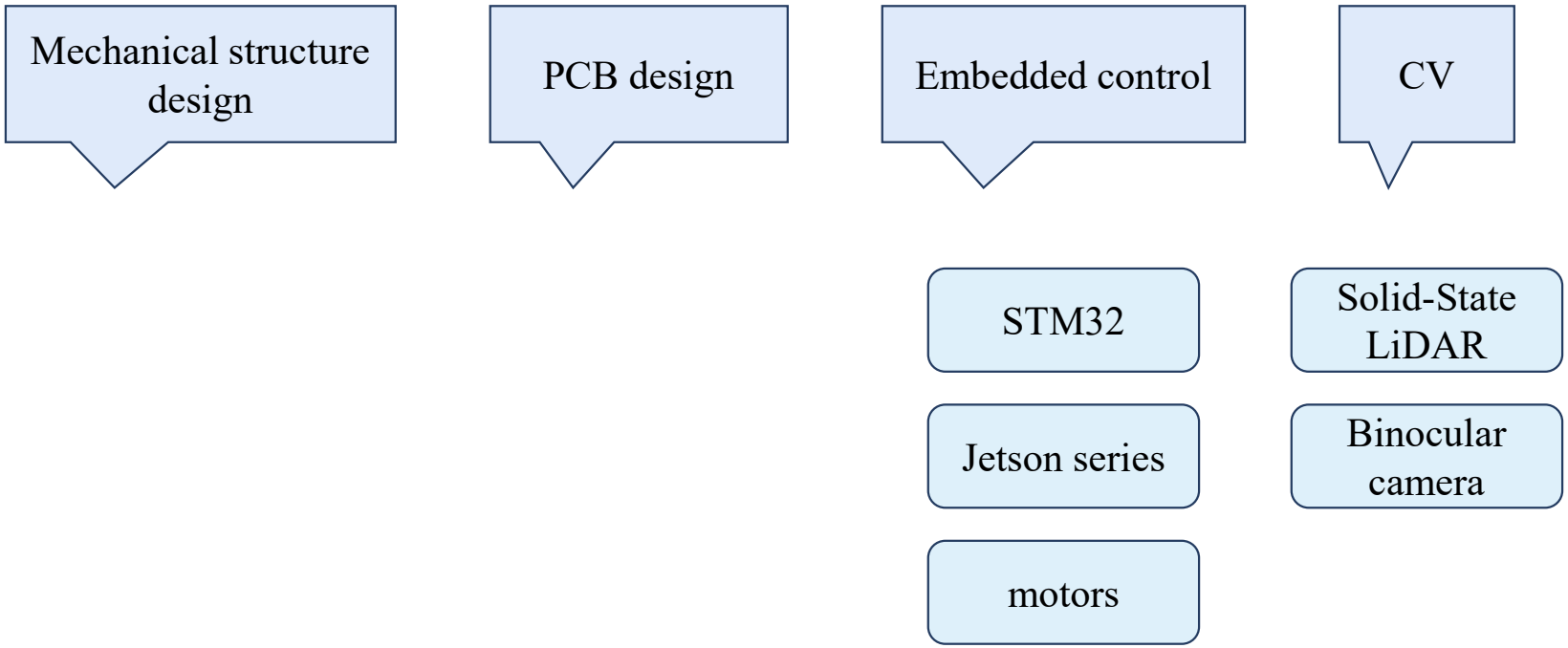
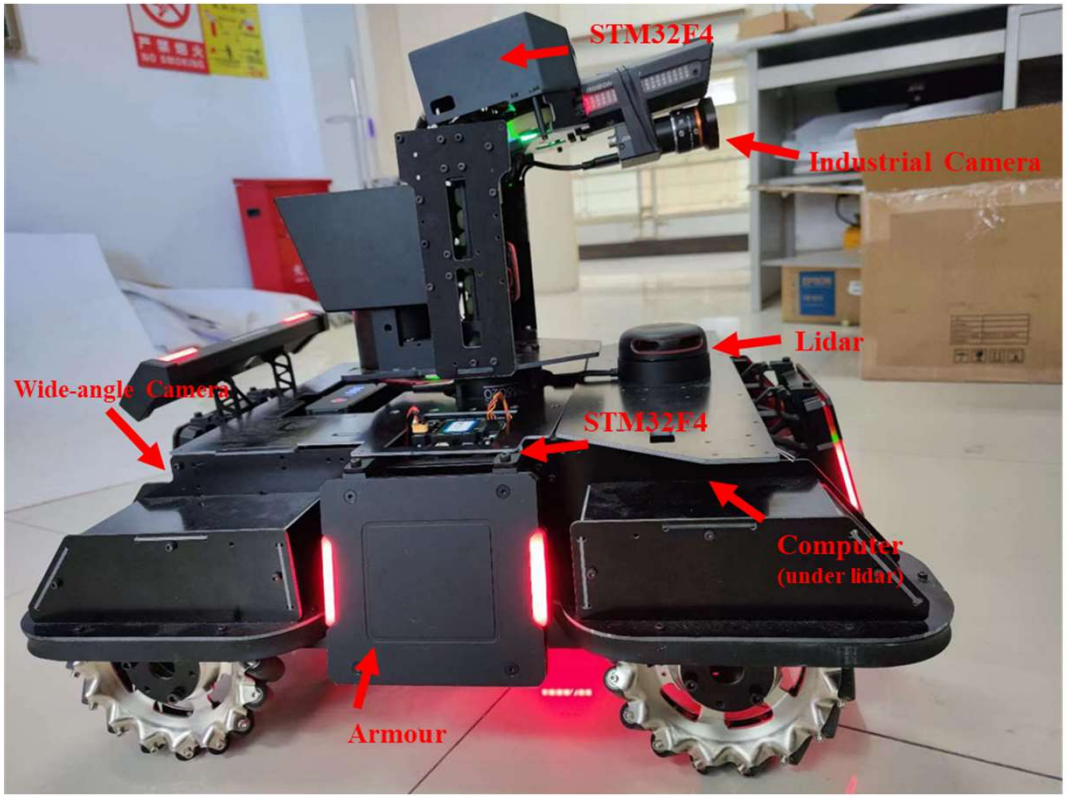
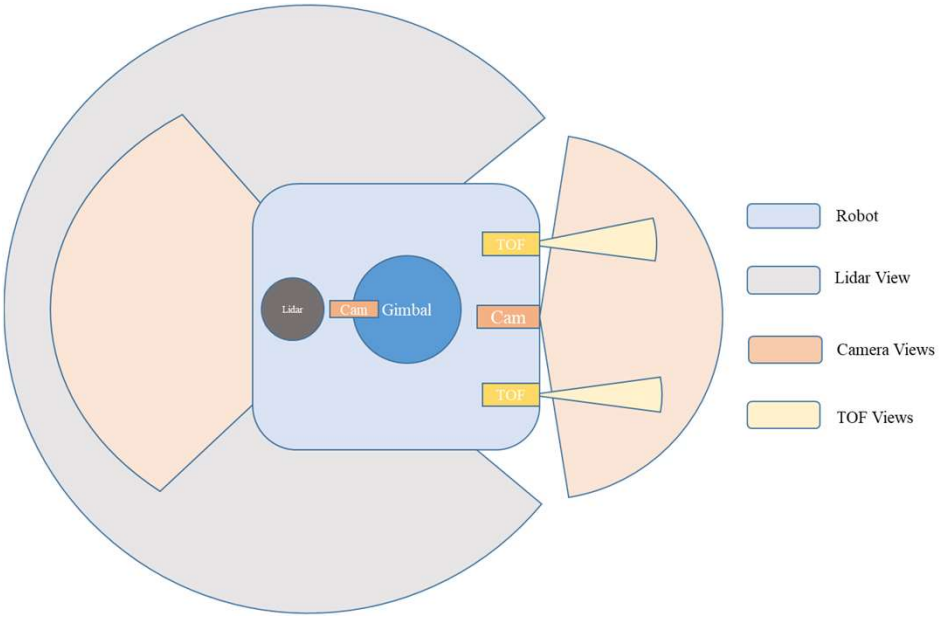
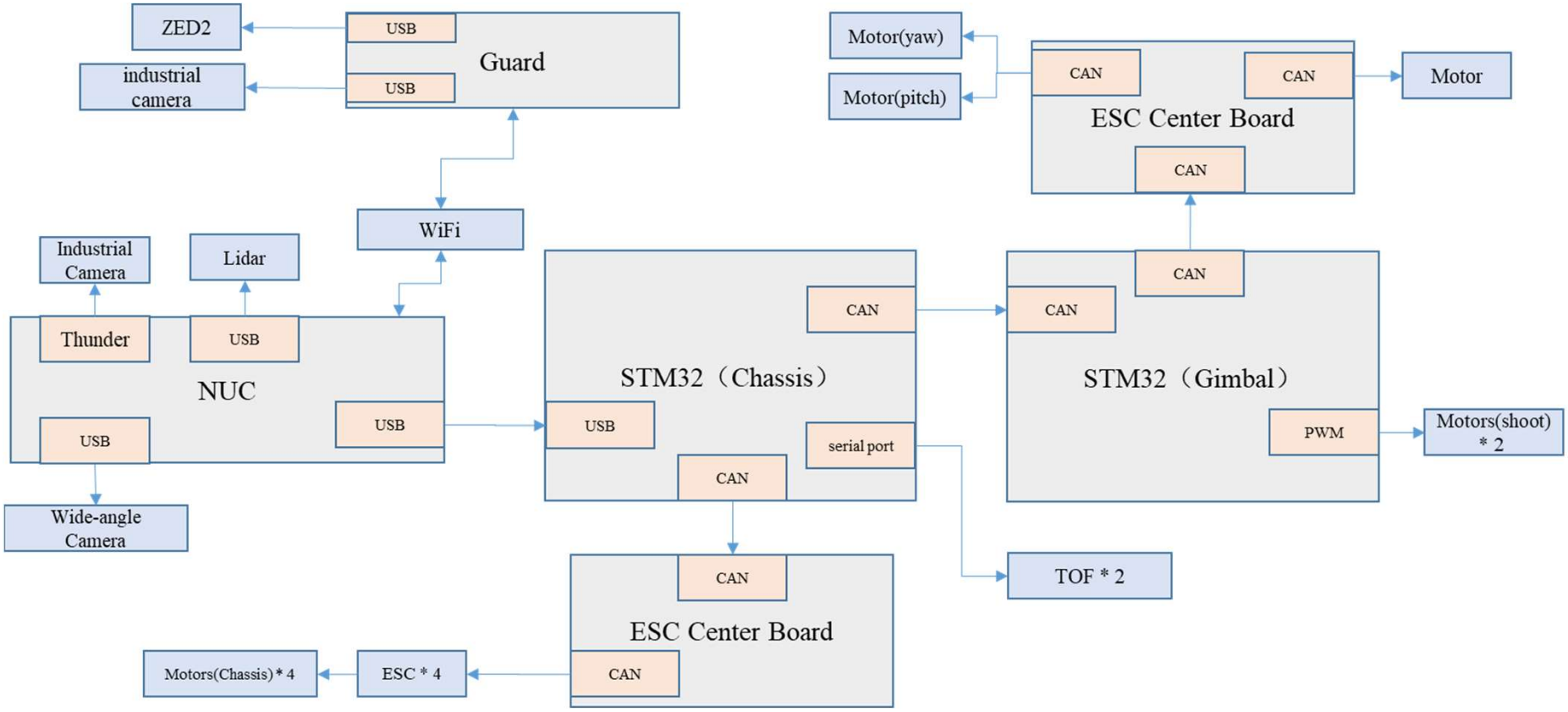
Robot engineer



