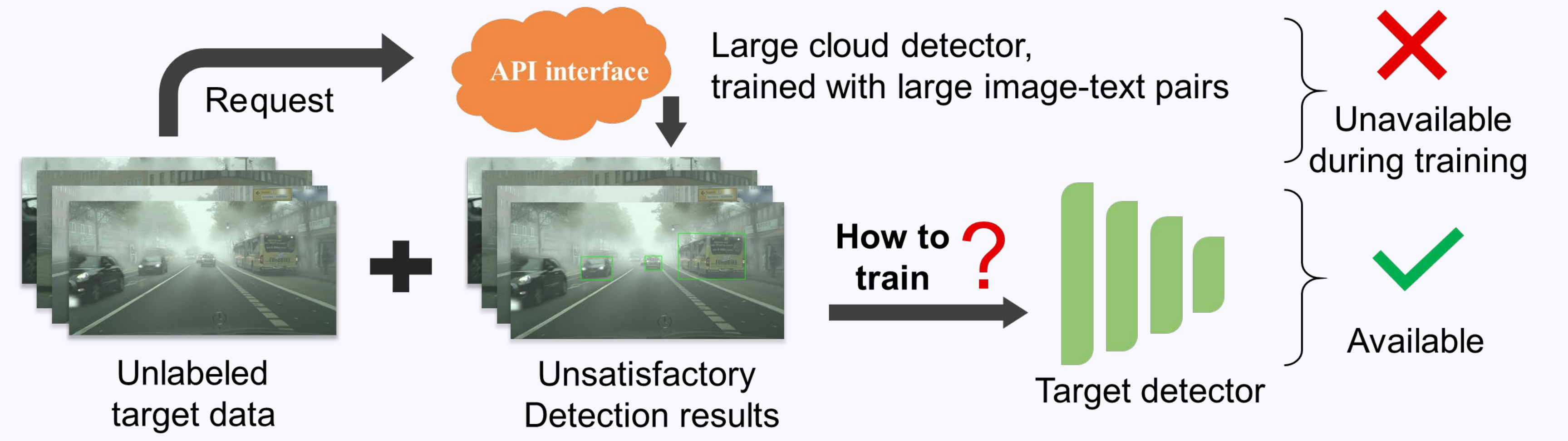
Cloud Object Detector Adaptation by Integrating Different Source Knowledge

Shuaifeng Li[1] Mao Ye[1*] Lihua Zhou[1] Nianxin Li[1] Siying Xiao[1] Song Tang[2] Xiatian Zhu[3]

[1] University of Electronic Science and Technology of China [2] University of Shanghai for Science and Technology [3] University of Surrey

