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]

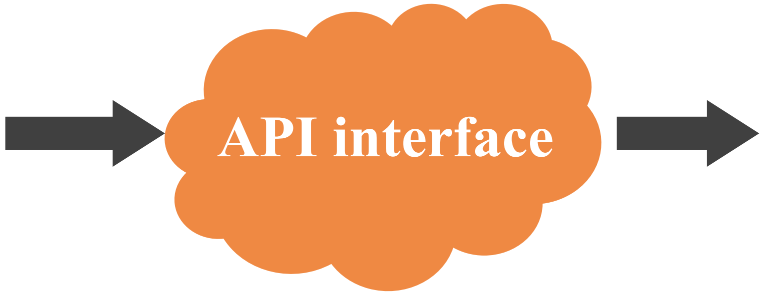
[1] University of Electronic Science and Technology of China
[2] University of Shanghai for Science and Technology
[3] University of Surrey

PROBLEM

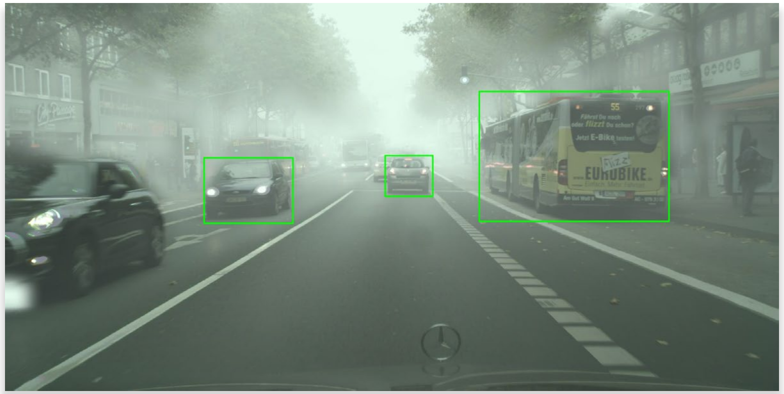
Cloud Object Detector Adaptation

Motivation

Target domain image



Large cloud detector



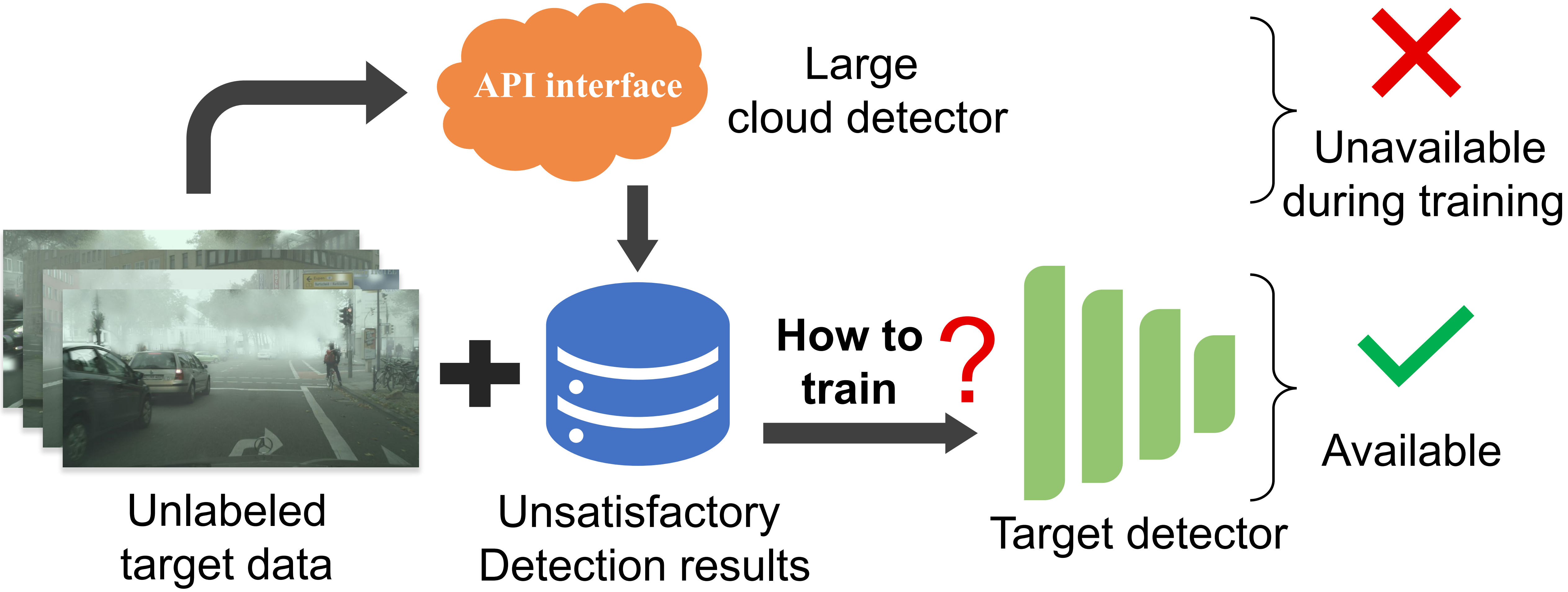
Detection results

Unsatisfactory results due to domain shift