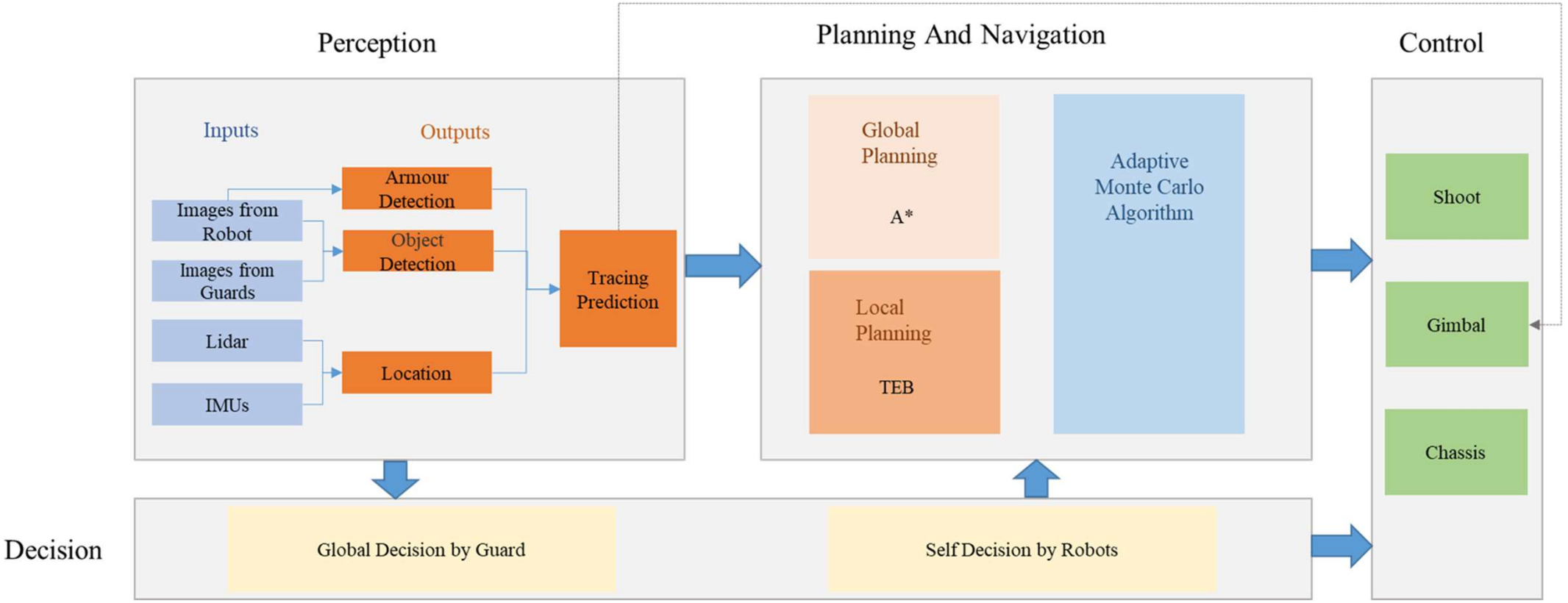

STM32F4
Industrial Camera
Lidar
Wide-angle Camera
STM32F4
Computer
(under lidar)
Armour

