JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Semi-DETR: Semi-Supervised Object Detection with Detection Transformers

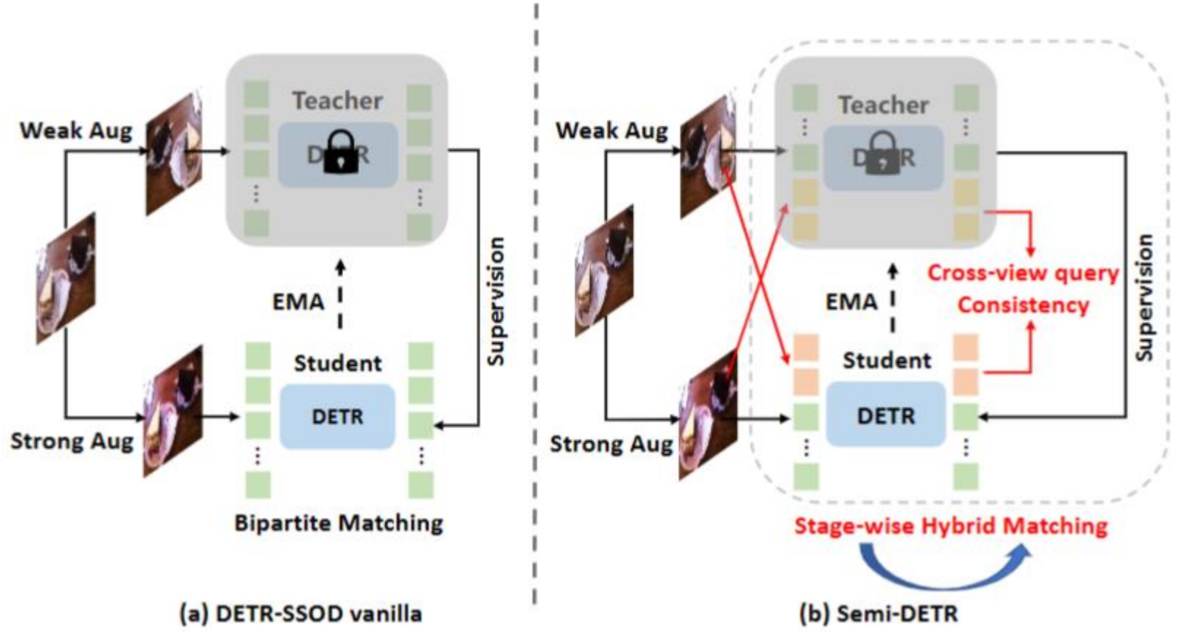
Jiacheng Zhang*, Xiangru Lin*, Wei Zhang, Kuo Wang, Xiao Tan,
Junyu Han, Errui Ding, Jingdong Wang, Guanbin Li