Poor detection speed due to numerous parameters and network latency

Cloud Object Detector Adaptation

Definition



Cloud Object Detector Adaptation

Difference

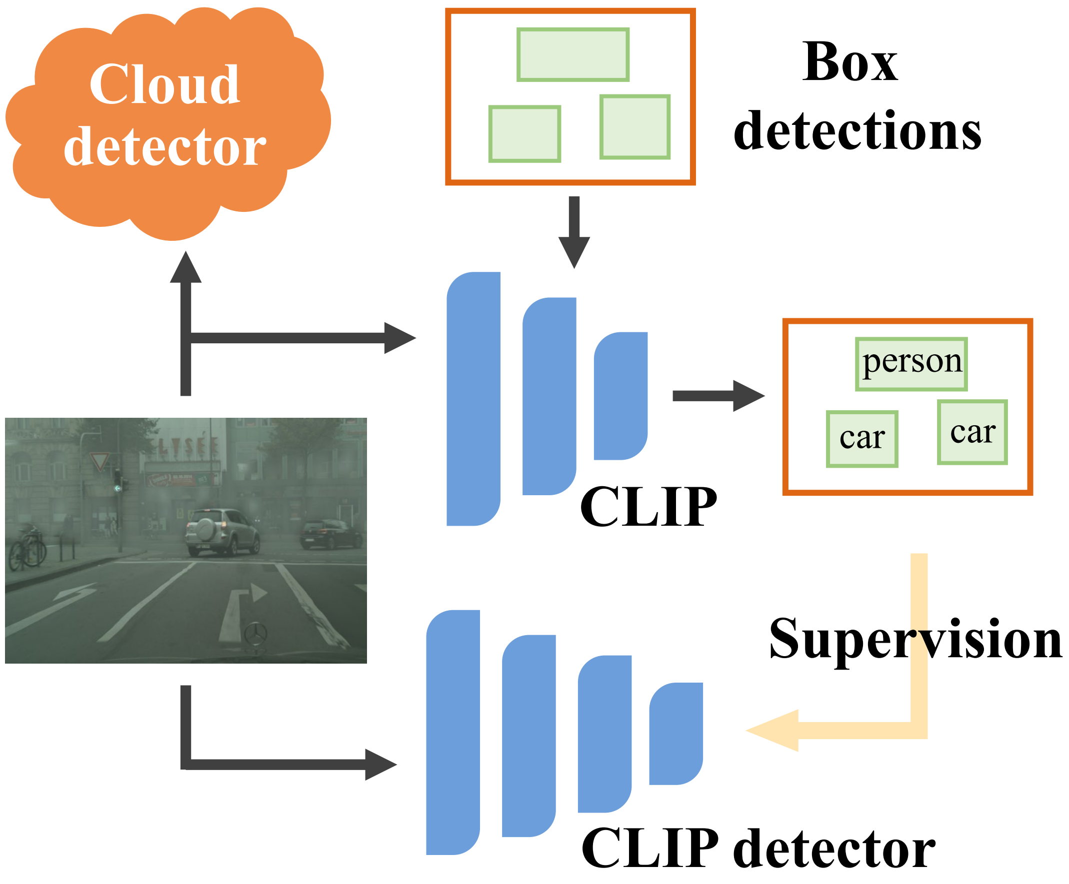
Conditions	UDAOD	SFOD	Black-box DAOD	**CODA**
Source data access	✓	✗	✗	✗
Source model access	✓	✓	✗	✗
Cloud API access	✗	✗	✓	✓
High domain similarity	✓	✓	✓	✗
Ability				
Flexible architecture	✗	✗	✓	✓
Open categories	✗	✗	✗	✓
Open scenarios	✗	✗	✗	✓