NEURAL INFORMATION PROCESSING SYSTEMS

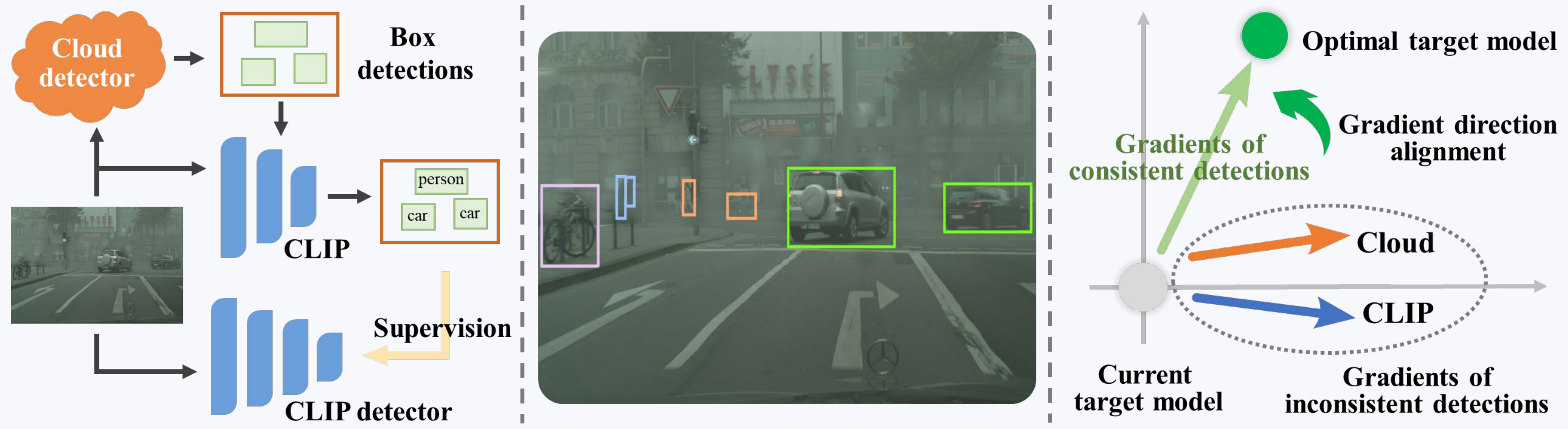
CODA: Cloud Object Detector Adaptation



CODA enables **open target scenarios** and **open object categories** adaptation due to large grounded pre-training of cloud detector.

Conditions	UDAOD	SFOD	Black-box DAOD	**CODA**
Source data access	✓	✗	✗	✗
Source model access	✓	✓	✗	✗
Cloud API access	✗	✗	✓	✓
High domain similarity	✓	✓	✓	✗
Ability				
Flexible architecture	✗	✗	✓	✓
Open categories	✗	✗	✗	✓
Open scenarios	✗	✗	✗	✓

Idea and Contributions



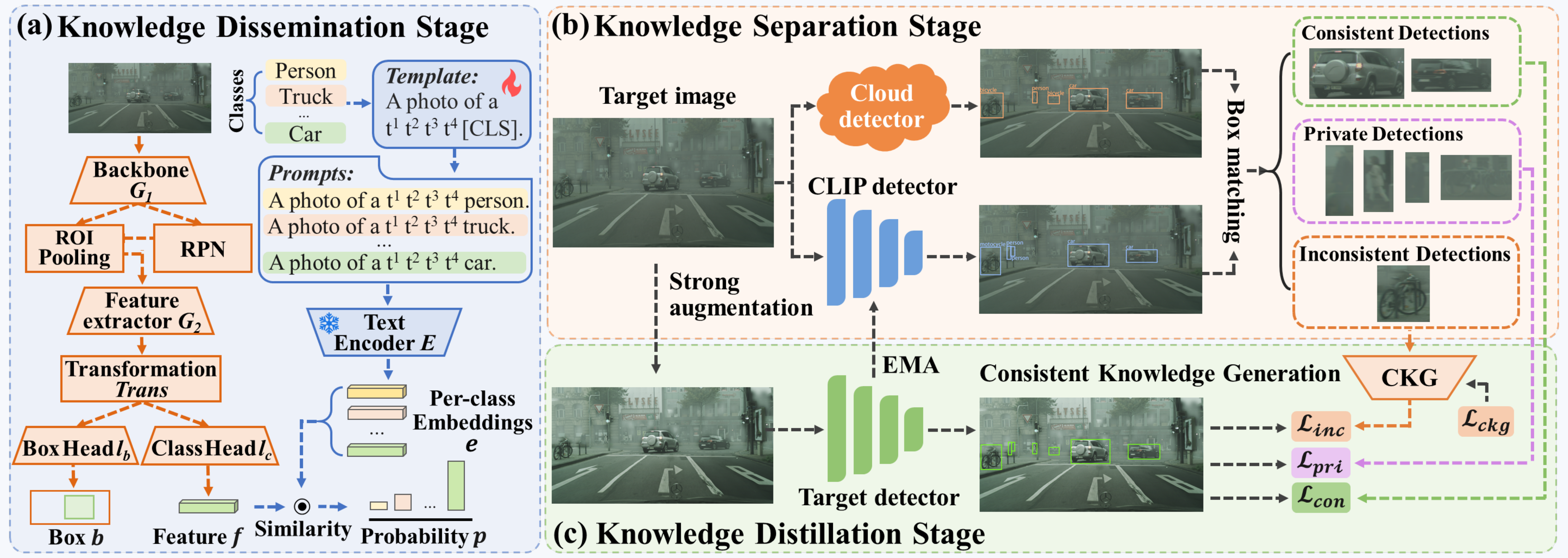
(a) Knowledge Dissemination (b) Knowledge Separation (c) Knowledge Distillation

1. ***Knowledge dissemination*** disseminates knowledge to a CLIP detector.
2. ***Knowledge separation*** separates detection results into three kinds.
3. ***Knowledge distillation*** fuses inconsistent detections by learning a CKG network using a self-promotion gradient direction alignment.

