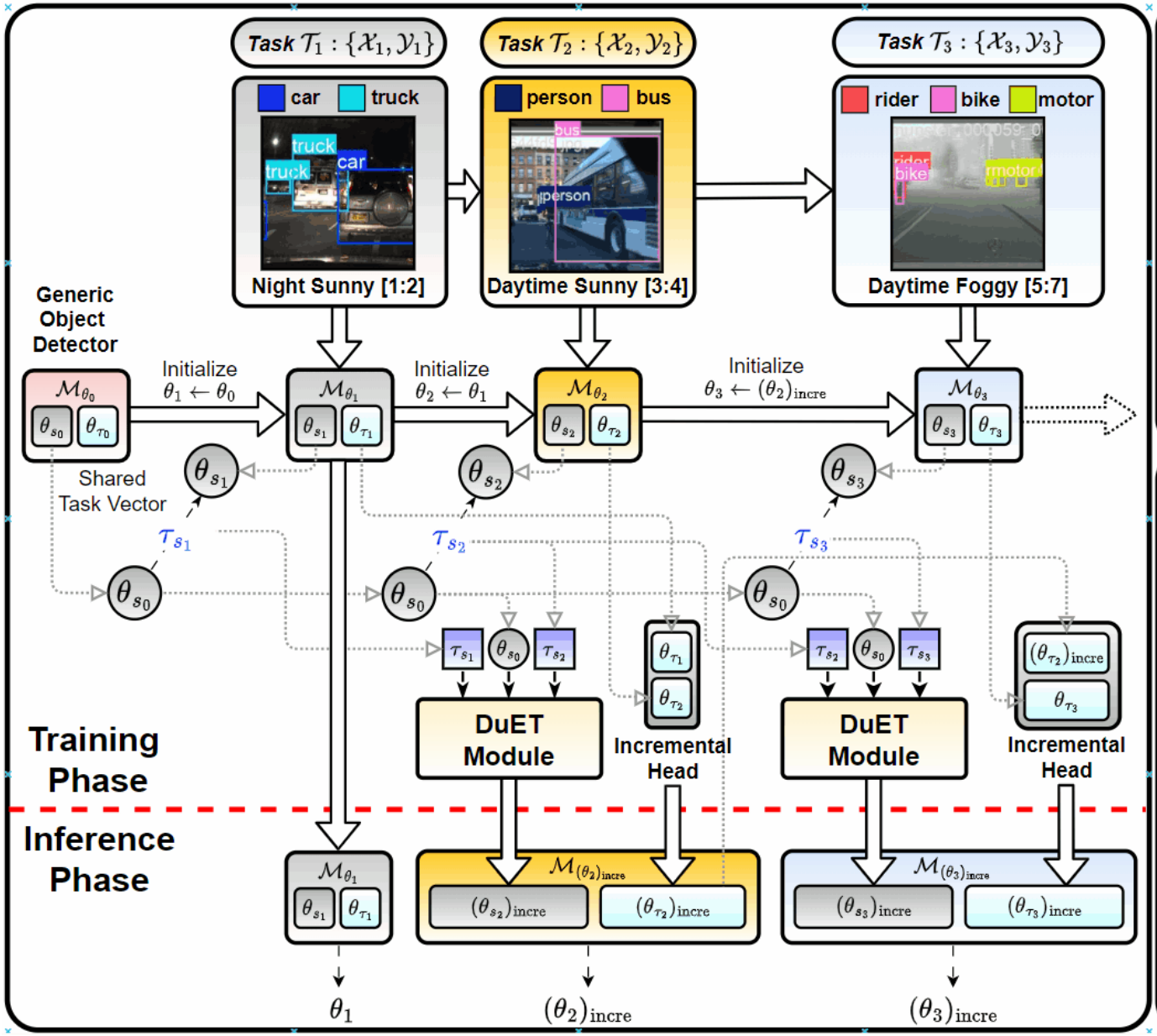
Building up the DuET Framework

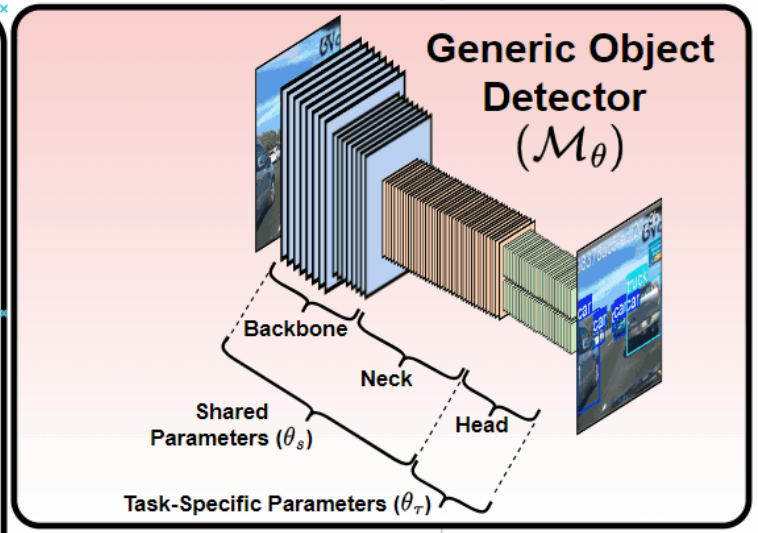


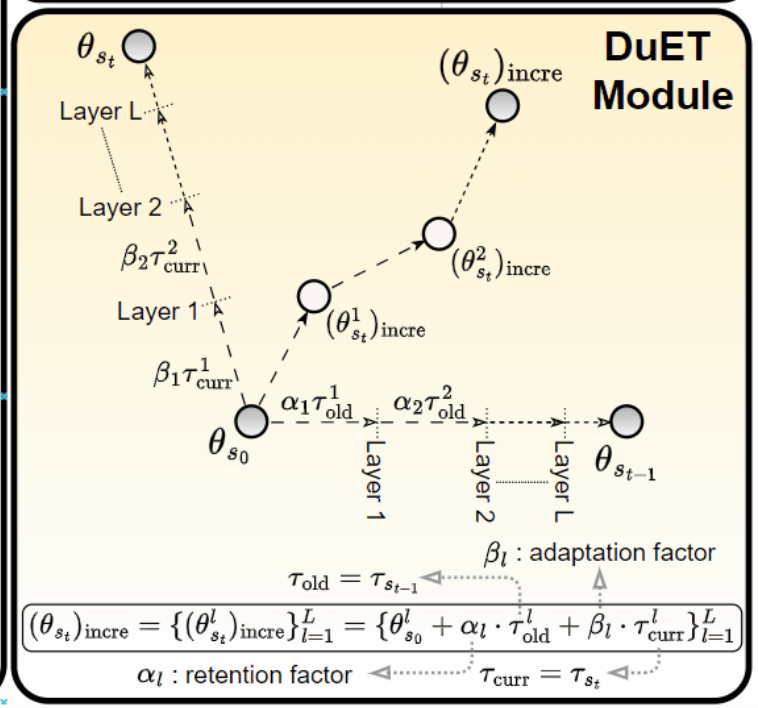




DuET Framework







DuET Module & Incremental Head

- **DuET Module** fuses task vectors on shared weights. It merges past and current task vectors dynamically using layer-wise retention (α_l) and adaptation (β_l) factors, **effectively mitigating catastrophic forgetting.**

$$\tau_{old} = \theta_{s_{t-1}} - \theta_{s_0}, \qquad \tau_{curr} = \theta_{s_t} - \theta_{s_0}$$

- For each layer, $l \in \{1, 2, \cdots L\}$, we define a **p-factor:**

$$p_l = \frac{\|\tau_{old}^l\| - \|\tau_{curr}^l\|}{\|\tau_{old}^l + \tau_{curr}^l\| + \varepsilon}$$