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

IDEA

IDEA

▶ Knowledge Dissemination   Knowledge Separation   Knowledge Distillation



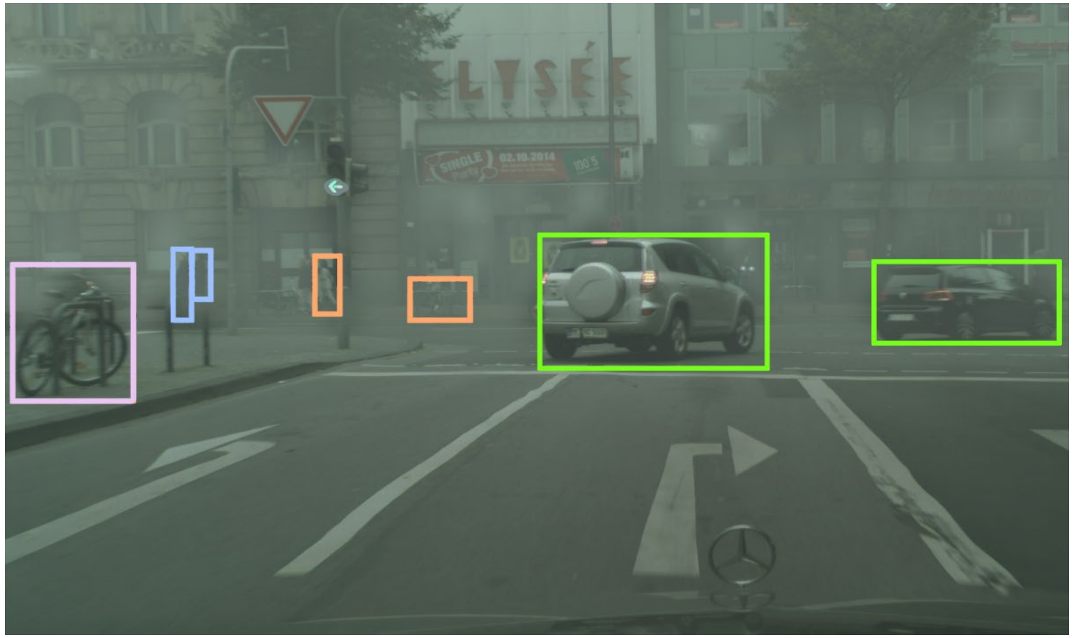
CLIP is leveraged to help adaptation.

Knowledge dissemination aims to **disseminate knowledge from cloud and CLIP to a CLIP detector**, as the existing domain shift and the lack of detection ability of CLIP

IDEA

Knowledge Dissemination ▶ Knowledge Separation Knowledge Distillation

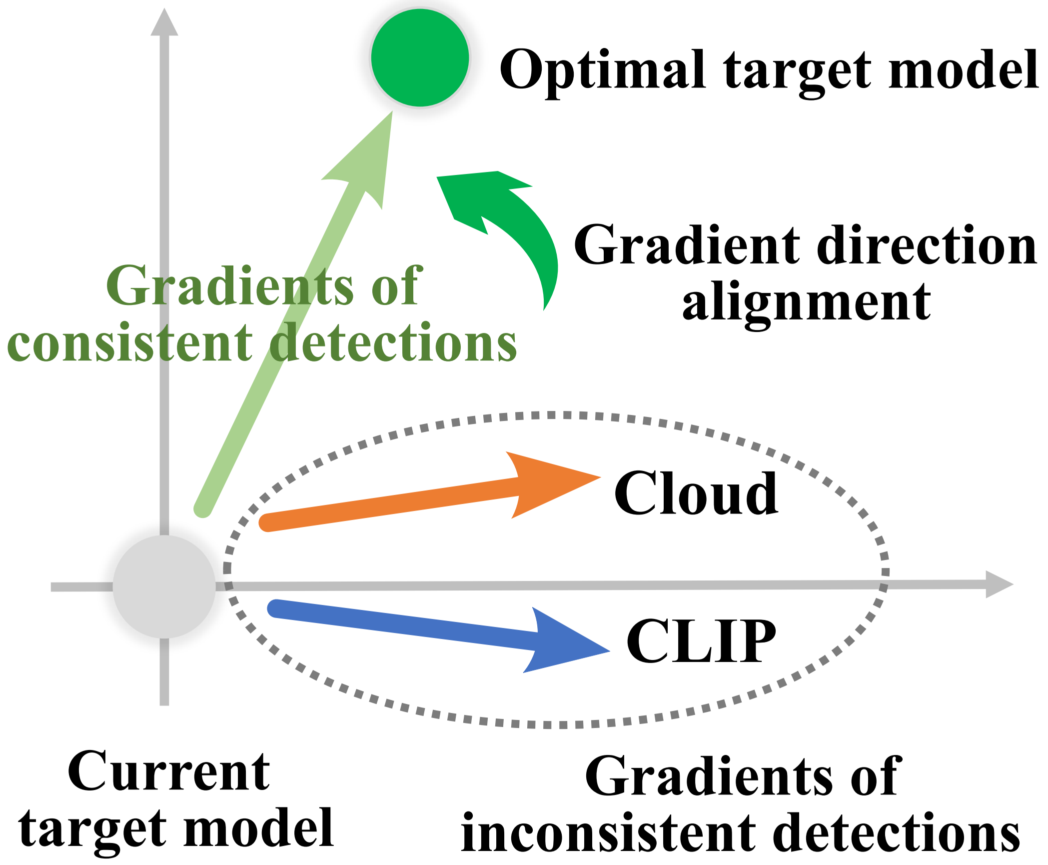
Knowledge separation and distillation adopts a divide-and-conquer manner.