Poster: THU-PM-306

Sun Yat-Sen University   Baidu Inc

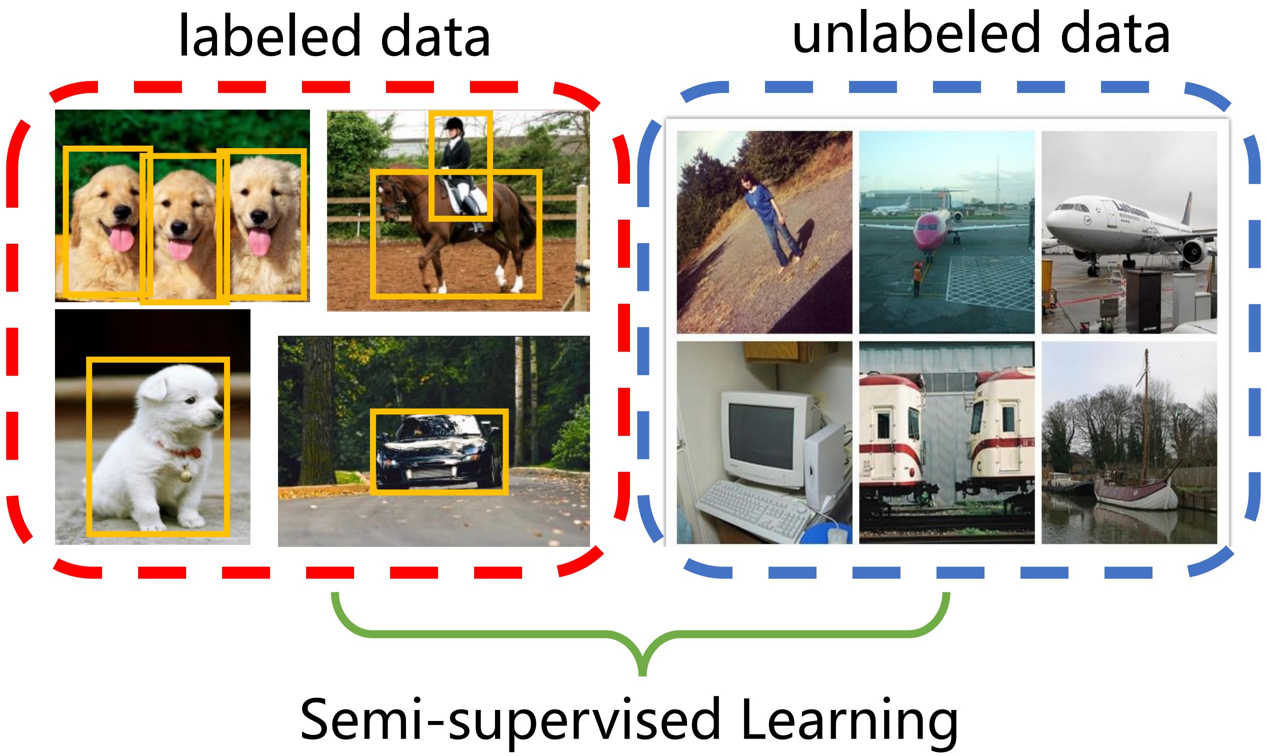
Semi-DETR Overview

- Semi-DETR is the first semi-supervised object detection method tailored for detection transformers.
- Semi-DETR can be used with various DETR-based detectors, e.g. Deformable DETR, DINO, etc.
- Semi-DETR achieves SOTA on both COCO and PASCAL VOC dataset.



Semi-Supervised Object Detection(SSOD)