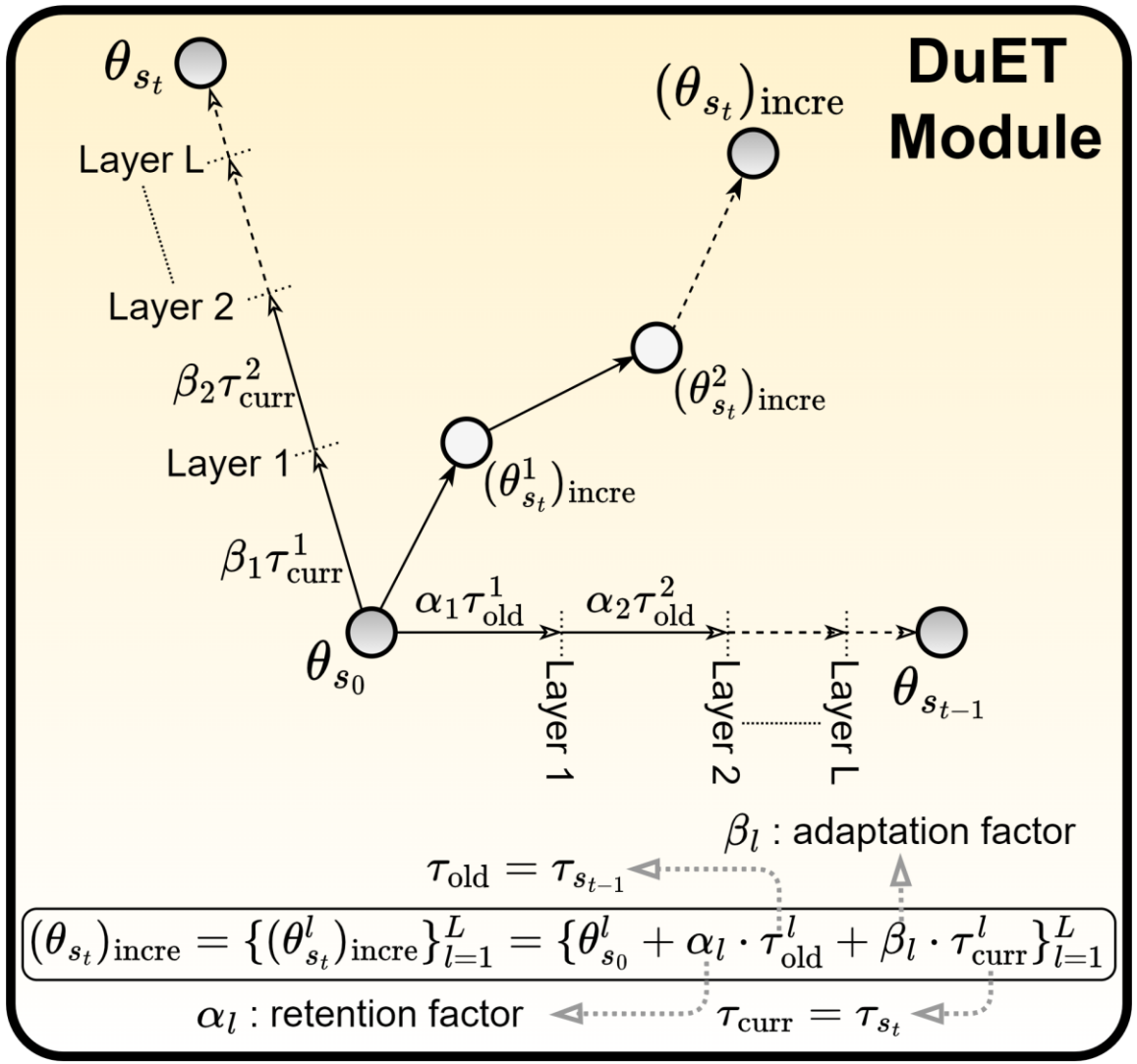
$$\delta_l = \gamma \tanh(p_l)$$

$$\alpha_l = \alpha_{base} + clamp(\delta_l, -\gamma, \gamma), \ \beta_l = 1 - \alpha_l$$

$$(\theta_{s_t}^l)_{incre} = \theta_{s_0}^l + \alpha_l \tau_{old}^l + \beta_l \tau_{curr}^l$$

- Applying this across all layers yields $(\theta_{s_t})_{incre}$ as shown.
- **Incremental Head** concatenates task-specific parameters from past and current tasks, thereby enhancing model's generalization across diverse detection domains.

$$(\theta_{\tau_t})_{incre} \leftarrow [\theta_{\tau_t}; (\theta_{\tau_{t-1}})_{incre}]$$



Loss Objectives

- For the **base task** $(t = 1)$, we optimize the object detector using the standard detection loss, $\mathcal{L}_{Detector}$,which is specific to the base detector and is responsible for object localization and classification.

$$\theta_1 \leftarrow \theta_0 - \eta \cdot \nabla_\theta \mathcal{L}_{Detector}$$

- For **incremental tasks** $(t \geq 2)$, we augment $\mathcal{L}_{Detector}$ with a **modified Distillation Loss** $(\mathcal{L}^*_{Distill})$ that also filters low confidence classification predictions and high variance bounding box predictions from the old model, along with our **Directional Consistency Loss** $(\mathcal{L}_{DC})$ scaled by scaling coefficients $\lambda_{Distill}$ & λ_{DC} respectively.