Knowledge separation aims to **separate detection results** from cloud detector and CLIP detector into three parts: consistent, inconsistent, and private (cloud and CLIP) detections

IDEA

Knowledge Dissemination   Knowledge Separation   ▶ Knowledge Distillation



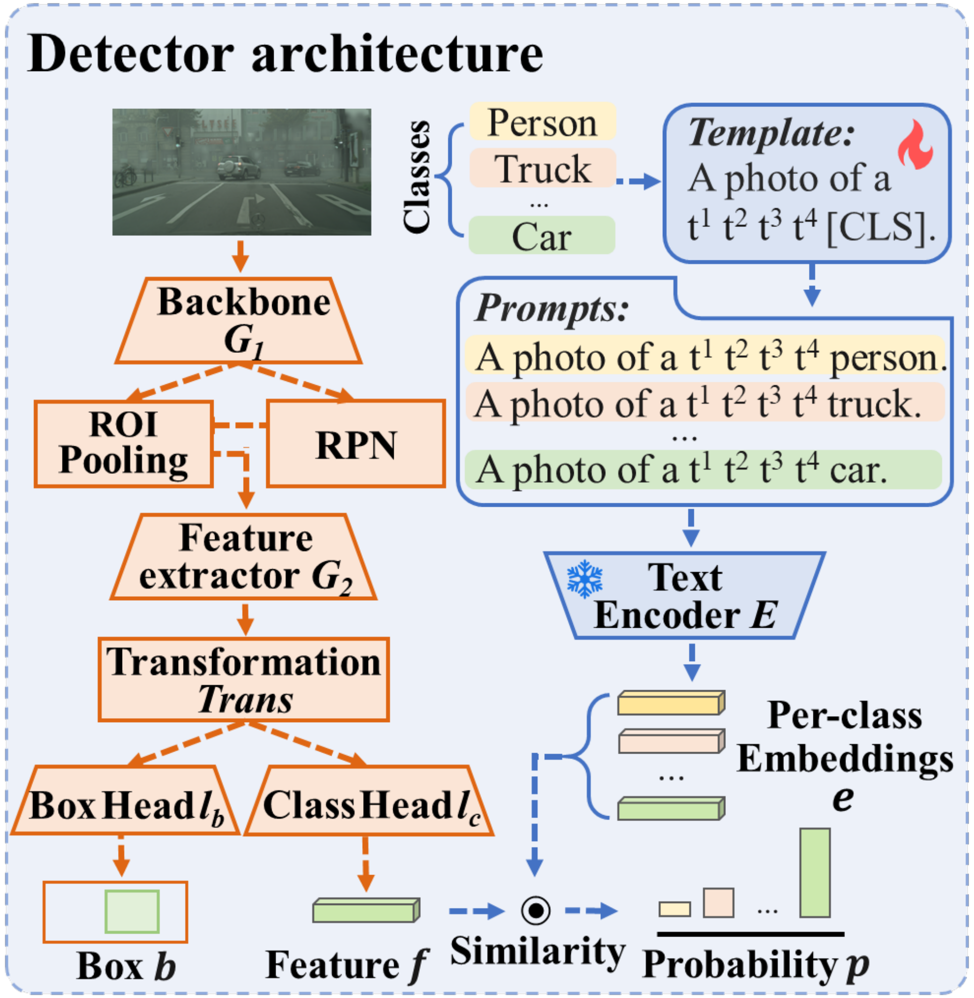
Decision-level fusion strategy

Knowledge distillation mainly focus on **fusing inconsistent detections**, by learning a Consistent Knowledge Generation network (CKG) using a **self-promotion gradient direction alignment**

DETAIL

DETAIL

▶ Knowledge Dissemination Knowledge Separation Knowledge Distillation



CLIP detector pre-train loss:

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

Prompt learning loss:

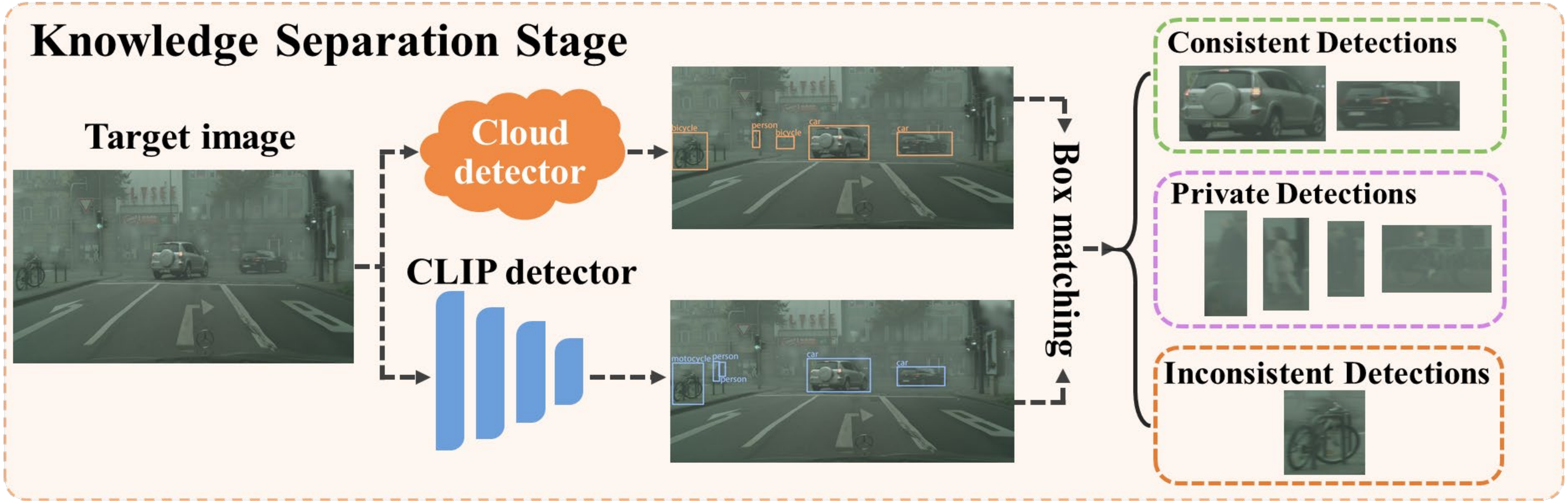
$$\mathcal{L}^1_{align} = ||\boldsymbol{e}_p - \boldsymbol{e}||_1.$$

Prototype updating by exponential moving average:

$$\boldsymbol{e}^i_p = \eta \cdot \boldsymbol{e}^i_p + (1-\eta) \cdot \mathbb{E}_{x \in \mathcal{D}} \frac{1}{|\mathcal{R}|} \sum \mathbb{1}(\boldsymbol{l} = i) \boldsymbol{f},$$

DETAIL

Knowledge Dissemination ▶ Knowledge Separation Knowledge Distillation



Box matching is used to categorize detections into consistent $\hat{\mathcal{P}}$, inconsistent $\tilde{\mathcal{P}}$, and private detections $\mathcal{Q}$:

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

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

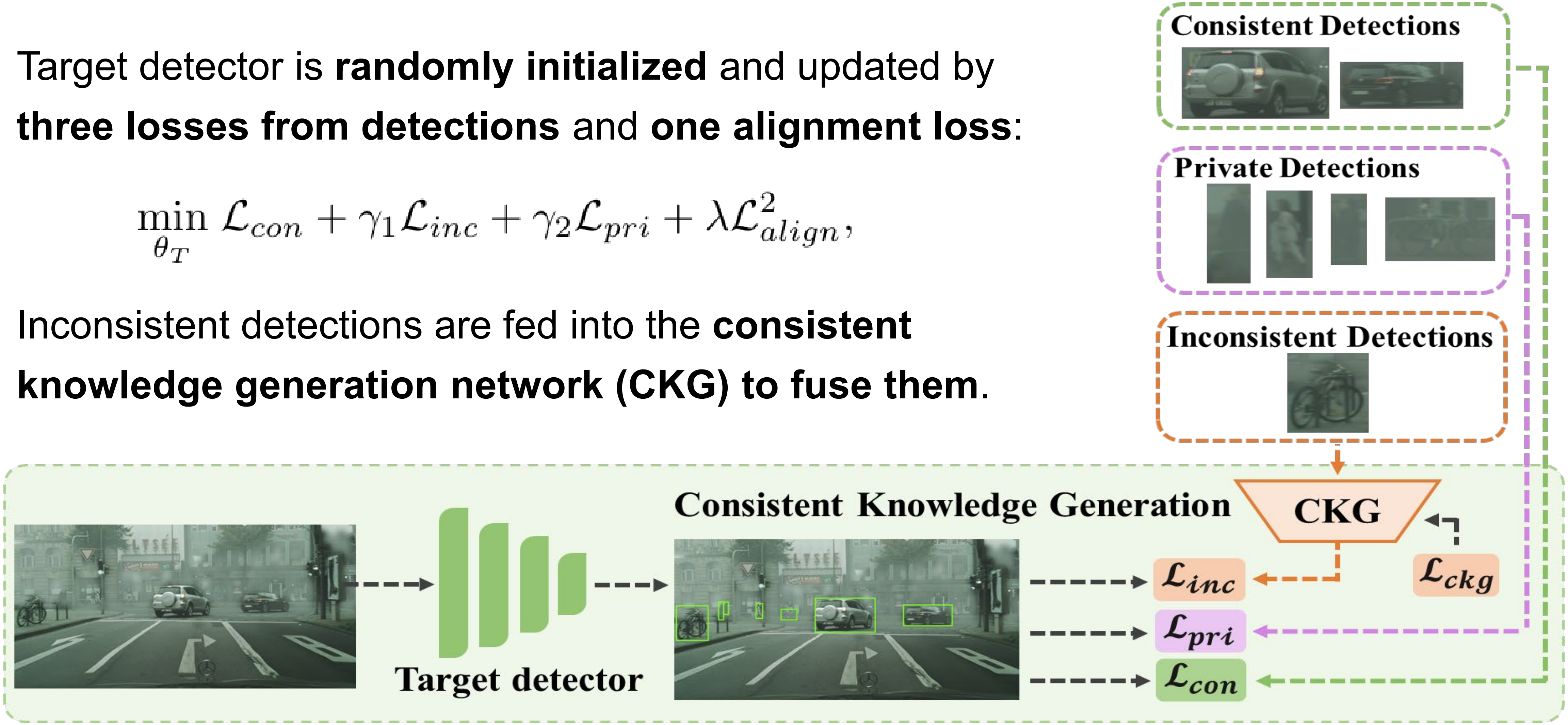
DETAIL

Knowledge Dissemination   Knowledge Separation   ▶ Knowledge Distillation

Target detector is **randomly initialized** and updated by **three losses from detections** and **one alignment loss**:

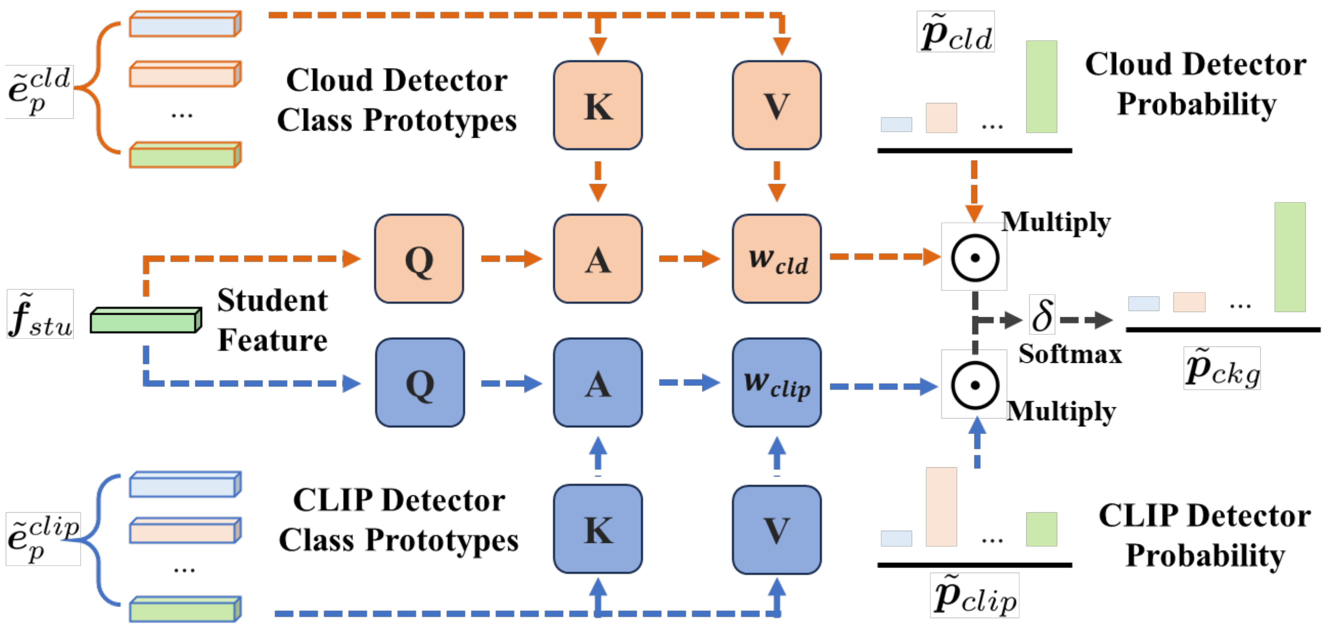
$$\min_{\theta_T} \mathcal{L}_{con} + \gamma_1 \mathcal{L}_{inc} + \gamma_2 \mathcal{L}_{pri} + \lambda \mathcal{L}^2_{align},$$

Inconsistent detections are fed into the **consistent knowledge generation network (CKG) to fuse them**.