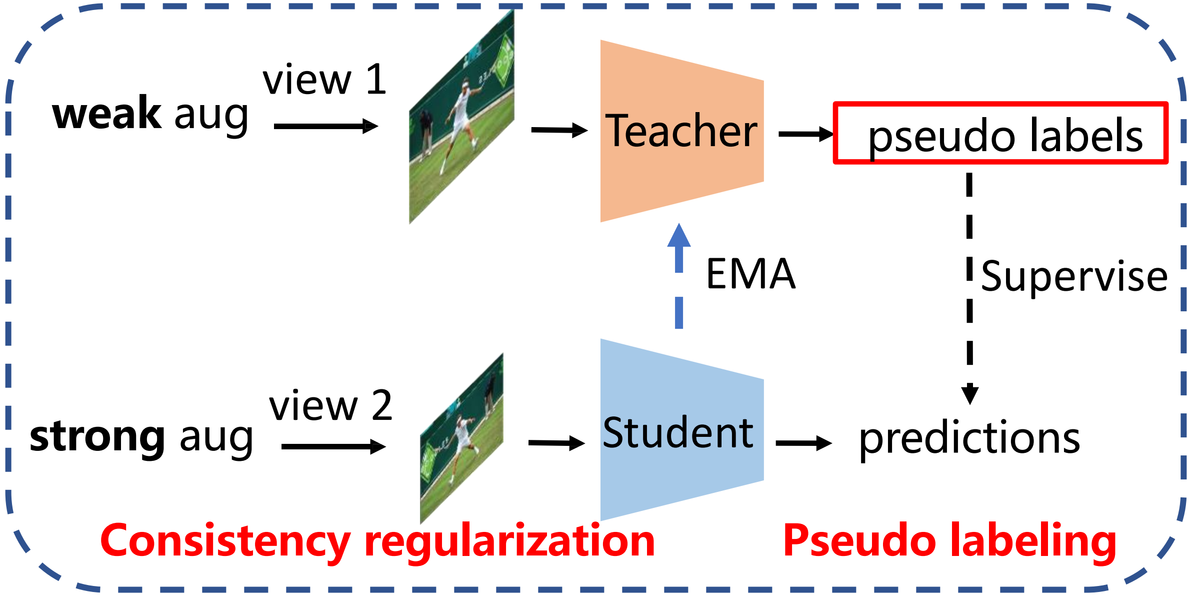
- Background
  - Problem definition



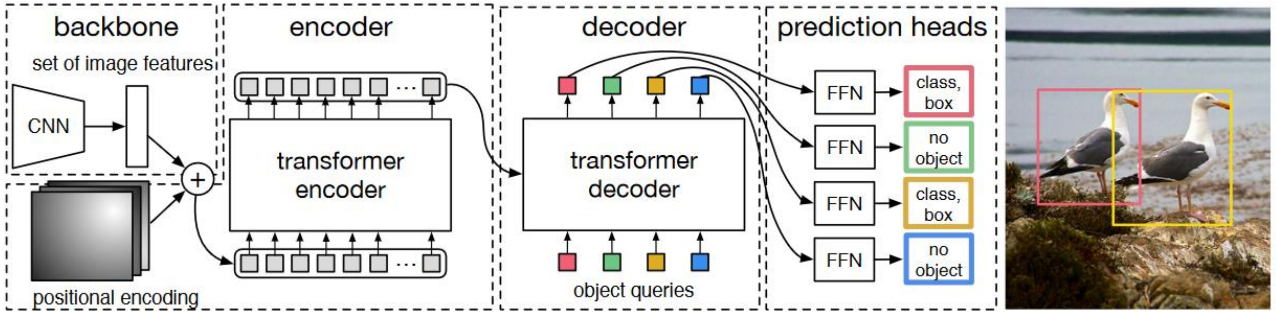
Settings:
1. labeled data is limited: Taking **10% coco** as labeled data, and **the rest** as unlabeled data;
2. labeled data is abundant: Taking **full coco** (118k images) as labeled data, and **unlabeled2017** (123k images) as unlabeled data;

Semi-Supervised Object Detection(SSOD)

- Background
  - Current Works



Mean-Teacher Framework



End-to-End Object Detection with Transformers(DETR)

- ✓ NMS Free
- ✓ Anchor Free
- ✓ Set-to-Set Prediction

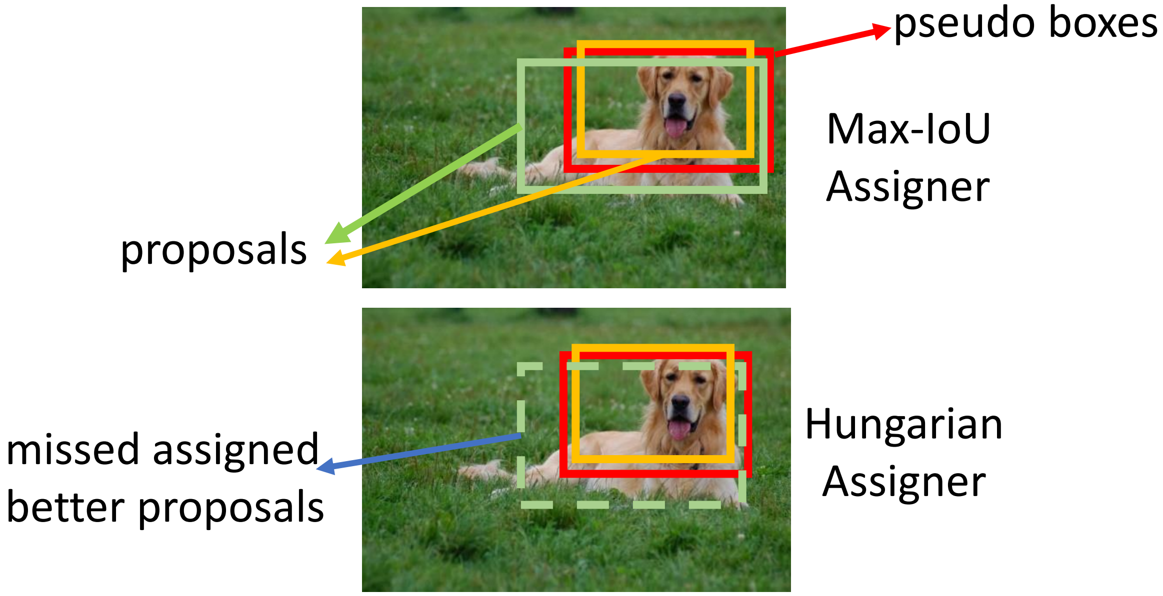
Soft-Teacher(Two-Stage Detector),Dense-Teacher(One-Stage Detector) ...