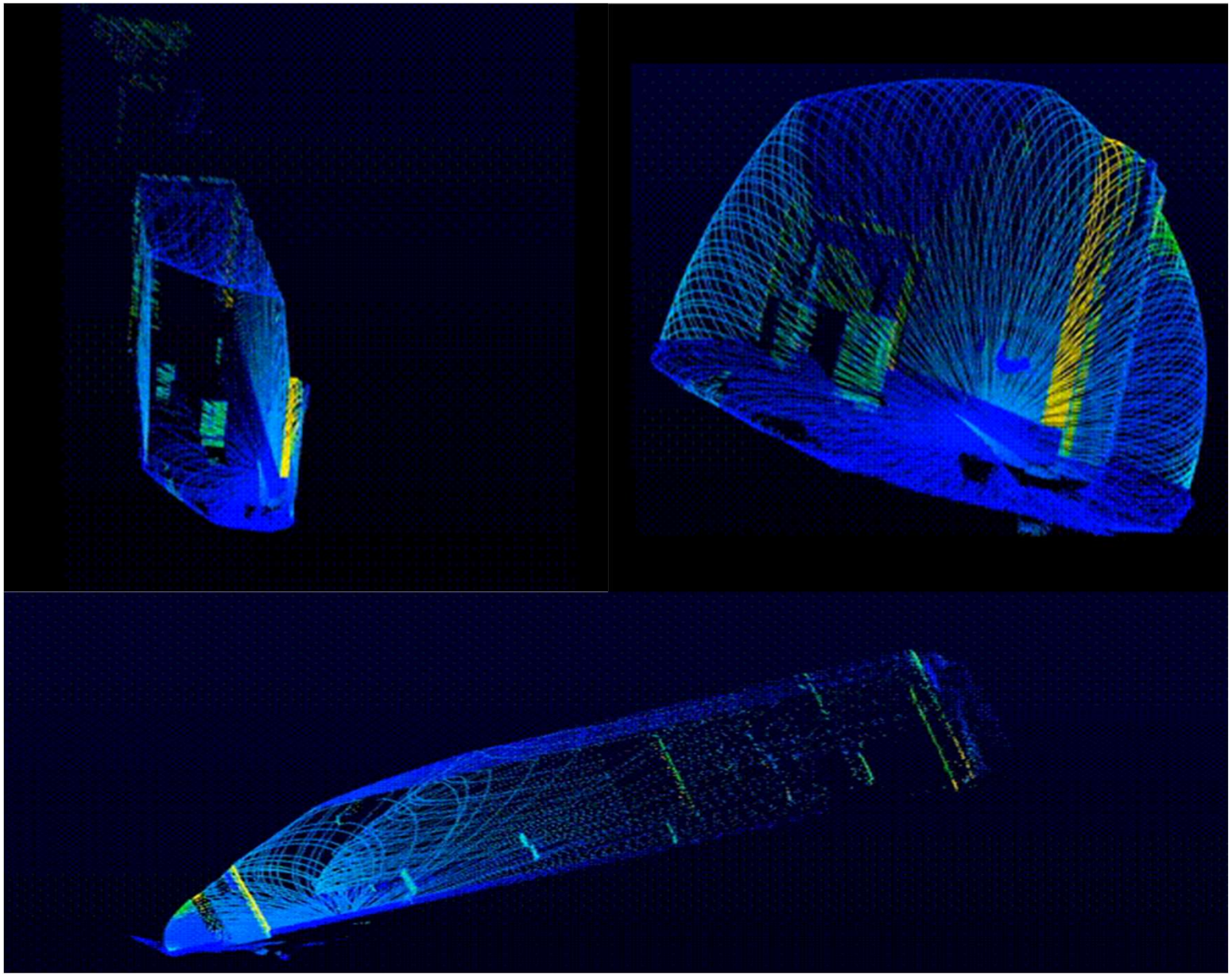

Lidar
Cam
Gimbal
TOF
Cam
TOF
Robot
Lidar View
Camera Views
TOF Views

Hardware Link

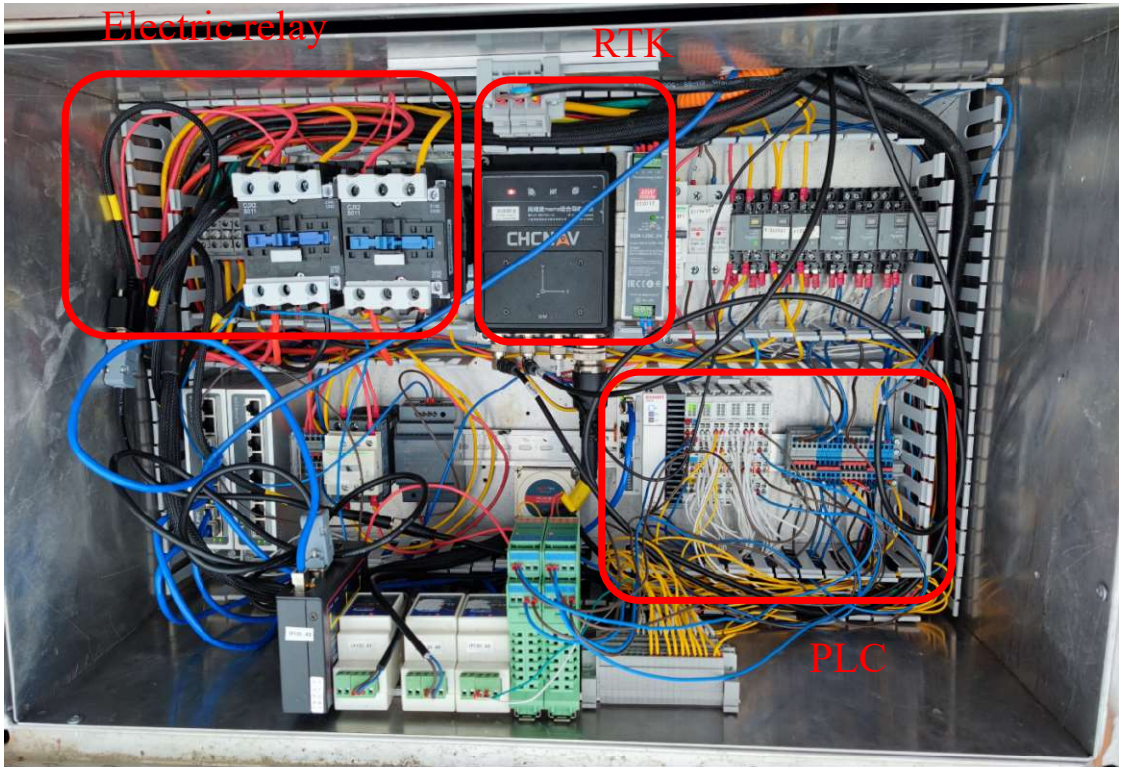


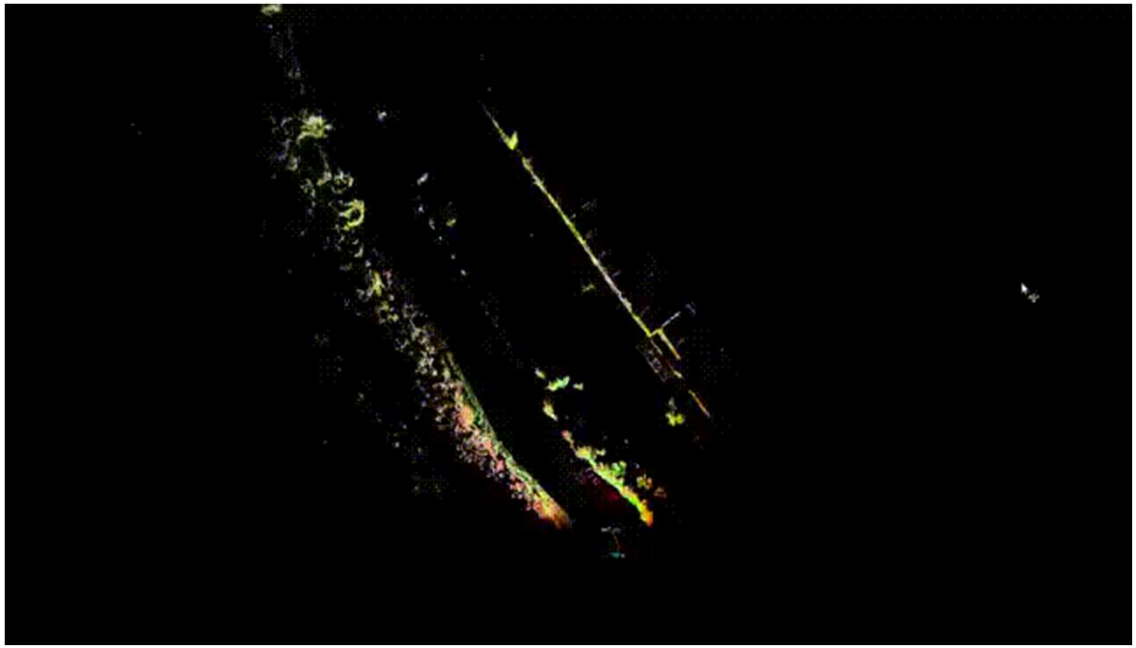
Software Architecture






科研用船
请勿靠近


Electric relay
RTK
PLC



452,000 points/s

NUC11PAHi7 (Intel i7 11370H)

模式分析与机器智能
工业和信息化部重点实验室
MIIT Key Laboratory of
Pattern Analysis & Machine Intelligence

ParNeC 模式识别与神经计算研究组
PAttern Recognition and NEural Computing

Efficient Teacher: Semi-Supervised Object Detection for YOLOv5

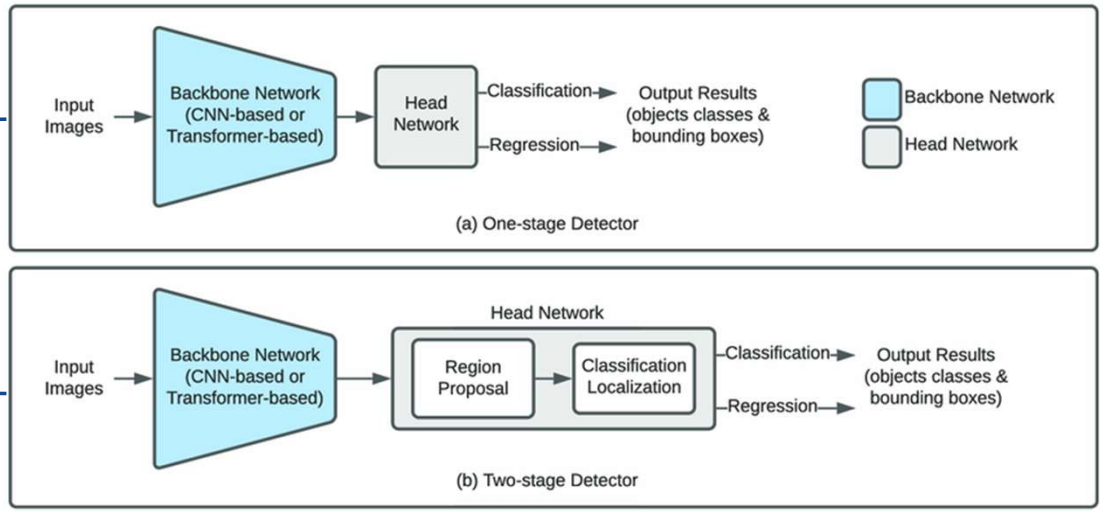
Bowen Xu Mingtao Chen Wenlong Guan Lulu Hu
Alibaba Group
{bowen.xbw,ruiyang.cmt,wenlong.gwl,chudu.hll}@alibaba-inc.com

Background

Object Detection

SSOD

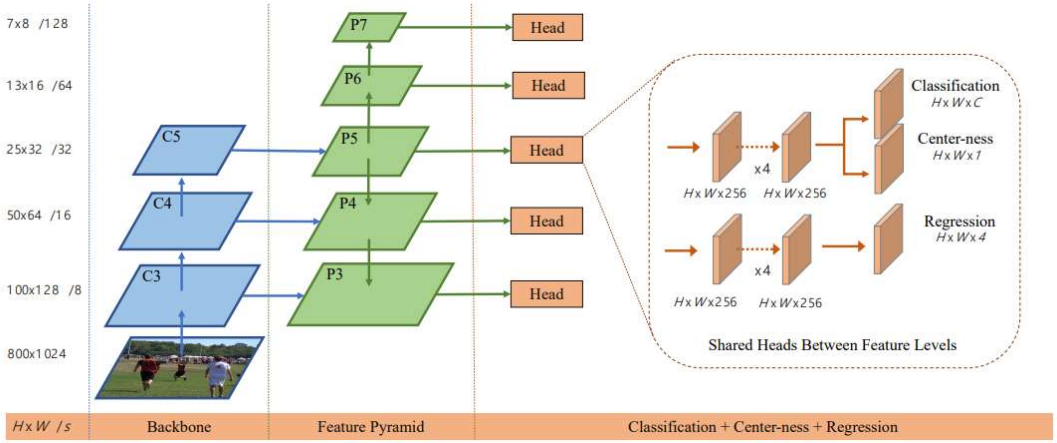
① **anchor-based**



✓

② **anchor-free**

(FCOS)



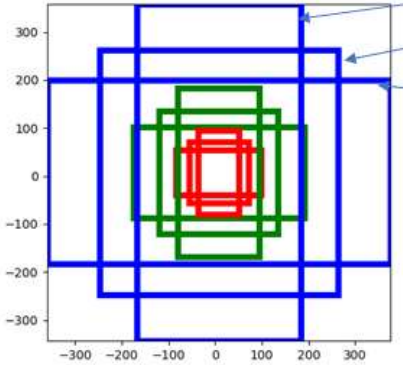
✓

Background

Generate Anchors

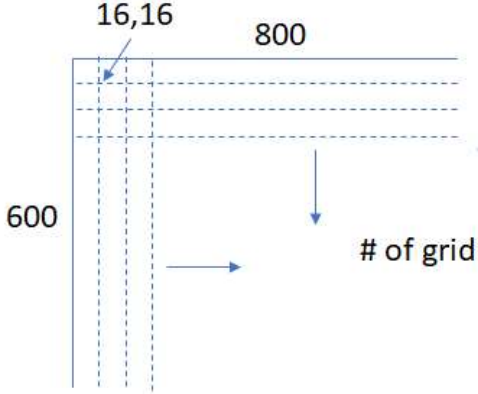
Given:

- Set of aspect ratios (0.5, 1, 2)
- Stride length (downscaling performed by resnet head: 16)
- Anchor Scales (8, 16, 32)