DETAIL

Knowledge Dissemination Knowledge Separation ▶ Knowledge Distillation



Decision-level fusion flow of CKG network:

$$\boldsymbol{w}_{cld} = CA_1(\tilde{\boldsymbol{f}}_{stu}, \tilde{\boldsymbol{e}}_p^{cld}),$$

$$\boldsymbol{w}_{clip} = CA_2(\tilde{\boldsymbol{f}}_{stu}, \tilde{\boldsymbol{e}}_p^{clip}),$$

$$\tilde{\boldsymbol{p}}_{ckg} = \delta(\boldsymbol{w}_{cld} \odot \tilde{\boldsymbol{p}}_{cld} + \boldsymbol{w}_{clip} \odot \tilde{\boldsymbol{p}}_{clip}),$$

Self-promotion gradient direction alignment for training CKG network:

Gradient from consistent detections is used as the supervised signal

$$\hat{\boldsymbol{g}} = \nabla_{\theta_T} \|\hat{\boldsymbol{p}}_{stu} - \mathbb{I}(\hat{\boldsymbol{l}}_m)\|_2, \quad \tilde{\boldsymbol{g}} = \nabla_{\theta_T} \|\tilde{\boldsymbol{p}}_{stu} - \tilde{\boldsymbol{p}}_{ckg}\|_2,$$

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

EXPERIMENT

EXPERIMENT

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											BDD100K							
Methods	Type	Tuck	Car	Rder	Pson	Tain	Mcle	Bcle	Bus	mAP	Methods	Type	Tuck	Car	Rder	Pson	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	SIGMA++ [34]	U	21.1	**65.6**	30.4	47.5	17.8	27.1	26.3	33.7
ICR-CCR[59]	U	27.2	49.2	43.8	32.9	36.4	30.3	34.6	45.1	37.4	PT [7]	U	25.8	52.7	39.9	40.5	23.0	28.8	33.8	34.9
SED [35]	SF	25.5	44.5	40.7	33.2	22.2	28.4	34.1	39.0	33.5	SED [35]	SF	20.6	50.4	32.6	32.4	18.9	25.0	23.4	29.0
LODS [33]	SF	27.3	48.8	45.7	34.0	19.6	33.2	37.8	39.7	35.8	PETS [39]	SF	19.3	62.4	34.5	42.6	17.0	26.3	16.9	31.3
A^2SFOD [10]	SF	28.1	44.6	44.1	32.3	29.0	31.8	38.9	34.3	35.4	A^2SFOD [10]	SF	33.2	36.3	**50.2**	26.6	28.2	24.4	22.5	31.6
IRG [53]	SF	24.4	51.9	45.2	37.4	25.2	31.5	41.6	39.6	37.1	BT [13]	SF	24.2	50.4	34.6	32.7	24.7	28.5	24.9	31.4
LPU [9]	SF	24.0	55.4	**50.3**	39.0	21.2	30.3	**44.2**	46.0	38.8	LPU [9]	SF	24.5	55.2	38.9	41.4	20.9	30.4	23.2	33.5
BiMem [67]	BB	23.4	56.9	42.5	**42.2**	28.5	32.4	41.3	39.7	38.4	DRU [28]	SF	27.1	62.7	36.9	45.8	22.7	32.5	28.1	36.6
Cloud det [40]	C	**30.8**	47.5	18.6	34.3	21.0	**34.6**	41.1	**47.4**	34.4	Cloud det [40]	C	38.7	46.0	11.4	**49.2**	**37.8**	**33.5**	**47.4**	37.7
CLIP [47]	C	9.7	28.6	11.5	19.5	1.1	12.8	17.9	21.9	15.4	CLIP [47]	C	23.6	31.1	4.4	6.7	18.0	11.4	27.7	17.5
CLIP det	C	8.2	46.9	27.5	34.1	16.5	24.9	31.5	36.2	28.2	CLIP det	C	34.3	53.4	14.1	31.7	28.7	24.6	36.7	31.9
COIN	C	27.4	**57.9**	42.3	41.6	25.9	32.7	41.2	43.1	**39.0**	**COIN**	C	**46.6**	56.8	23.5	45.5	32.0	33.0	40.6	**39.7**
Oracle	-	32.5	67.1	50.8	46.7	43.1	34.4	43.2	54.4	46.5	Oracle	-	54.0	70.6	42.3	51.4	35.8	41.5	53.2	49.8

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**

EXPERIMENT

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 2: Results on **Clipart** under GDINO. Object detection adaptation settings: `SF` – Source-free, `U` – Unsupervised, `C` – Cloud. det: detector.