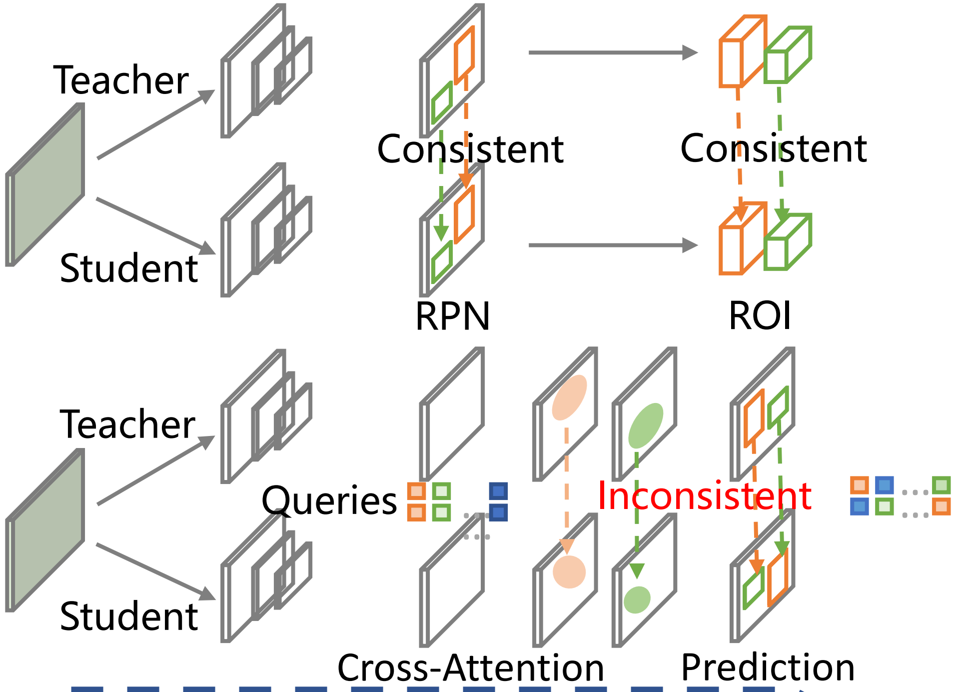
Problem: Anchor generation, Label assignment by various rules, NMS ...

[Xu 2021] End-to-End Semi-Supervised Object Detection with Soft Teacher ICCV2021
[Zhou 2022] Dense Teacher: Dense Pseudo-Labels for Semi-supervised Object Detection ECCV2022
[Nicolas 2020] End-to-End Object Detection with Transformers ECCV2020

Semi-DETR: SSOD with Detection Transformers

- Motivation
  - Bipartite matching make NMS-free but cause training inefficiency.
  - Set-to-Set Prediction cause consistency regularization infeasible.