Contributions:

- Propose to explore a promising problem CODA.
- Propose a novel method COIN that acts in a divide-and-conquer manner.
- Propose a novel decision-level fusion strategy driven by gradient alignment.

Overall Pipeline of the Proposed COIN



- ***Knowledge dissemination pre-trains the CLIP detector with prompt learning:***

$$\min_{\theta_{clip}} \mathcal{L}_{RPN} + \mathcal{L}_{ROI} + \lambda \mathcal{L}^{1}_{align},$$

- ***Knowledge separation divides detections by box matching, resulting:***

$$\hat{\mathcal{P}} = \{(\boldsymbol{y}^{i}_{cld}, \boldsymbol{y}^{j}_{clip}) \,|\, \Gamma_{i,j} = 1, \boldsymbol{l}^{i}_{cld} = \boldsymbol{l}^{j}_{clip}\}, \tilde{\mathcal{P}} = \{(\boldsymbol{y}^{i}_{cld}, \boldsymbol{y}^{j}_{clip}) \,|\, \Gamma_{i,j} = 1, \boldsymbol{l}^{i}_{cld} \neq \boldsymbol{l}^{j}_{clip}\}$$

$$\mathcal{Q} = \{\boldsymbol{y}^{i}_{cld} \,|\, \Gamma_{i,*} = 0\} \cup \{\boldsymbol{y}^{j}_{clip} \,|\, \Gamma_{*,j} = 0\}$$

- ***Knowledge distillation distills detections to target detector, and fuses inconsistent detections with a CKG network, which is trained by a gradient direction alignment:***

$$\hat{g} = \nabla_{\theta_T} \|\hat{\boldsymbol{p}}_{stu} - \mathbb{I}(\hat{\boldsymbol{l}}_m)\|_2, \quad \tilde{g} = \nabla_{\theta_T} \|\tilde{\boldsymbol{p}}_{stu} - \tilde{\boldsymbol{p}}_{ckg}\|_2 \quad \min_{\theta_{ckg}} \mathcal{L}_{ckg} = (1 - sim(\hat{g}, \tilde{g})) + L_{kl}(\hat{\boldsymbol{p}}_{ckg}, \mathbb{I}(\hat{\boldsymbol{l}}_m))$$

Experiments on Benchmarks

Table 1: Results on **Foggy-Cityscapes** and **BDD100K** under GDINO. Object detection adaptation settings: U – Unsupervised, SF – Source-free, BB – Black-Box, C – Cloud. det: detector.

		Foggy-Cityscapes								
Methods	Type	Tuck	Car	Rder	Pson	Tain	Mcle	Bcle	Bus	mAP
MTOR [3]	U	21.9	44.0	41.4	30.6	**40.6**	28.3	35.6	38.6	35.1
ICR-CCR[59]	U	27.2	49.2	43.8	32.9	36.4	30.3	34.6	45.1	37.4
SED [35]	SF	25.5	44.5	40.7	33.2	22.2	28.4	34.1	39.0	33.5
LODS [33]	SF	27.3	48.8	45.7	34.0	19.6	33.2	37.8	39.7	35.8
A^{2}SFOD [10]	SF	28.1	44.6	44.1	32.3	29.0	31.8	38.9	34.3	35.4
IRG [53]	SF	24.4	51.9	45.2	37.4	25.2	31.5	41.6	39.6	37.1
LPU [9]	SF	24.0	55.4	**50.3**	39.0	21.2	30.3	**44.2**	46.0	38.8
BiMem [67]	BB	23.4	56.9	42.5	**42.2**	28.5	32.4	41.3	39.7	38.4
Cloud det [40]	C	**30.8**	47.5	18.6	34.3	21.0	**34.6**	41.1	**47.4**	34.4
CLIP [47]	C	9.7	28.6	11.5	19.5	1.1	12.8	17.9	21.9	15.4
CLIP det	C	8.2	46.9	27.5	34.1	16.5	24.9	31.5	36.2	28.2
COIN	C	27.4	**57.9**	42.3	41.6	25.9	32.7	41.2	43.1	**39.0**
Oracle	-	32.5	67.1	50.8	46.7	43.1	34.4	43.2	54.4	46.5