# of anchors: 9

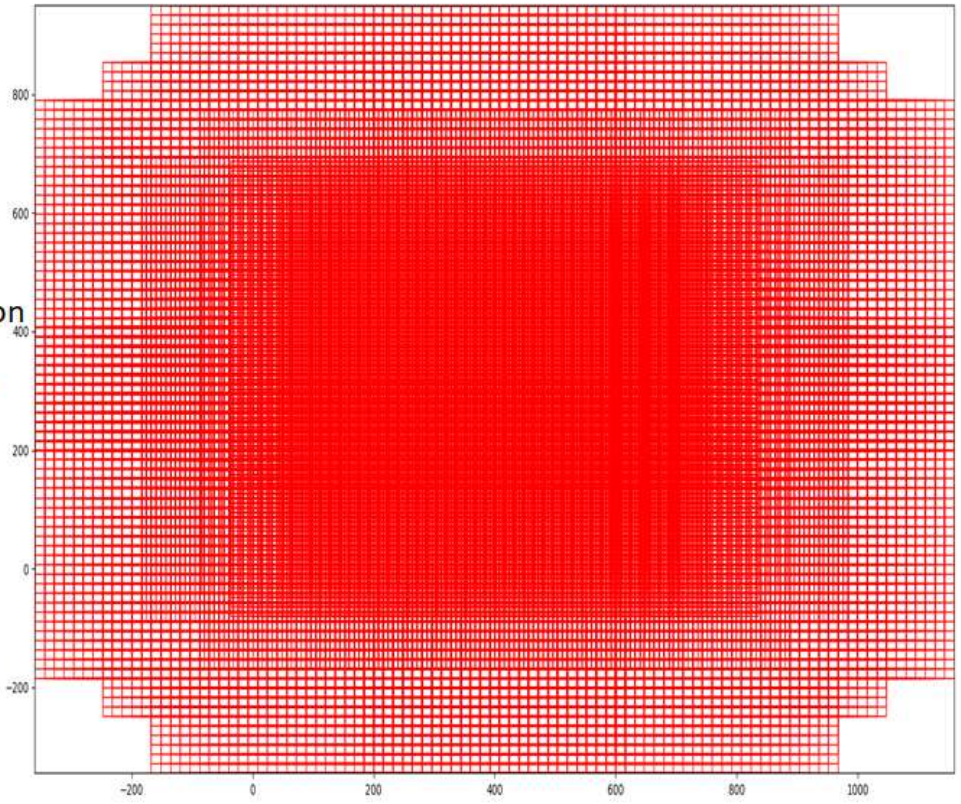
Add anchors to every grid location



# of grid locations: $\frac{800}{16} * \frac{600}{16} = 1900$

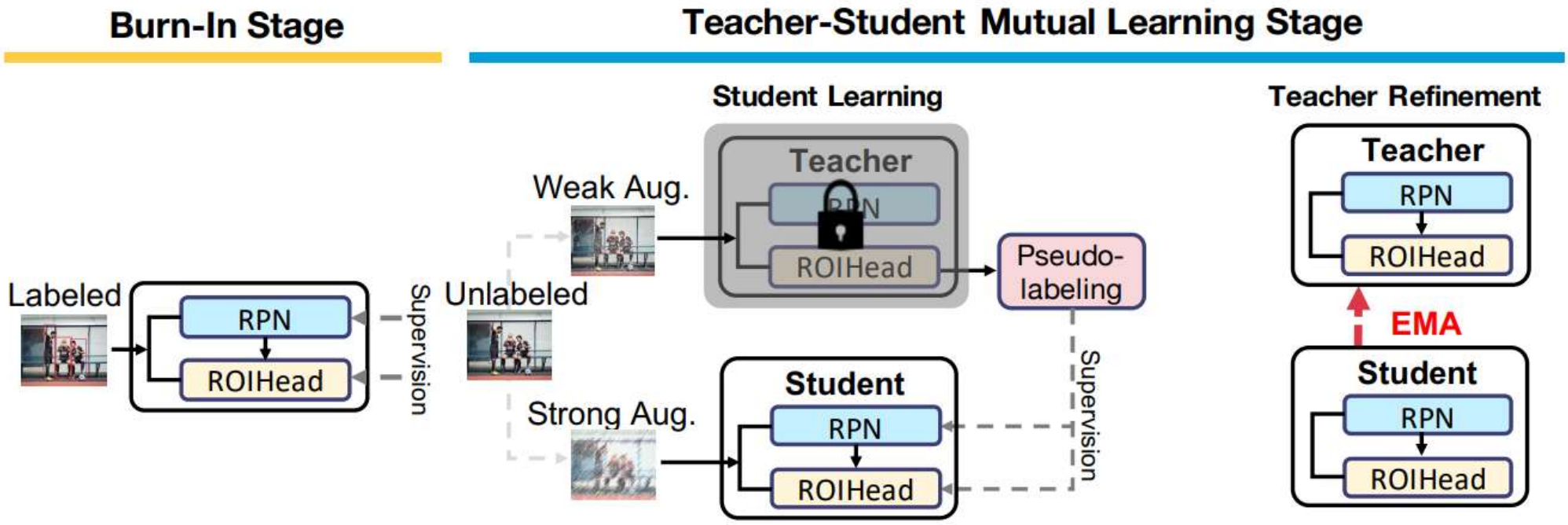
Create uniformly spaced grid with spacing = stride length

Total number of anchors: 1900*9 = 17100
Some boxes lie outside the image boundary



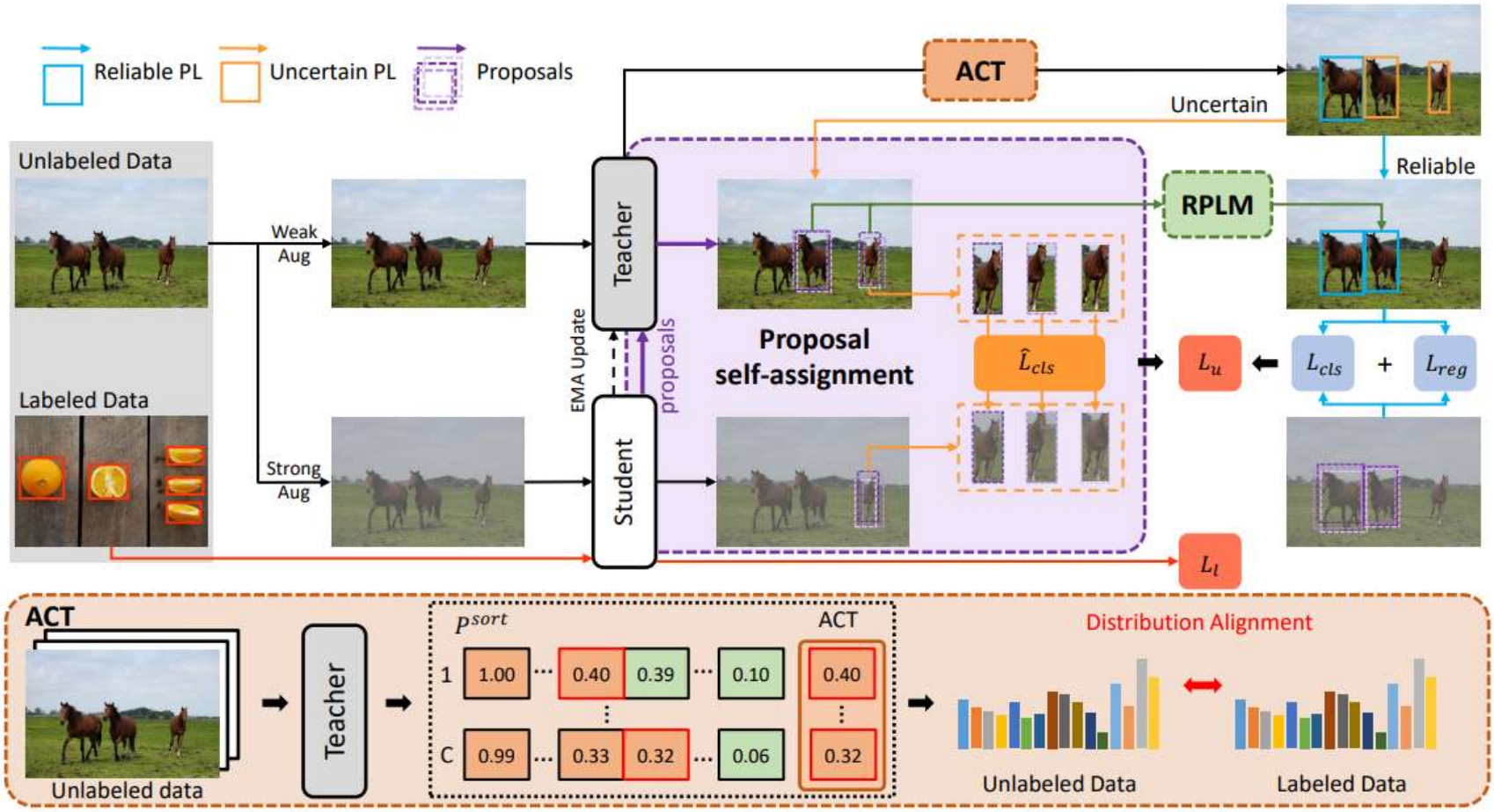
Related works

Unbiased Teacher (ICLR 2021)



Related works

LabelMatch (CVPR 2022)



Motivation

Challenge

- For the one-stage models, there is positive and negative samples imbalance during supervised training and poor quality of pseudo labels during semi-supervised training.

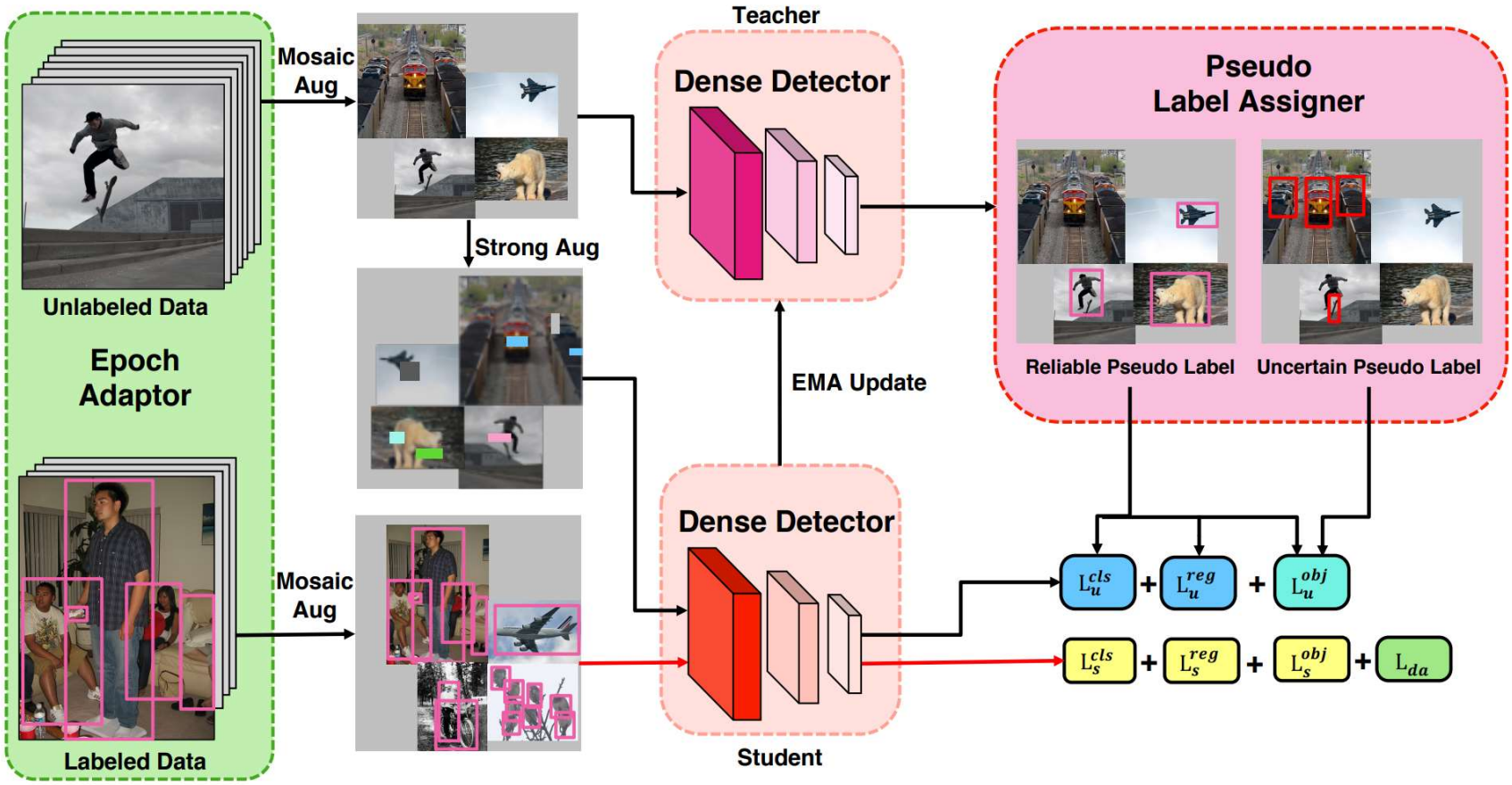
multi-stage coarse-to-fine prediction mechanism