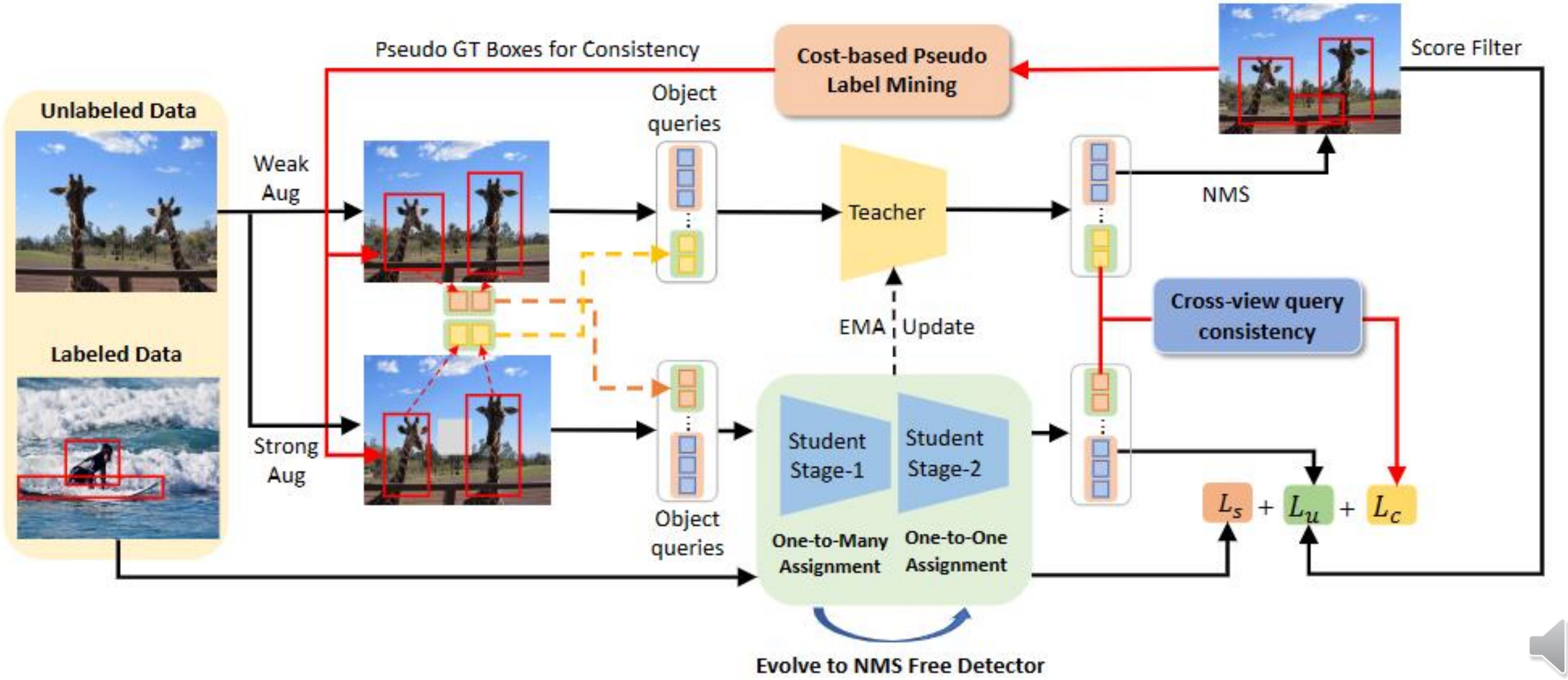


one-to-one assigner treats the better proposal as background while one-to-many assigner not.

attention mechanism in the decoder makes the updated proposal features lack explicit correspondence.

Semi-DETR: SSOD with Detection Transformers

- Method



Semi-DETR: SSOD with Detection Transformers

- Stage-wise Hybrid Matching(*SHM*)

Stage-1

$$\hat{\sigma}_{o2m} = \left\{ \underset{\sigma_i \in C_N^M}{\arg\min} \sum_{j=1}^{M} \mathcal{C}_{\text{match}}\left(\hat{y}_i^t, \hat{y}^s_{\sigma_i(j)}\right) \right\}_{i=1}^{|\hat{y}^t|}.$$

$$L_{cls}^{o2m} = \sum_{i=1}^{N_{\text{pos}}} \boxed{|\hat{m}_i - s_i|^{\gamma}}\, BCE\left(s_i, \boxed{\hat{m}_i}\right) + \sum_{j=1}^{N_{\text{neg}}} s_j^{\gamma} BCE\left(s_j, 0\right) \quad (4)$$

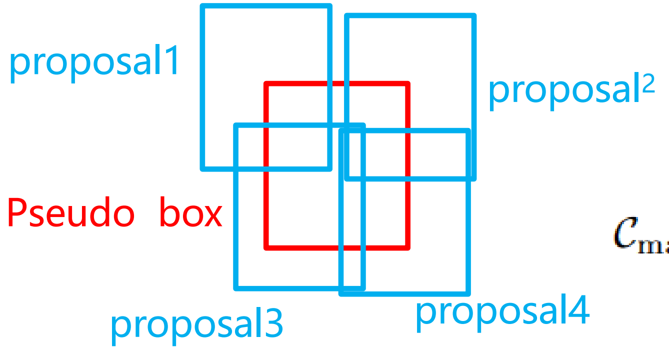
soft weighting soft target

$$L_{reg}^{o2m} = \sum_{i=1}^{N_{\text{pos}}} \boxed{\hat{m}_i} L_{GIoU}\left(b_i, \hat{b}_i\right) + \sum_{i=1}^{N_{\text{pos}}} \boxed{\hat{m}_i} L_{L_1}\left(b_i, \hat{b}_i\right) \quad (5)$$

$$L^{o2m} = L_{cls}^{o2m} + L_{reg}^{o2m} \quad (6)$$

Stage-2

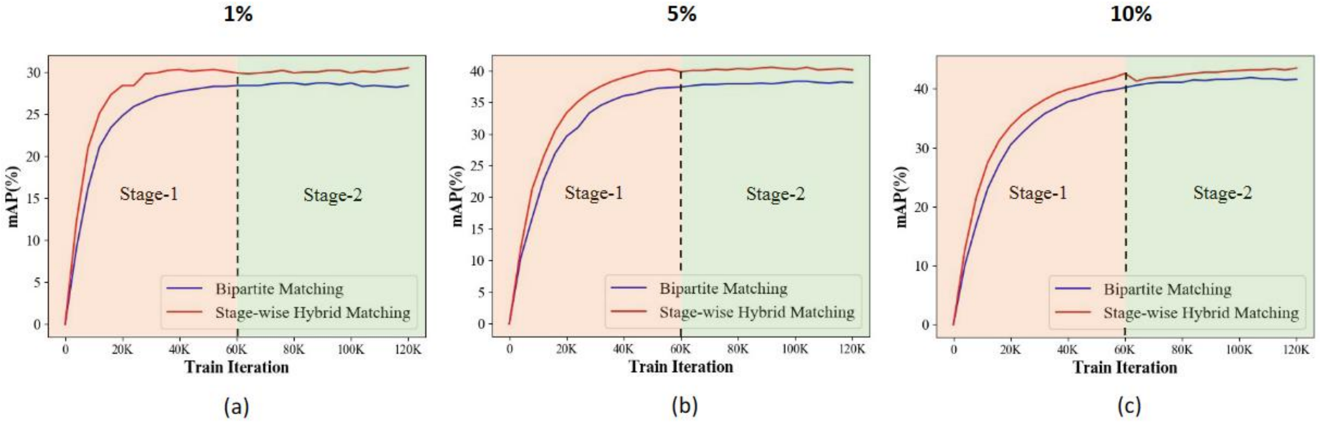
$$\hat{\sigma}_{o2o} = \underset{\sigma \in \xi_N}{\arg\min} \sum_{i=1}^{N} \mathcal{C}_{\text{match}}\left(\hat{y}_i^t, \hat{y}^s_{\sigma(i)}\right)$$