$$\mathcal{L}_{Total} = \begin{cases} \mathcal{L}_{Detector}, & t = 1 \\ \mathcal{L}_{Detector} + \lambda_{Distill}\mathcal{L}^*_{Distill} + \lambda_{DC}\mathcal{L}_{DC} & t \geq 2 \end{cases}$$

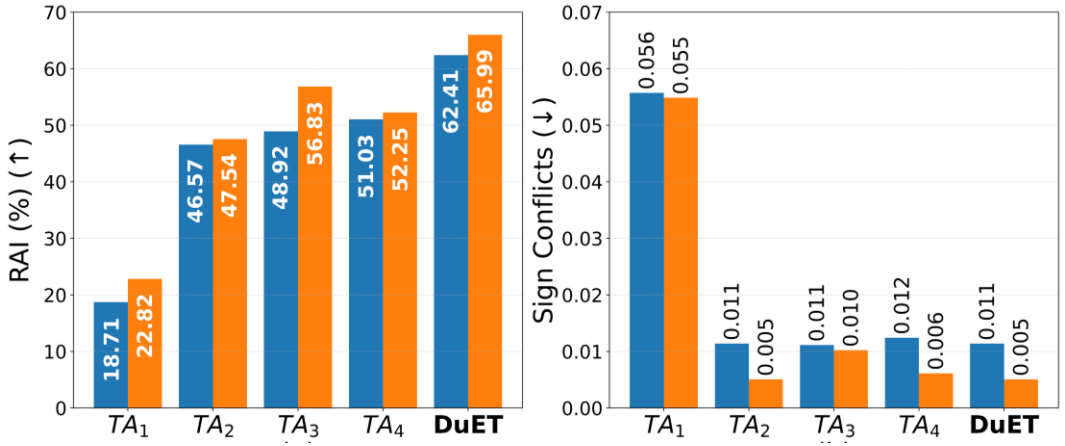
$$\theta_t \leftarrow \theta_{t-1} - \eta \cdot \nabla_\theta \mathcal{L}_{Total}$$

Directional Consistency Loss

- Incremental learning suffers from conflicting weight updates during model-merging [1], which can destabilize the merging of shared parameters.
- We introduce the **Directional Consistency Loss** ($\mathcal{L}_{DC}$) to address this. This loss penalizes updates in the shared parameter space that diverge in direction relative to previous incremental changes, thus promoting a consistent evolution of the shared task vectors.
- Specifically, for ($t \geq 2$) we define the DC Loss between consecutive tasks t & $t-1$ as:

$$\mathcal{L}_{DC} = \sum_{i \,\epsilon\, \theta_s} ReLU\left[-((\tau_{s_t}^{(i)} - \tau_{s_{t-1}}^{(i)}) \cdot \left(\tau_{s_{t-1}}^{(i)} - \tau_{s_{t-2}}^{(i)}\right)\right]$$

- $\mathcal{L}_{DC}$ effectively **reduces catastrophic forgetting** by stabilizing weight updates.
- $\mathcal{L}_{DC}$ remains **object detector-agnostic** – can be applied to different detection backbones.