Methods	Type	Aero	Bcle	Bird	Boat	Botl	Bus	Car	Cat	Chair	Cow	Tble	Dog	Hrs	Bike	Pson	Plnt	Shep	Sofa	Tain	Tv	mAP
MGADA [75]	U	35.5	64.6	27.8	34.5	41.6	66.4	49.8	26.8	43.6	56.7	24.3	20.9	43.2	84.3	74.2	41.1	17.4	27.6	56.5	57.6	44.8
SIGMA++ [34]	U	36.3	54.6	40.1	31.6	58.0	60.4	46.2	33.6	44.4	66.2	25.7	25.3	44.4	58.8	64.8	55.4	36.2	38.6	54.1	59.3	46.7
CIGAR [41]	U	35.2	55.0	39.2	30.7	60.1	58.1	46.9	31.8	47.0	61.0	21.8	26.7	44.6	52.4	68.5	54.4	31.3	38.8	56.5	63.5	46.2
TFD [54]	U	27.9	64.8	28.4	29.5	25.7	64.2	47.7	13.5	47.5	50.9	50.8	21.3	33.9	60.2	65.6	42.5	15.1	40.5	45.5	48.6	41.2
LODS [33]	SF	43.1	61.4	40.1	36.8	48.2	45.8	48.3	20.4	44.8	53.3	32.5	26.1	40.6	86.3	68.5	48.9	25.4	33.2	44.0	56.5	45.2
IRG [53]	SF	20.3	47.3	27.3	19.7	30.5	54.2	36.2	10.3	35.1	20.6	20.2	12.3	28.7	53.1	47.5	42.4	9.1	21.1	42.3	50.3	31.5
WSCoL [61]	SF	42.8	57.2	34.9	43.2	41.5	78.9	44.7	3.0	50.8	54.0	40.1	19.6	48.7	**88.2**	61.2	46.5	30.3	43.0	52.6	46.2	46.4
Cloud det [40]	C	76.2	**91.8**	67.4	**62.7**	60.2	**82.2**	68.4	43.7	**77.9**	52.9	**69.8**	39.3	64.4	85.6	**88.1**	**78.9**	30.8	**56.9**	72.9	66.5	66.8
CLIP [47]	C	62.3	70.1	42.5	42.7	50.9	50.0	44.8	47.8	22.8	59.5	28.6	34.2	43.7	51.4	61.1	59.8	24.1	28.1	50.4	50.5	46.3
CLIP det	C	61.4	56.5	46.9	48.8	57.4	54.1	49.7	40.2	32.7	48.7	16.6	33.8	51.4	50.4	62.8	60.6	25.7	28.8	43.9	52.6	46.2
COIN	C	**82.0**	87.6	**70.1**	58.1	**63.7**	63.8	**68.7**	**55.2**	70.5	**76.3**	59.0	**58.8**	**68.6**	82.9	88.0	67.3	**43.1**	53.3	**78.7**	**73.4**	**68.5**
Oracle	-	100	99.1	98.7	96.5	96.3	100	99.5	99.7	100	99.9	99.4	100	99.4	100	99.8	99.4	100	100	100	100	99.4

EXPERIMENT

Ablation study

Table 5: Ablation study on **Foggy-Cityscapes** and **Cityscapes** under GDINO. det: detector.

Methods	Losses				mAP	
	$\mathcal{L}_{align}$	$\mathcal{L}_{con}$	$\mathcal{L}_{inc}$	$\mathcal{L}_{pri}$	Foggy-Cityscapes	Cityscapes
Cloud det [40]	×	×	×	×	34.4	42.1
CLIP [47]	×	×	×	×	15.4	20.6
	×	×	×	×	27.4	35.1
CLIP det	√	×	×	×	28.2	35.7
	×	√	×	×	36.7	41.7
	√	√	×	×	37.1	42.4
	√	√	×	√	37.5	42.9
	√	√	√	×	38.4	43.8
COIN	√	√	√	√	39.0	44.5

EXPERIMENT

Ablation study

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**

Thank you!

To learn more about this paper,

welcome to our poster session 2