anchor-free detection head

- Current mainstream SSOD approaches following a teacher-student mutual learning manner is difficult for an one-stage anchor-based detector to train due to the serious pseudo label inconsistency problem.

- SSOD model with both higher accuracy and better efficiency.

Method

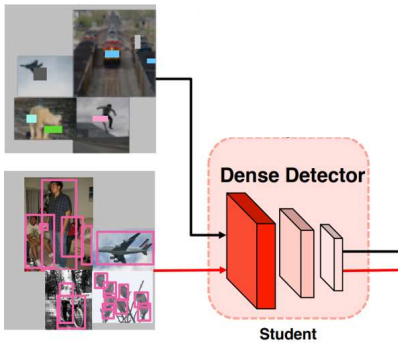

Unlabeled Data
Epoch Adaptor
Labeled Data
Mosaic Aug
Strong Aug
Mosaic Aug
Teacher
Dense Detector
EMA Update
Dense Detector
Student
Pseudo Label Assigner
Reliable Pseudo Label
Uncertain Pseudo Label
L_u^{cls} + L_u^{reg} + L_u^{obj}
L_s^{cls} + L_s^{reg} + L_s^{obj} + L_{da}

Method

Dense Detector

Method	Resolution	Mosaic	Param.	FLOPs	$AP_{50:95}(\%)$
Faster R-CNN [24]	[1333,800]		39.8M	202.31G	40.3
FCOS [31]	[1333,800]		32.02M	200.59G	38.5
YOLOv5 w/o	[640,640]		46.56M	109.59G	41.2
YOLOv5 [14]	[640,640]	✓	46.56M	109.59G	49.0
YOLOv7 [33]	[640,640]	✓	37.62M	106.59G	51.5
RetinaNet [19]	[1333,800]		37.74M	239.32G	39.5
Dense Detector	[640,640]	✓	42.13M	169.61G	44.86

Table 1. Comparison with Faster R-CNN, FCOS, YOLOv5, YOLOv7, RetinaNet and Dense Detector. The top section shows results for object detectors without Mosaic augmentation, the middle section shows results with Mosaic augmentation during training. Dense Detector achieves comparable results to RetinaNet baseline, having lower FLOPs but greatly improved $AP_{50:95}$. Both Faster R-CNN, FCOS, RetinaNet and Dense Detector uses ResNet-50-FPN as backbone. $AP_{50:95}$ is reported on COCO val dataset.



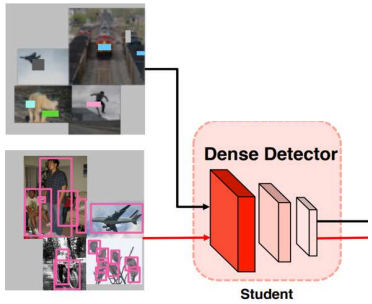
+ Mosaic

+ Dense flow of information

+ density of inputs

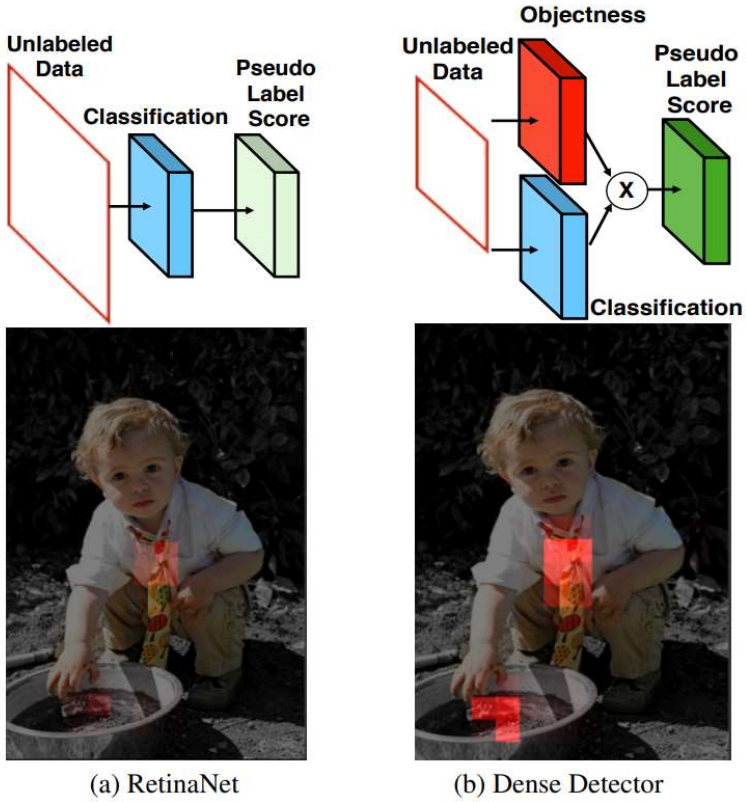
Method

Dense Detector



Base: RetinaNet with ResNet-50-FPN backbone

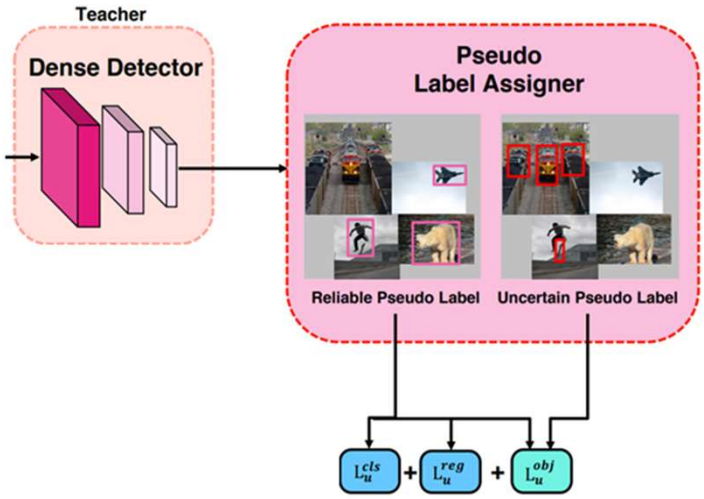
- Change FPN output from 5 to 3
- Eliminating the weight sharing between detection headers
- Reducing the input resolution from 1333 to 640
- Calculating the CIoU between the predicted and GT boxes to obtains objectness score(location quality of the predicted boxes)