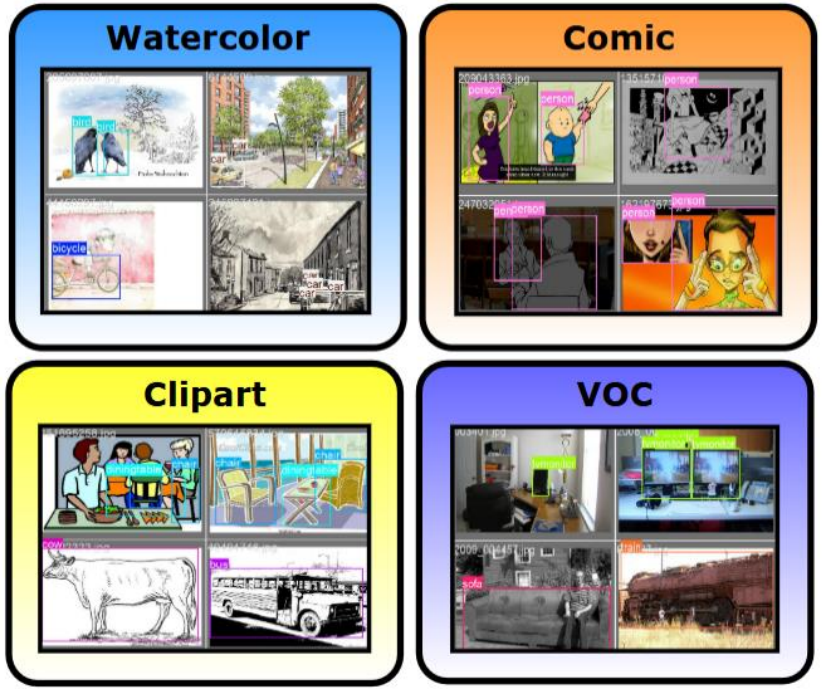
[1] Yadav, P., Tam, D., Choshen, L., Raffel, C. A., & Bansal, M. (2023). Ties-merging: Resolving interference when merging models. *Advances in Neural Information Processing Systems*, *36*, 7093-7115..

Datasets: Pascal Series

Sony AI

SONY

Pascal Series Datasets Statistics			
Dataset	Total Classes	Train Images	Val Images
Pascal VOC [1]	20	16551	4952
Clipart [2]	20	500	500
Watercolor [2]	6	1000	1000
Comic [2]	6	1000	1000



	Class ID	Class Name	Domains			
			Watercolor	Comic	Clipart	VOC
3	1	bicycle	✓	✓	✓	✓
	2	bird	✓	✓	✓	✓
	3	car	✓	✓	✓	✓
3	4	cat	✓	✓	✓	✓
	5	dog	✓	✓	✓	✓
	6	person	✓	✓	✓	✓
7	7	aeroplane			✓	✓
	8	boat			✓	✓
	9	bottle			✓	✓
	10	bus			✓	✓
	11	chair			✓	✓
	12	cow			✓	✓
	13	diningtable			✓	✓
7	14	horse			✓	✓
	15	motorbike			✓	✓
	16	pottedplant			✓	✓
	17	sheep			✓	✓
	18	sofa			✓	✓
	19	train			✓	✓
	20	tvmonitor			✓	✓

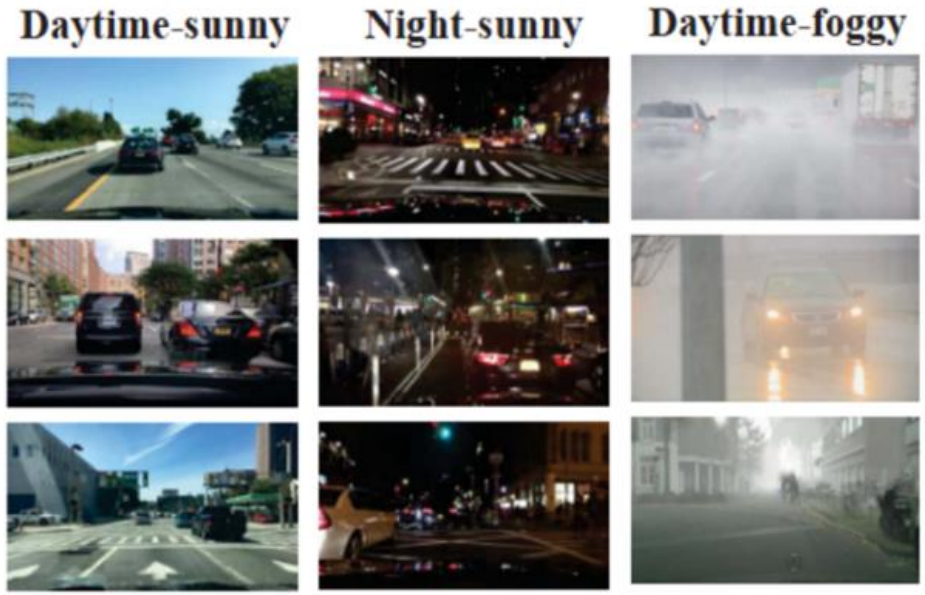
[1] Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International journal of computer vision*, *88*, 303-338.