		BDD100K							
Methods	Type	Tuck	Car	Rder	Pson	Mcle	Bcle	Bus	mAP
SIGMA++ [34]	U	21.1	**65.6**	30.4	47.5	17.8	27.1	26.3	33.7
PT [7]	U	25.8	52.7	39.9	40.5	23.0	28.8	33.8	34.9
SED [35]	SF	20.6	50.4	32.6	32.4	18.9	25.0	23.4	29.0
PETS [39]	SF	19.3	62.4	34.5	42.6	17.0	26.3	16.9	31.3
A^{2}SFOD [10]	SF	33.2	36.3	**50.2**	26.6	28.2	24.4	22.5	31.6
BT [13]	SF	24.2	50.4	34.6	32.7	24.7	28.5	24.9	31.4
LPU [9]	SF	24.5	55.2	38.9	41.4	20.9	30.4	23.2	33.5
DRU [28]	SF	27.1	62.7	36.9	45.8	22.7	32.5	28.1	36.6
Cloud det [40]	C	38.7	46.0	11.4	**49.2**	**37.8**	**33.5**	**47.4**	37.7
CLIP [47]	C	23.6	31.1	4.4	6.7	18.0	11.4	27.7	17.5
CLIP det	C	34.3	53.4	14.1	31.7	28.7	24.6	36.7	31.9
COIN	C	**46.6**	56.8	23.5	45.5	32.0	33.0	40.6	**39.7**
Oracle	-	54.0	70.6	42.3	51.4	35.8	41.5	53.2	49.8

Experiments on Benchmarks

Table 3: Quantitative results on **KITTI** under GDINO. U – Unsupervised, C – Cloud. det: detector.

Type	Methods	AP of Car	Methods	AP of Car	Methods	AP of Car	Methods	AP of Car
U	DA-Faster [8]	64.1	MAF [23]	72.1	SCL [50]	72.7	ATF [24]	73.5
C	Cloud det [40]	45.2	CLIP [47]	62.1	CLIP det	79.9	**COIN**	**80.8**

Table 4: Quantitative results on **Cityscapes** and **Sim10K** under GDINO. C – Cloud. det: detector.

		Cityscapes									Sim10K
Methods	Type	Truck	Car	Rider	Person	Train	Mcycle	Bcycle	Bus	mAP	Car
Cloud det [40]	C	**37.5**	59.9	16.4	43.4	26.1	42.7	**48.4**	**62.6**	42.1	46.5
CLIP [47]	C	15.9	36.9	15.5	27.8	0.9	15.7	20.5	31.8	20.6	46.4
CLIP det	C	11.3	55.8	35.1	39.1	**33.8**	32.0	33.7	44.7	35.7	60.0
COIN	C	26.9	**64.3**	**47.5**	**47.0**	26.4	**44.4**	46.9	52.8	**44.5**	**62.4**
Oracle	-	34.7	70.4	56.4	50.5	43.0	38.7	46.9	58.9	49.9	79.2

Table 6: Ablation study for decision-level fusion of inconsistent detections on **Foggy-Cityscapes** under GDINO. Detections are filtered by $\pi = 0.7$ for fair comparison. det: detector. probs: probabilities. avg: average. s-avg: score-weighted average.

Methods	Truck	Car	Rider	Person	Train	Mcycle	Bcycle	Bus	mAP
COIN w/ cloud det probs	25.1	56.1	45.3	40.1	20.5	**33.7**	**41.3**	39.3	37.7
COIN w/ CLIP det probs	22.1	56.4	44.5	39.5	**26.8**	32.4	40.4	42.4	38.1
COIN w/ avg	24.8	55.8	44.1	39.9	21.7	32.8	40.9	**43.7**	38.0
COIN w/ s-avg	24.2	56.4	**45.9**	40.7	24.1	31.3	40.4	41.7	38.1
COIN w/ CKG	**27.4**	**57.9**	42.3	**41.6**	25.9	32.7	41.2	43.1	**39.0**