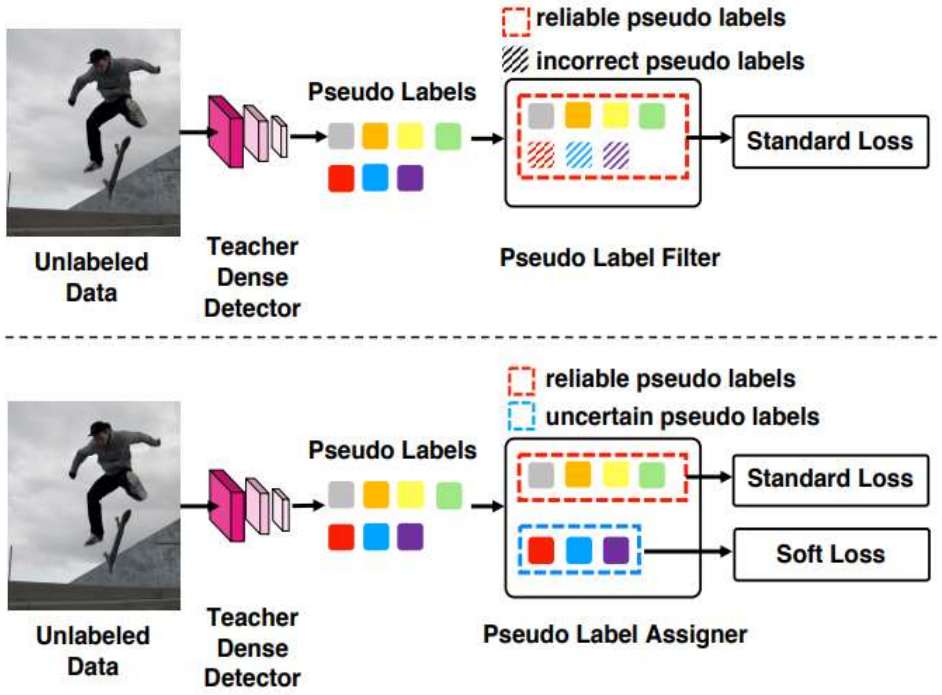


(a) RetinaNet (b) Dense Detector

Method

Pseudo Label Assigner



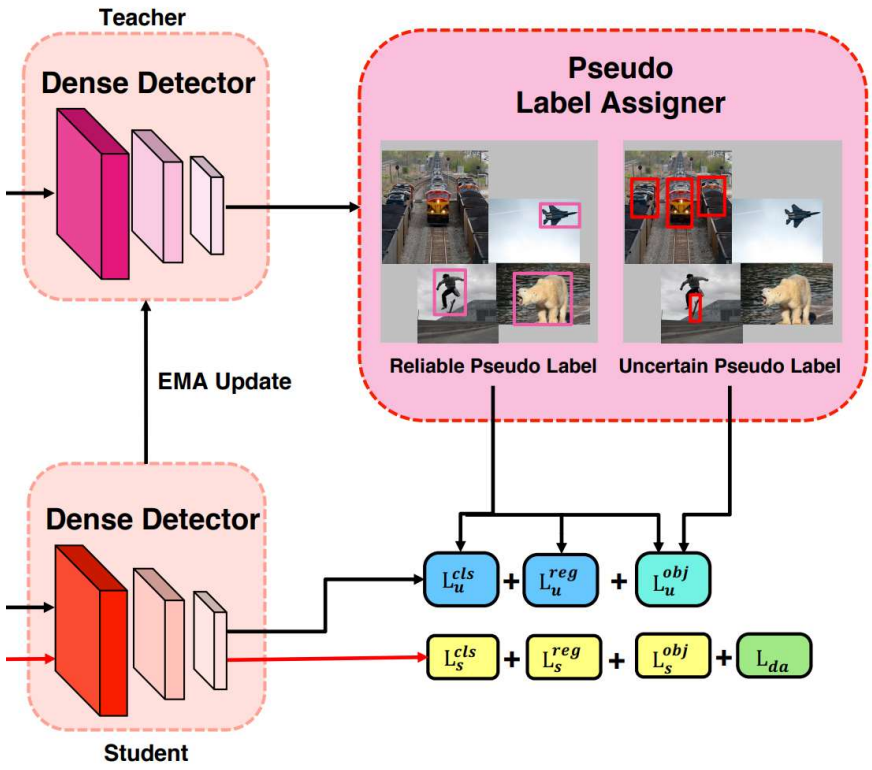


Sub-optimal assignments can lead to **inconsistent pseudo labels** and deteriorating performance of the **mutual learning mechanism**.

τ_1^k < uncertain < τ_2^k reliable

Method

Pseudo Label Assigner



YOLOv5

$$L_s = \sum_{h,w}(CE(X^{cls}_{(h,w)}, Y^{cls}_{(h,w)}) + CIoU(X^{reg}_{(h,w)}, Y^{reg}_{(h,w)}) + CE(X^{obj}_{(h,w)}, Y^{obj}_{h,w}))$$

balance

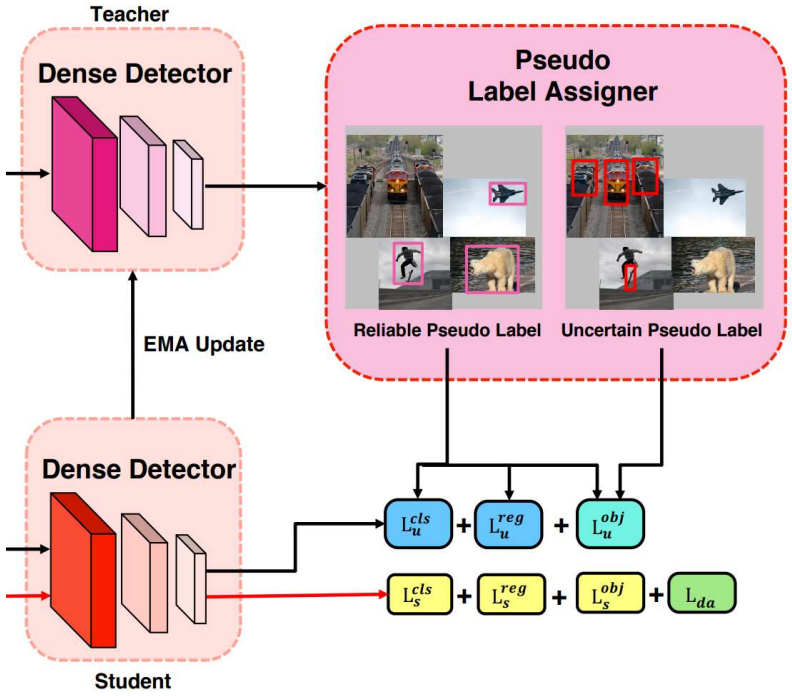
$$L = L_s + \lambda L_u$$

$$L_u = L_u^{cls} + L_u^{reg} + L_u^{obj}$$

improve the quality of pseudo labels

Method

Pseudo Label Assigner



$$L_u = L_u^{cls} + L_u^{reg} + L_u^{obj}$$

classification score

$$L_u^{cls} = \sum_{h,w} (\mathbb{1}_{\{p_{(h,w)} >= \tau_2\}} CE(X_{(h,w)}^{cls}, \hat{Y}_{(h,w)}^{cls}))$$

regression score

$$L_u^{reg} = \sum_{h,w} (\mathbb{1}_{\{p_{(h,w)} >= \tau_2 \text{ or } \hat{obj}_{(h,w)} > 0.99\}} CIoU(X_{(h,w)}^{reg}, \hat{Y}_{(h,w)}^{reg}))$$

objectness score

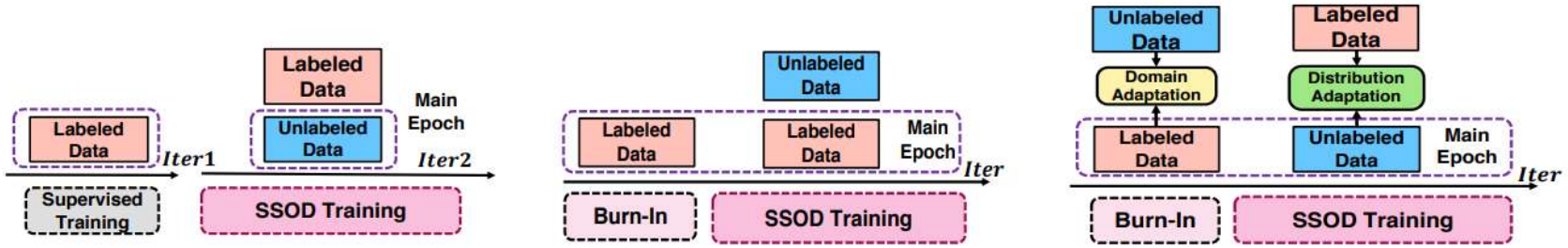
$$L_u^{obj} = \sum_{h,w} (\mathbb{1}_{\{p_{(h,w)} <= \tau_1\}} CE(X_{(h,w)}^{obj}, \mathbf{0}) + \mathbb{1}_{\{p_{(h,w)} >= \tau_2\}} CE(X_{(h,w)}^{obj}, \hat{Y}_{(h,w)}^{obj})) + \mathbb{1}_{\{\tau_1 < p_{(h,w)} < \tau_2\}} CE(X_{(h,w)}^{obj}, \hat{obj}_{(h,w)}))$$

objectness score of pseudo label (soft label)

Since they are not involved in classification and regression, these pseudo-labels are not categorized into positive and negative samples.

Method

Epoch Adaptor



(a) Alternating Training (b) Joint Training with Burn-In (c) Joint Training with Epoch Adaptor

Figure 4. Main epoch denotes a full training period that remains uninterrupted and without any reloading of new weights during its execution. Training strategies for Efficient Teacher: (a) supervised training on labeled data followed by SSOD training on unlabeled data; (b) supervised training on labeled data with additional SSOD training on unlabeled data; (c) end-to-end training on both labeled and unlabeled data with Epoch Adaptor incorporating Domain and Distribution Adaptation for improved convergence and feature distribution.

- **Domain adaptation**: closing the gap between the distribution of labeled and unlabeled data
- **Distribution adaptation**: dynamically estimating the threshold of pseudo-labeling for each epoch

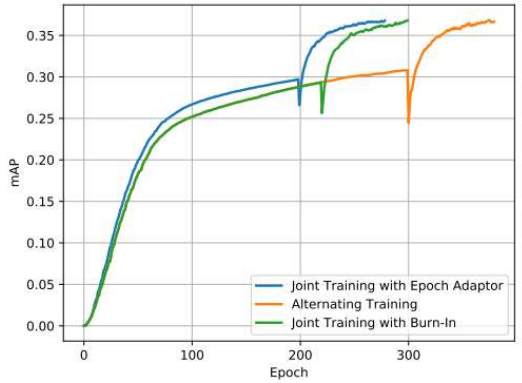


Figure 5. Performance ($AP_{50:95}$) comparisons of Epoch Adaptor, Alternating Training and Joint Training with Burn-In methods on COCO standard 10%.

Method

Domain adaptation

mitigates the overfitting effect

output of the domain classifier

$$L_s = \sum_{h,w}(CE(X^{cls}_{(h,w)}, Y^{cls}_{(h,w)}) + CIoU(X^{reg}_{(h,w)}, Y^{reg}_{(h,w)}) + CE(X^{obj}_{(h,w)} + Y^{obj}_{h,w})) + \lambda L_{da}$$

$$L_{da} = -\sum_{h,w}\left[D \log p_{(h,w)} + (1-D)\log(1-p_{(h,w)})\right]$$

unlabeled labeled

Distribution adaptation

τ_1^k < uncertain < τ_2^k < reliable

number of unlabeled

$$\tau_1^k = P_c^k[n_c^k \cdot \frac{N_u}{N_l}]$$

$$\tau_2^k = P_c^k[\alpha\% \cdot n_c^k \cdot \frac{N_u}{N_l}],$$

reliable ratio

The use of Mosaic data augmentation disrupts the label distribution ratio.

Proposal Self-Assignment. It is worth noting that the quality of pseudo labels cannot be guaranteed, especially at the early beginning of self-training. Inspired by the noise label learning [5, 14], we divide pseudo labels into reliable ones and uncertain ones according to the confidence score. Denoting $\alpha\%$ as the pre-defined proportion of reliable pseudo labels, the confidence thresholds t_c^r to filter reliable pseudo labels for the c-th category can be written as:

$$t_c^r = P_c^{sort}[\alpha\% \cdot n_c^l \cdot \frac{N_u}{N_l}], \quad (6)$$

Pseudo labels with confidence higher than t_c^r are regarded as hard labels for student model optimization in a supervised manner. In contrast, the remaining uncertain ones are treated as soft labels for soft learning.