[2] Inoue, N., Furuta, R., Yamasaki, T., & Aizawa, K. (2018). Cross-domain weakly-supervised object detection through progressive domain adaptation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5001-5009).

ICCV OCT 19-23, 2025 HONOLULU HAWAII

Datasets: Diverse Weather Series

Diverse Weather Series Datasets Statistics			
Dataset	**Total Classes**	**Train Images**	**Val Images**
Daytime-Sunny [1]	7	19317	8289
Night-Sunny [1]	7	25868	7756
Daytime-Foggy [2]	7	1829	688



	Class ID	Class Name	Domains				
			Daytime Sunny	Night Sunny	Daytime Foggy	Night Rainy	Dusk Rainy
2	1	bus	✓	✓	✓	✓	✓
	2	bike	✓	✓	✓	✓	✓
2	3	car	✓	✓	✓	✓	✓
	4	motor	✓	✓	✓	✓	✓
3	5	person	✓	✓	✓	✓	✓
	6	rider	✓	✓	✓	✓	✓
	7	truck	✓	✓	✓	✓	✓

[1] Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., ... & Darrell, T. (2020). Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 2636-2645).

[2] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., ... & Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3213-3223).

Proposed Evaluation Metrics

- Since DuIOD goes beyond preserving old knowledge (***retention***), it also demands evaluating how well the model can adapt to unseen categories that emerge within known domains (***adaptability***).
- To address this requirement, we introduce the **Retention-Adaptability Index (RAI)**, which is defined as the mean of the **Average Retention Index (Avg RI)** and the **Average Generalization Index (Avg GI).**

$$RAI = \frac{Avg\ RI + Avg\ GI}{2}$$

- For each domain D_i corresponding to task T_i where $i \in \{1, 2, \cdots T-1\}$, we define **Retention Index** as RI_{D_i} and calculate Avg RI as:

$$RI_{D_i} = \frac{mAP_{old}^{T_T}(D_i[C_i])}{mAP_{new}^{T_i}(D_i[C_i])}, \qquad Avg\ RI = \frac{1}{T-1}\sum_{i=1}^{T-1} RI_{D_i}$$

- Similarly, for a given domain D_i at task T_j, we define **Generalization Index** as GI_{D_i,T_j} and calculate Avg GI as:

$$GI_{D_i,T_j} = \frac{mAP_{unseen}^{T_j}(D_i[C_{unseen}])}{mAP_{ref}(D_i[C_{unseen}])}, \qquad Avg\ GI = \frac{1}{N}\sum_{(D_i,T_j)} GI_{D_i,T_j}$$

- **Significance:** A higher Avg RI implies better retention of past knowledge while lower values indicate **catastrophic forgetting**. Similarly, Avg GI quantifies the model's ability to generalize to unseen classes across all encountered domains so far and a higher value of Avg GI indicates **better generalization**.

Quantitative Results

- On an average, DuET achieves an RAI improvement of **+9.82%** and **+13.12%** while preserving **87.44%** and **89.30% Avg RI** on the Pascal Series datasets, two-phase and multi-phase experiments respectively.
- Similarly, on the Diverse Weather Series datasets, it achieves an RAI improvement of **+12.59%** and **+11.39%** while preserving **88.06%** and **88.57% Avg RI** on the two-phase and multi-phase experiments, respectively.

Detailed quantitative results on Daytime Sunny [1:4] → Night Sunny [5:7].

Method	Base Detector	Trainable Params (M)	T1 Daytime Sunny [1:4]	T2: Night Sunny [5:7]				Avg RI (%)	Avg GI (%)	RAI (%)
				Old	New	Unseen				
				Daytime Sunny [1:4]	Night Sunny [5:7]	Night Sunny [1:4]	Daytime Sunny [5:7]			
Sequential FT	YOLO11n	2.58	$49.40_{\pm 0.3}$	$0.00_{\pm 0.0}$	$62.20_{\pm 0.5}$	$12.60_{\pm 0.4}$	$35.90_{\pm 0.3}$	$0.00_{\pm 0.0}$	$45.88_{\pm 0.6}$	$22.94_{\pm 0.3}$
$\text{LwF}_{\text{ECCV}'16}$ [1]	YOLO11n	2.58	$49.40_{\pm 0.2}$	$27.60_{\pm 0.4}$	$0.34_{\pm 0.6}$	$21.30_{\pm 0.3}$	$0.67_{\pm 0.5}$	$55.87_{\pm 0.3}$	$21.88_{\pm 0.7}$	$38.88_{\pm 0.6}$
$\text{ERD}_{\text{CVPR}'22}$ [2]	YOLO11n	2.58	$49.40_{\pm 0.5}$	$33.00_{\pm 0.4}$	$34.00_{\pm 0.3}$	$26.10_{\pm 0.6}$	$29.10_{\pm 0.7}$	$66.80_{\pm 0.5}$	$53.04_{\pm 0.3}$	$59.92_{\pm 0.4}$
$\text{LDB}_{\text{AAI}'24}$ [3]	VitDet	110.52	$45.30_{\pm 0.6}$	$0.50_{\pm 0.3}$	$15.10_{\pm 0.4}$	$0.30_{\pm 0.5}$	$16.90_{\pm 0.7}$	$1.10_{\pm 0.2}$	$22.41_{\pm 0.3}$	$11.76_{\pm 0.6}$
$\text{CL} - \text{DETR}_{\text{CVPR}'23}$ [4]	Deformable DETR	39.85	$46.29_{\pm 0.4}$	$27.41_{\pm 0.5}$	$31.94_{\pm 0.6}$	$19.85_{\pm 0.3}$	$32.55_{\pm 0.4}$	$59.21_{\pm 0.2}$	$54.96_{\pm 0.5}$	$57.09_{\pm 0.4}$
DuET (Ours)	YOLO11n	2.58	$49.40_{\pm 0.2}$	$43.50_{\pm 0.1}$	$22.20_{\pm 0.3}$	$31.60_{\pm 0.2}$	$27.40_{\pm 0.1}$	$\mathbf{88.06}_{\pm 0.2}$	$\mathbf{56.95}_{\pm 0.1}$	$\mathbf{72.51}_{\pm 0.2}$