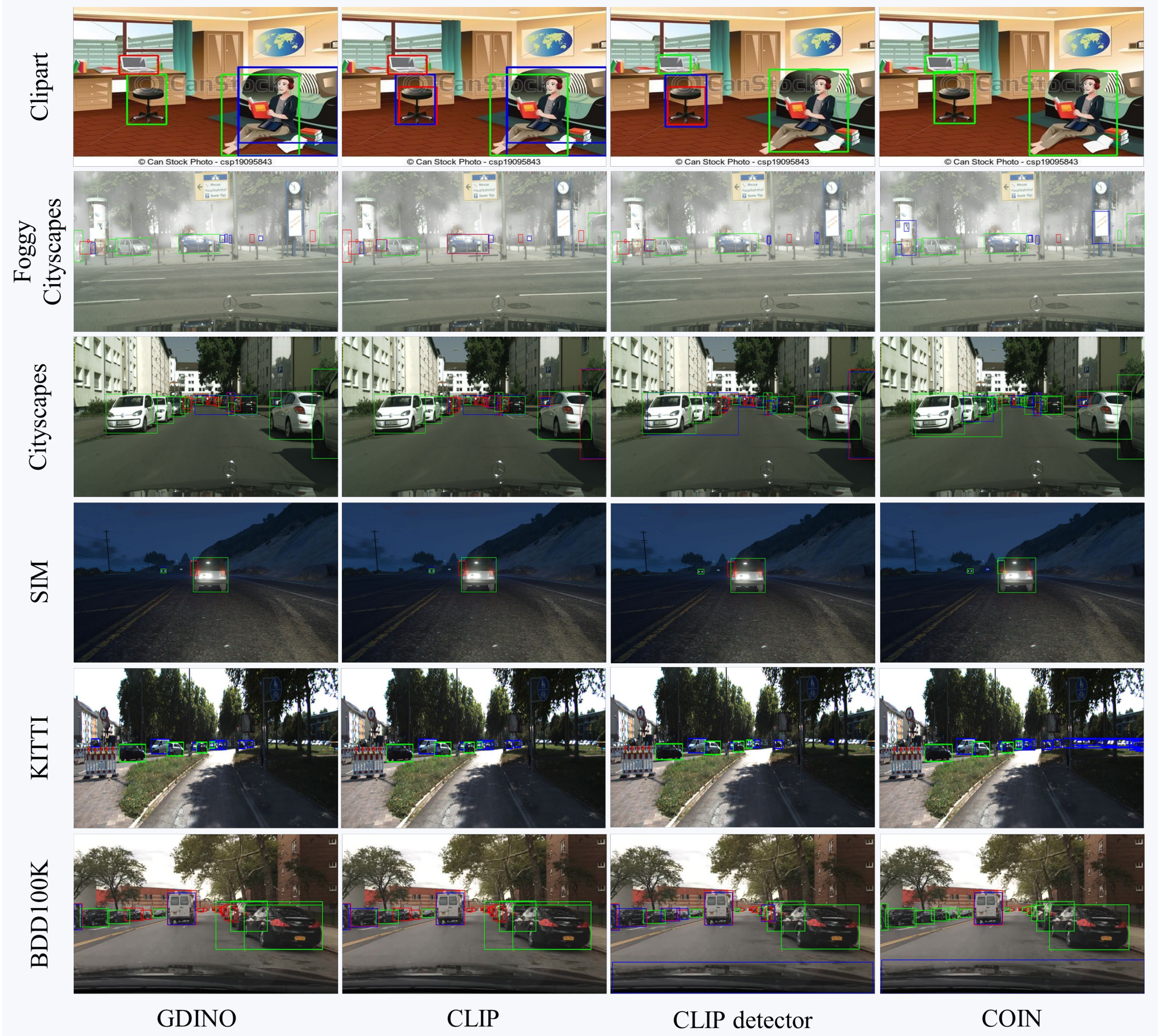


Figure 5: Qualitative results on Clipart, Foggy-Cityscapes, Cityscapes, SIM, KITTI and BDD100K. Green , red and blue boxes represent true positives (TP), false negatives (FN) and false positives (FP), respectively. *Zoom in for best view.*

Our COIN achieves the state-of-the-art performance on all datasets, and the proposed CKG works as above. For more information about this work, please refer to the full paper or slides with the following links.

Contact:

hotwindlsf@gmail.com

Project Code WeChat

Feel free to contact me!