$$m = s^{\alpha} \cdot u^{\beta}$$

$$\mathcal{C}_{\text{match}}\left(\hat{y}_i^t, \hat{y}_j^s\right) = -m_{ij} = s_{ij}^{\alpha} \cdot u_{ij}^{\beta}$$

where, ***s*** is the classification score of the proposal predicted by student, ***u*** is the IoU between the predicted box and the pseudo box.

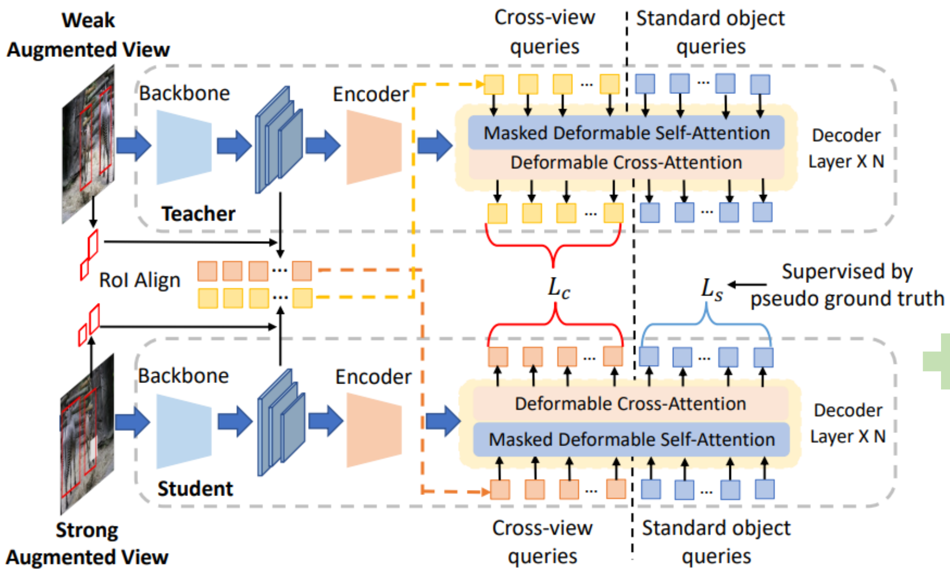


$$L = \mathbb{I}(t \le T_1) \cdot \left(L_{sup}^{o2m} + w_u \cdot L_{unsup}^{o2m}\right) + \mathbb{I}(t > T_1) \cdot \left(L_{sup}^{o2o} + w_u \cdot L_{unsup}^{o2o}\right) + w_c \cdot L_c$$

Our total loss function

Semi-DETR: SSOD with Detection Transformers

- Cross-view query consistency(*CQC*)
  - How to find the correspondence for proposal features?
    - Bipartite matching? × Time consuming when query num N is large(e.g. 300)
    - Semantic prior? √ Enforce the decoder to learning the semantic invariance



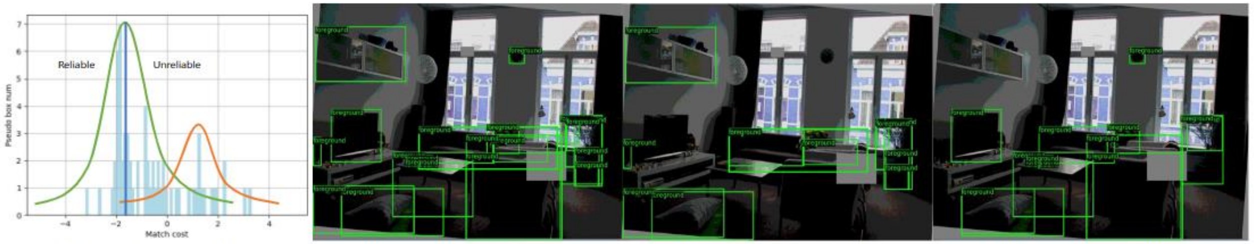
***Probs**: The sparse cross-view queries results in limited improvement!*