[1] Li, Z., & Hoiem, D. (2017). Learning without forgetting. IEEE transactions on pattern analysis and machine intelligence, 40(12), 2935-2947.

[2] Feng, T., Wang, M., & Yuan, H. (2022). Overcoming catastrophic forgetting in incremental object detection via elastic response distillation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 9427-9436).

[3] Song, X., He, Y., Dong, S., & Gong, Y. (2024, March). Non-exemplar domain incremental object detection via learning domain bias. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, No. 13, pp. 15056-15065).

[4] Liu, Y., Schiele, B., Vedaldi, A., & Rupprecht, C. (2023). Continual detection transformer for incremental object detection.

Evaluating DuET across different Base Detectors

Results of proposed DuET framework on Daytime Sunny [1:4] → Night Sunny [5:7] with different base detectors.

Method	Base Detector	Trainable Params (M)	GFLOPs	Avg RI (%)	Avg GI (%)	RAI (%)
DuET (Ours)	VitDet [1]	110.52	1829.61	$27.55_{\pm 0.3}$	$28.22_{\pm 0.8}$	$27.89_{\pm 0.5}$
	Deformable DETR [2]	39.85	11.77	$84.45_{\pm 0.2}$	$33.45_{\pm 0.1}$	$58.95_{\pm 0.2}$
	RT-DETR-l [3]	32	103.4	$47.73_{\pm 0.2}$	$21.00_{\pm 0.1}$	$34.37_{\pm 0.2}$
	RT-DETR-x [3]	65.49	222.5	$56.39_{\pm 0.2}$	$24.15_{\pm 0.1}$	$40.27_{\pm 0.2}$
	YOLO11n [4]	**2.58**	**6.3**	$88.06_{\pm 0.2}$	$\mathbf{56.95_{\pm 0.1}}$	$\mathbf{72.51_{\pm 0.3}}$
	YOLO11x [4]	56.84	194.4	$\mathbf{96.88_{\pm 0.2}}$	$42.41_{\pm 0.1}$	$69.18_{\pm 0.2}$

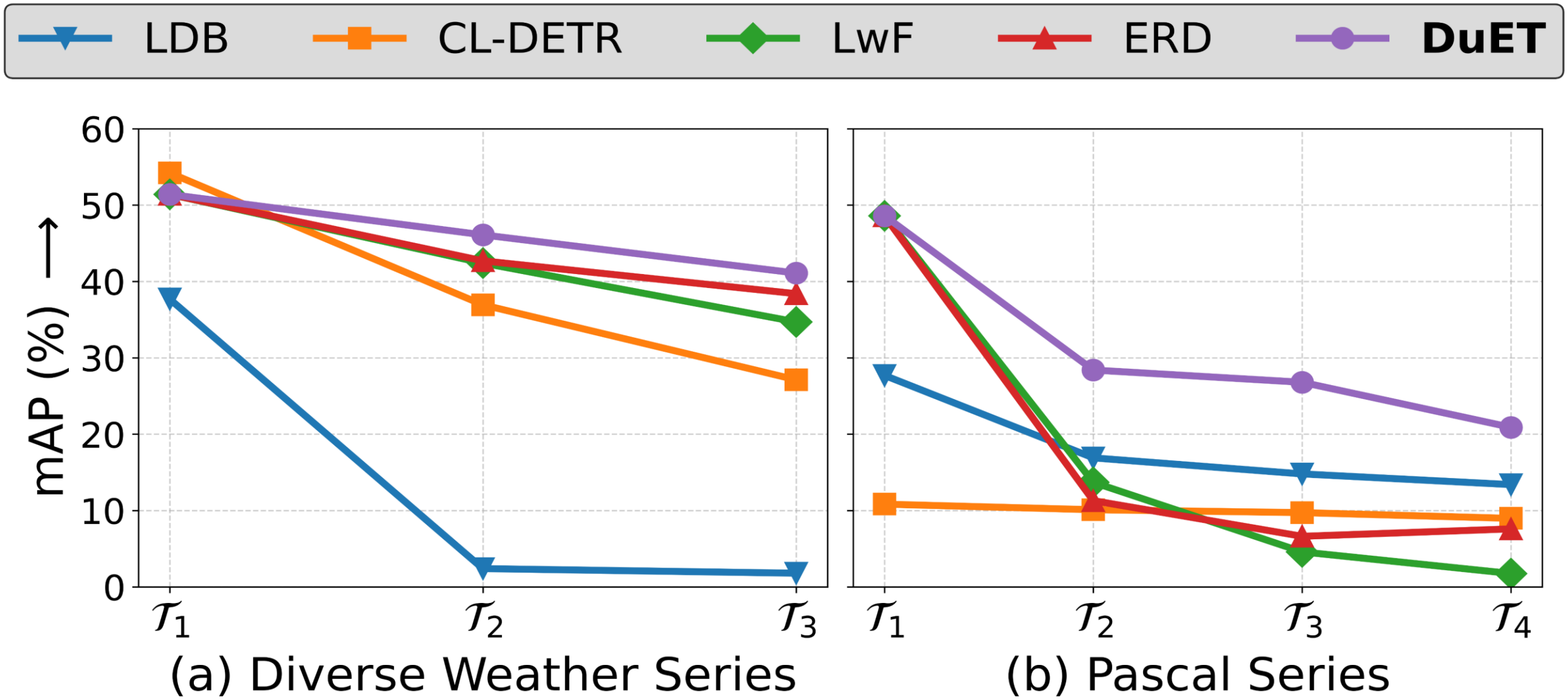
[1] Li, Y., Mao, H., Girshick, R., & He, K. (2022, October). Exploring plain vision transformer backbones for object detection. In European conference on computer vision (pp. 280-296). Cham: Springer Nature Switzerland.