LabelMatch

Experiment

COCO

Method		%1	%2	%5	%10	FLOPs
Two-stage anchor-based	Supervised	9.05	12.70	18.47	23.86	202.31G
	STAC [27]	13.97 ± 0.35(+4.92)	18.25 ± 0.25 (+5.91)	24.38 ± 0.12 (+5.91)	28.64 ± 0.21 (+4.78)	202.31G
	Instant Teaching [40]	18.05 ± 0.15 (+9.00)	22.45 ± 0.15 (+9.75)	26.75 ± 0.05 (+8.28)	30.40 ± 0.05 (+6.54)	202.31G
	Humber teacher [29]	16.96 ± 0.38 (+7.91)	21.72 ± 0.24 (+9.02)	27.70 ± 0.15 (+9.23)	31.61 ± 0.28 (+7.75)	202.31G
	Unbiased Teacher [21]	20.75 ± 0.12 (+11.70)	24.30 ± 0.07 (+9.80)	28.27 ± 0.11 (+9.80)	31.50 ± 0.10 (+7.64)	204.13G
	Soft Teacher [35]	20.46 ± 0.39 (+11.41)	-	30.74 ± 0.08 (+12.27)	34.04 ± 0.14 (+10.18)	202.31G
	LabelMatch [4]	**25.81** ± 0.28 (+16.76)	-	32.70 ± 0.18 (+14.23)	35.49 ± 0.17 (+11.63)	202.31G
	PseCo [17]	22.43 ± 0.36 (+13.38)	27.77 ± 0.18 (+15.07)	32.50 ± 0.08 (+14.03)	36.06 ± 0.24 (+12.20)	202.31G
One-stage anchor-free	Supervised	9.53	11.71	18.74	23.70	200.59G
	Unbiased Teacher v2 [22]	22.71 ± 0.42 (+13.18)	26.03 ± 0.12 (+14.32)	30.08 ± 0.04 (+11.34)	32.61 ± 0.03 (+8.91)	200.59G
	DSL [5]	22.03 ± 0.28 (+12.50)	25.19 ± 0.37 (+13.48)	30.87 ± 0.24 (+12.13)	36.22 ± 0.18 (+12.52)	200.59G
	Dense Teacher [39]	22.38 ± 0.31 (+12.85)	27.20 ± 0.20 (+15.49)	33.01 ± 0.21 (+14.27)	37.13 ± 0.12 (+13.43)	200.59G
One-stage anchor-based	Supervised	10.29	13.12	19.28	24.04	169.61G
	Unbiased Teacher∗ [21]	18.81 ± 0.28 (+9.07)	22.72 ± 0.21 (+9.60)	28.35 ± 0.12 (+8.15)	30.34 ± 0.09 (+6.30)	169.61G
	Ours	21.51 ± 0.21 (+11.22)	27.15 ± 0.13 (+14.03)	31.1 ± 0.08 (+11.82)	34.09 ± 0.11 (+10.05)	169.61G
	Ours †	23.76 ± 0.13 (+12.47)	**28.70** ± 0.14 (+15.58)	**34.11** ± 0.09 (+14.83)	**37.90** ± 0.04 (+13.86)	109.59G

Table 2. Experimental results on COCO-standard ($AP_{50:95}$), $*$ means re-implemented results on Dense Detector, † means Efficient Teacher with YOLOv5l [14]. All the results are the average of 5 folds.

Experiment

COCO

Method	$AP_{50:95}$	AP_{50}	FLOPs
STAC [27]	44.64	77.45	202.31G
Instant Teacher [40]	50.00	79.20	202.31G
Unbiased Teacher [21]	48.69	77.37	204.13G
Dense Teacher [39]	55.87	79.89	200.59G
DSL [5]	56.80	80.70	200.59G
Unbiased Teacher v2 [22]	56.87	81.29	200.59G
LabelMatch [4]	55.11	85.48	202.31G
Ours †	58.30	81.60	109.59G
Ours ‡	**60.56**	**86.54**	109.59G

Table 4. Experimental results on PASCAL-VOC. The ‡ indicates using a ImageNet pre-trained backbone to initialize the Efficient Teacher

Method	$AP_{50:95}$	AP_{50}
Supervised	30.45	44.65
Unbiased Teacher [21]	32.10 (+1.65)	47.30 (+2.65)
Ignore uncertain pseudo label [5]	35.20 (+4.75)	52.00 (+7.35)
Pseudo Label Assigner	**37.90 (+7.45)**	**54.19 (+9.54)**

Table 5. Ablation study about different pseudo label assignment methods.

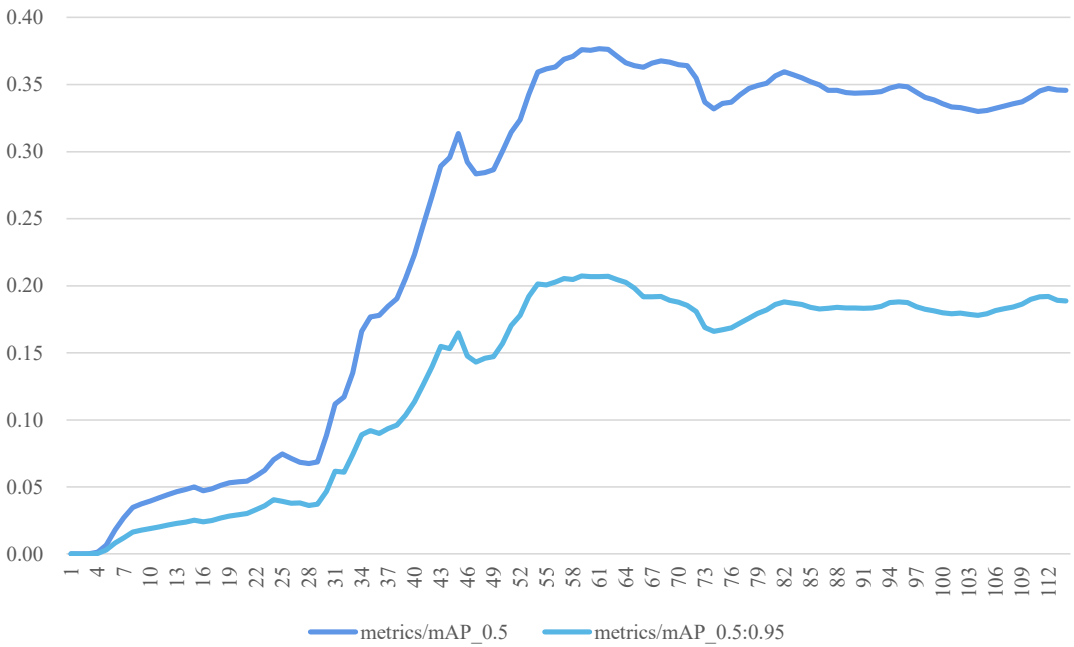
Method	$AP_{50:95}$	AP_{50}
w/o domain adaptation	37.25	54.16
domain adaptation	**37.90**	**54.80**

Table 7. Ablation studies on domain adaptation in EA.

τ_2	$AP_{50:95}$	AP_{50}
0.4	37.20	54.08
0.5	37.20	54.10
0.6	36.90	53.77
0.7	35.10	51.60
EA	**37.90**	**54.80**

Table 6. Ablation studies on threshold value τ_2 , EA indicates τ_2 is calculated by Epoch Adaptor.

Experiment



2*4090: need about 10 days to train

Implementation Details. We use 8 NVIDIA-V100 GPUs with 16G memory per GPU. We randomly sample 32 images from labeled data and 32 images from unlabeled data with ratio 1:1 in each iteration. For training configurations, the learning rate is 0.01 all the time, the τ_1 and τ_2 are calculated by EA. We used both weak and strong data augmentation. Mosaic is used in weak data augmentation. In the strong data augmentation, Mosaic, left-right flip, large scale jittering, graying, Gaussian blur, cutout, and color space conversion are selected. The max epoch is 300. Smoothing hyper-parameter in EMA is 0.999.

Summary

1. The YOLO-style models are more flexible compared to other object detection models and is suitable for deployment on edge computing devices such as Nvidia Jetson;

2. Dense Detector obtains higher quality pseudo-tags in a simple and efficient way;

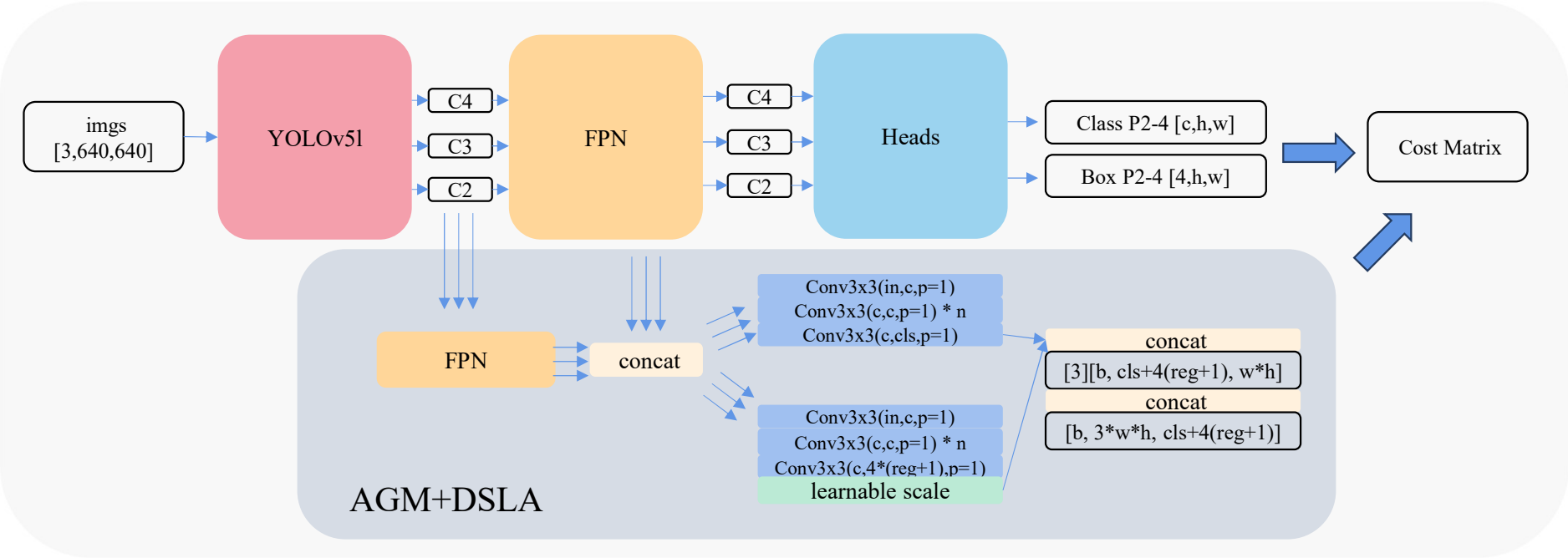
3. The Efficient Teacher is able to strike a good balance between performance and computational efficiency;