Cost-based Pseudo Label Mining(*CPM*)

$$C_{ij} = \lambda_1 C_{Cls}(p_i, \hat{p_j}) + \lambda_2 C_{GIoU}(b_i, \hat{b_j}) + \lambda_3 C_{L_1}(b_i, \hat{b_j})$$

$$P(c|\theta) = w_r \mathcal{N}_r(c, \mu_r, \sigma_r) + w_u \mathcal{N}_u(c, \mu_u, \sigma_u)$$

$$\tau_c = \arg\max_{c} P_{reliable}(c|c, \theta)$$



(a) Distribution of Cost (b) Initial Pseudo Boxes (c) Unreliable Pseudo Boxes (d) Reliable Pseudo Boxes

Semi-DETR: SSOD with Detection Transformers

- Main Results

Category	Method	ID	1%	5%	10%
Two-Stage	Unbiased Teacher	1	20.75 ± 0.12	28.27 ± 0.11	31.50 ± 0.10
	Soft-Teacher	2	20.46 ± 0.39	30.74 ± 0.08	34.04 ± 0.14
	PseCo	3	22.43 ± 0.36	32.50 ± 0.08	36.06 ± 0.24
One-Stage	DSL	4	22.03 ± 0.28	30.87 ± 0.24	36.22 ± 0.18
	Dense Teacher	5	22.38 ± 0.31	33.01 ± 0.14	37.13 ± 0.12
	Unbiased Teacher v2	6	22.71 ± 0.42	30.08 ± 0.04	32.61 ± 0.03
End-to-End	Omi-DETR(Def-DETR)	7	18.60	30.20	34.10
	Def-DETR(Sup only)	8	11.00 ± 0.24	23.70 ± 0.13	29.20 ± 0.11
	Def-DETR SSOD(Baseline)	9	19.40 ± 0.31	31.10 ± 0.21	34.80 ± 0.09
	Semi-DETR(Def-DETR)	10	**25.20 ± 0.23**	**34.50 ± 0.18**	**38.10 ± 0.14**
	DINO(Sup only)	11	18.00 ± 0.21	29.50 ± 0.16	35.00 ± 0.12
	DINO SSOD (Baseline)	12	28.40 ± 0.21	38.00 ± 0.13	41.60 ± 0.11
	Omi-DETR(DINO)	13	27.60	37.70	41.30
	Semi-DETR(DINO)	14	**30.50 ± 0.30**	**40.10 ± 0.15**	**43.50 ± 0.10**

COCO Partial

Method	100%
Unbiased Teacher	$40.2 \xrightarrow{+1.1} 41.3$
Soft-Teacher	$40.9 \xrightarrow{+3.6} 44.5$
PseCo	$41.0 \xrightarrow{+5.1} 46.1$
DSL	$40.2 \xrightarrow{+3.6} 43.8$
Dense Teacher	$41.2 \xrightarrow{+3.6} 46.1$
Semi-DETR(Def-DETR)	$\mathbf{43.6} \xrightarrow{+3.6} \mathbf{47.2}$
Semi-DETR(DINO)	$\mathbf{48.6} \xrightarrow{+1.8} \mathbf{50.4}$

COCO Full

Category	Method	AP_{50}	$AP_{50:95}$
Two-Stage	Unbiased Teacher	77.37	48.69
	Soft-Teacher	-	-
	PseCo	-	-
One-Stage	DSL	80.70	56.80
	Dense Teacher	79.89	55.87
	Unbiased Teacher v2	81.29	56.87
End-to-End	Def-DETR(Sup only)	74.50	46.20
	Def-DETR SSOD(Baseline)	78.90	53.40
	Semi-DETR(Def-DETR)	**83.50**	**57.20**
	DINO(sup only)	81.20	59.60
	DINO SSOD (Baseline)	84.30	62.20
	Semi-DETR(DINO)	**86.10**	**65.20**

PASCAL VOC

Semi-DETR: SSOD with Detection Transformers

- Ablation Study

Table 4. Component effectiveness of Semi-DETR. SHM denotes the Stage-wise Hybrid Matching, CQC means Cross-view Query Consistency, and CPM represents Cost-based Pseudo Label Mining, respectively.

ID	SHM	CQC	CPM	mAP	AP_{50}	AP_{75}
1				41.6	58.3	45.1
2	✓			42.7	59.3	46.2
3	✓	✓		43.1	59.6	46.6
4	✓	✓	✓	**43.5**	**59.7**	**46.8**

Table 7. Effects of the training iteration T_1 of the first stage using one-to-many assignment strategy in Stage-wise Hybrid Matching.