[2] Zhu, X., Su, W., Lu, L., Li, B., Wang, X., & Dai, J. (2020). Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159.

[3] Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., ... & Chen, J. (2024). Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16965-16974).

[4] Jocher, G., Qiu, J., & Chaurasia, A. (2024). Ultralytics YOLO11, Version 11.0. 0

Catastrophic Forgetting of old classes



Qualitative Analysis: Two Phase Experiments

Sony AI

SONY

ICCV OCT 19-23, 2025 HONOLULU HAWAII

Qualitative Analysis: Multi Phase Experiments

Sony AI

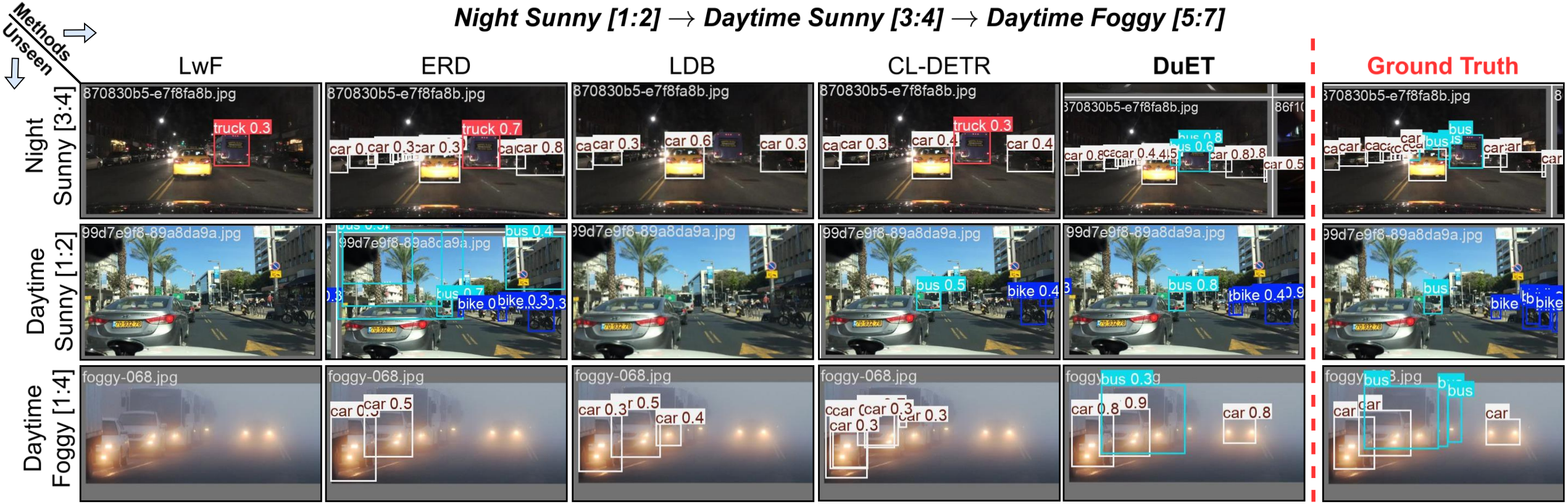
SONY

ICCV OCT 19-23, 2025 HONOLULU HAWAII

Qualitative Analysis: Multi Phase Experiments

Sony AI

SONY



Ablations: Role of each component

Ablations of DuET framework with different components and losses

Seq FT	Incremental Head	DuET Module	$\mathcal{L}^*_{Distill}$	$\mathcal{L}_{DC}$	Avg RI (%)	Avg GI (%)	RAI (%)
✗	✗	✗	✗	✗	0.5	9.13	4.82
✓	✗	✗	✗	✗	0.75	12.86	6.81
✓	✓	✗	✗	✗	24.75	33.36	29.06
✓	✓	✓	✗	✗	75.00	37.26	56.13
✓	✓	✓	✓	✗	87.06	37.75	62.41
✓	✓	✓	✓	✓	**87.44**	**44.54**	**65.99**

Ablations: Influence of random class-domain order

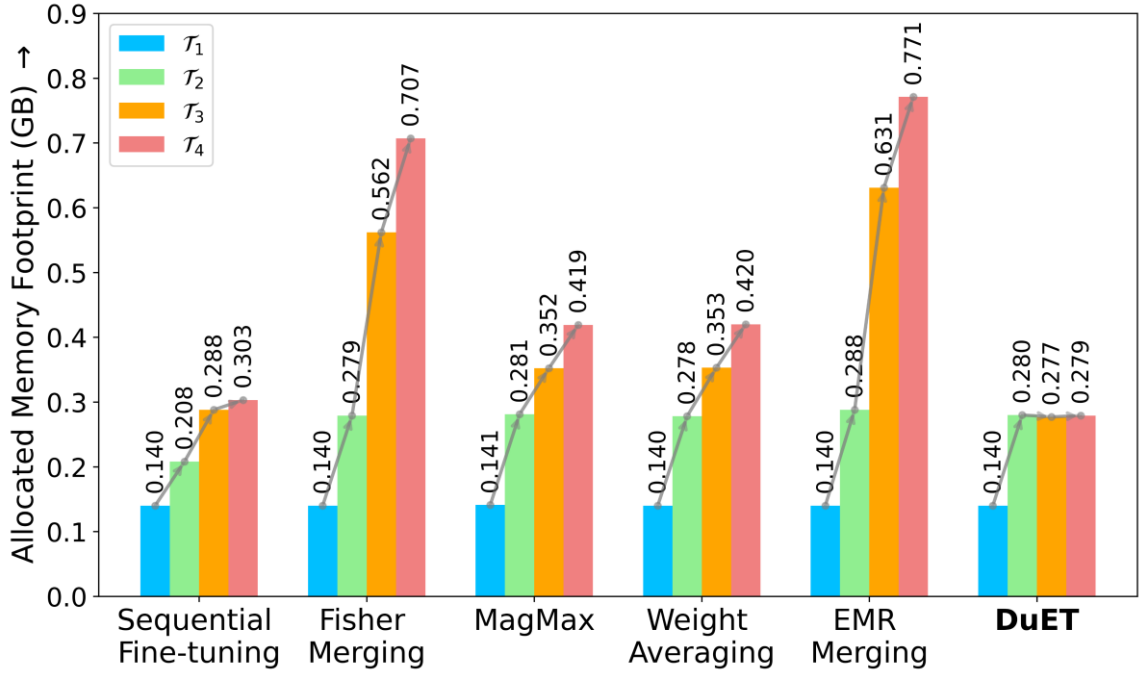
Influence of random permutations across three incremental tasks on Diverse Weather Series datasets.

T_1	T_2	T_3	Avg RI (%)	Avg GI (%)	RAI (%)
Night Sunny [5:7]	Daytime Sunny [1:2]	Daytime Foggy [3:4]	83.49	51.01	67.25
Night Sunny [3:4]	Daytime Sunny [5:7]	Daytime Foggy [1:2]	80.39	51.76	66.08
Night Sunny [1:2]	Daytime Sunny [3:4]	Daytime Foggy [5:7]	88.57	41.92	65.25
Daytime Foggy [1:2]	Night Sunny [3:4]	Daytime Sunny [5:7]	78.34	50.54	64.44
Daytime Sunny [1:2]	Daytime Foggy [3:4]	Night Sunny [5:7]	88.33	35.97	62.15
Standard Deviation			4.12	6.26	1.72

Ablations: Complexity Analysis

Computational complexity analysis of various methods

Method	Base Detector	GFLOPs	Trainable Params (M)	Avg. Inference Speed (ms)	Avg. Memory Footprint (GB)
Sequential FT	YOLO11n	6.3	2.58	9.15	0.235
LwF	YOLO11n	6.3	2.58	9.125	0.261
ERD	YOLO11n	6.3	2.58	9.075	0.257
LDB	ViTDet	1829.61	110.52	137.92	1.818
CL-DETR	Deformable DETR	11.77	39.85	39.075	0.789
DuET (Ours)	YOLO11n	6.3	2.58	4.4	0.244



Key Contributions

- To the best of our knowledge, **DuET** is the first method to tackle the proposed task of Dual Incremental Object Detection using the concept of Task Merging.
- **DuET** framework offers a simple yet effective **Task-Arithmetic based solution**, which is **object-detector agnostic**, validated on YOLO11 and RT-DETR, and enables real-time incremental object detection.
- **Directional Consistency Loss** helps to **mitigate sign conflicts** during model merging, and ensures stable incremental learning.
- **Retention-Adaptability Index** quantifies both **catastrophic forgetting** and **domain generalization** performance effectively.
- **DuET** opens a new research direction in IOD, providing a promising framework for continual learning and domain adaptation.

Thank You

Scan to view paper!