T_1	40k	60k	80k	100k	120k
mAP	42.9	**43.5**	43.2	43.0	**44.0**
NMS-Free	Y	Y	Y	Y	N

Semi-DETR can achieve better performance and retain the NMS-Free

Table 5. Effects of different methods to filter pseudo labels for cross-view consistency training.

Method	mAP	Precision	Recall
Fixed(0.4)	42.8	81.5%	41.3%
Top-K(K=9)	42.9	80.2%	39.4%
Mean + Std	43.1	60.2%	54.0%
Cost-based GMM	43.5	77.6%	52.1%

The effect of different pseudo boxes used in cross-view query consistency.

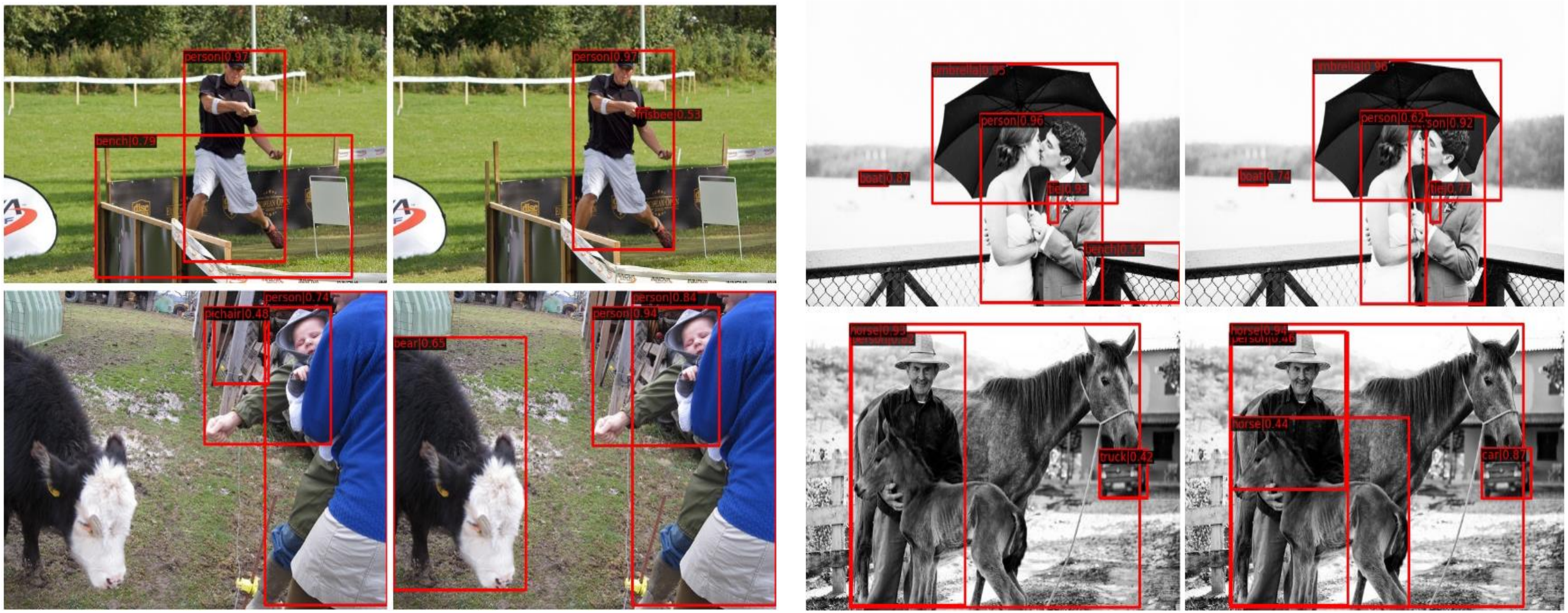
Table 3. Experiments about usage extension of the pseudo labels from Cost-baed Pseudo Label mining(CPM). Cls means classification training and Reg means regression training. Consistency represents the cross-view query consistency.

Method	Cls + Reg	Consistency	mAP
CPM(Ours)		✓	**43.5**
CPM(Extension)	✓	✓	42.4

The pseudo labels are the boxes covering the semantic areas, but not the one localized precisely.

Semi-DETR: SSOD with Detection Transformers

- Visualization



First column: supervised baseline; Second column: ours

Conclusion

- We present Semi-DETR, the first semi-supervised object detection method tailor designed for detection transformers.
- Semi-DETR analyze and solve the main obstacles which hinder the performance improvement for the SSOD with detection transformers.
- Semi-DETR can be applied with various detection transformers, for example Deformable DETR, DINO, etc. and achieves the new state-of-the-art performance on both COCO and PASCAL VOC dataset.