TODO:

1. Explore why Domain Adaptation enables the model to converge earlier and whether better ways exist to accelerate model training;
2. How to obtain higher quality pseudo-labels and whether there exists a more efficient way than objectness branch;
3. Better model performance;

TODO

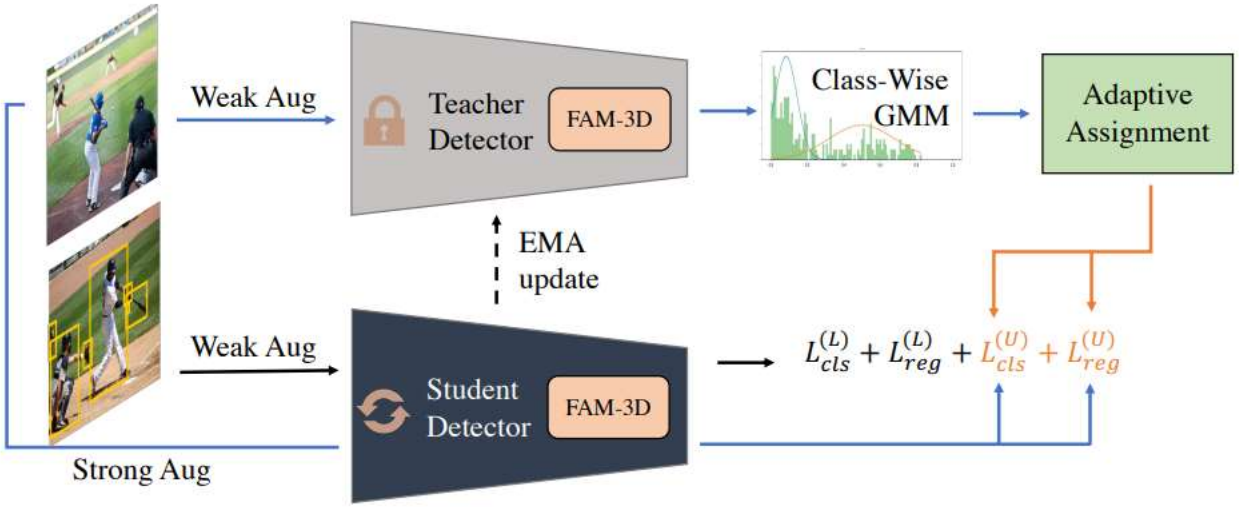
Better ways to accelerate model training

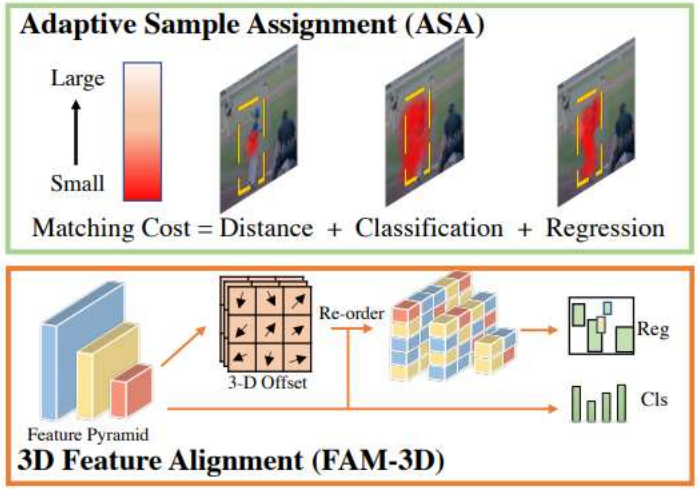


+ Assign Guidance Module
+ Dynamic Soft Label Assigner

TODO

How to obtain higher quality pseudo-labels

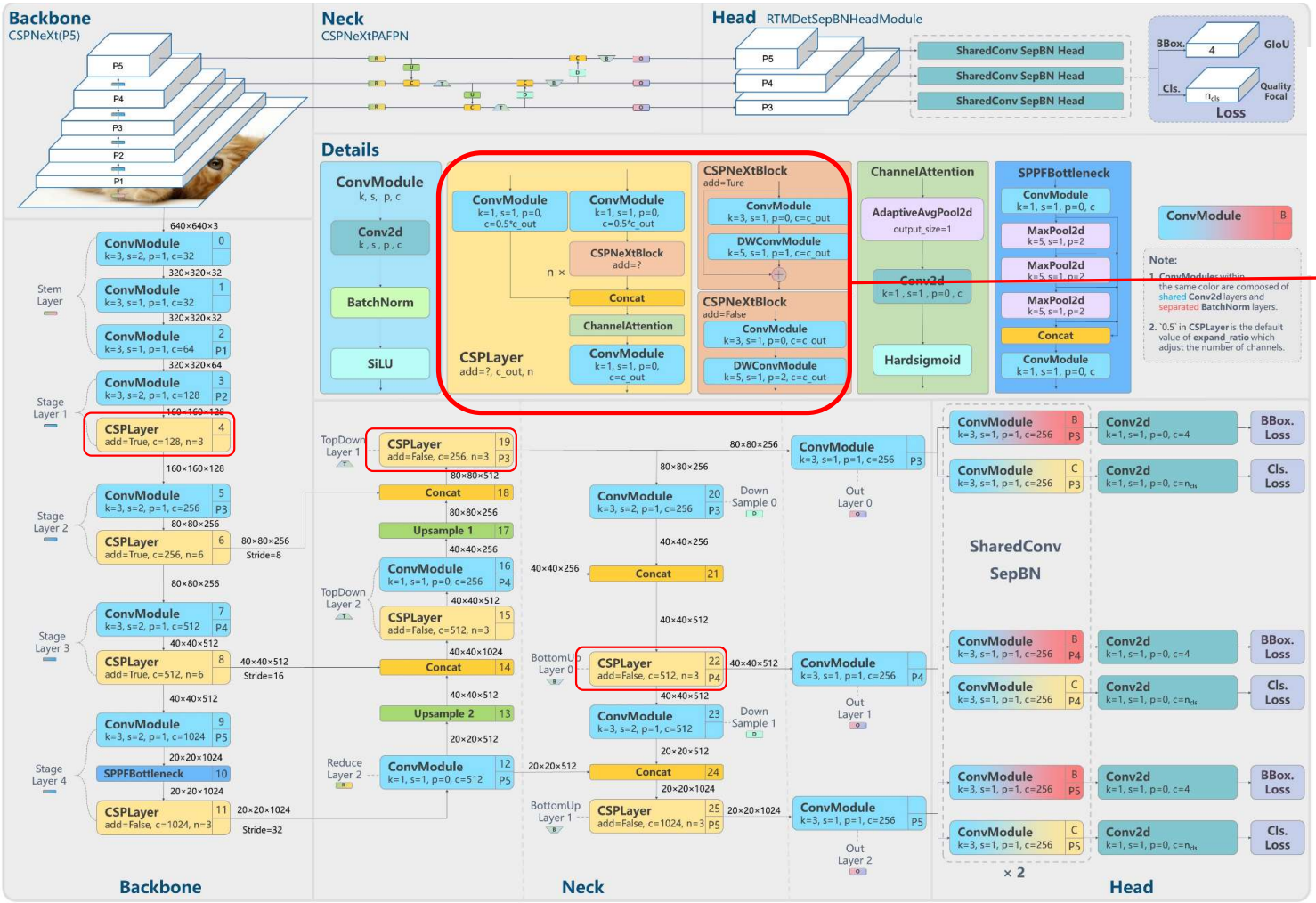


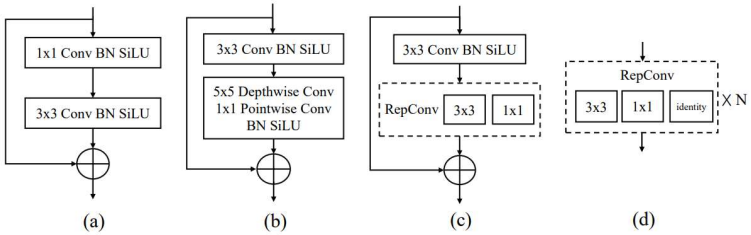


Wang X, Yang X, Zhang S, et al. Consistent-Teacher: Towards Reducing Inconsistent Pseudo-Targets in Semi-Supervised Object Detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 3240-3249.

TODO

Better model performance

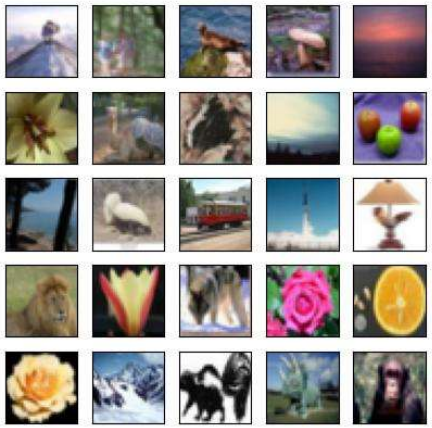




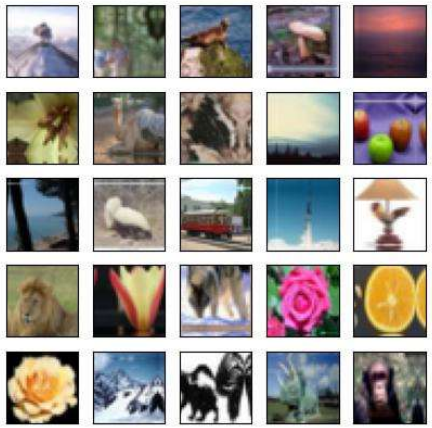
Lyu C, Zhang W, Huang H, et al. Rtmdet: An empirical study of designing real-time object detectors[J]. arXiv preprint arXiv:2212.07784, 2022.

Anything else

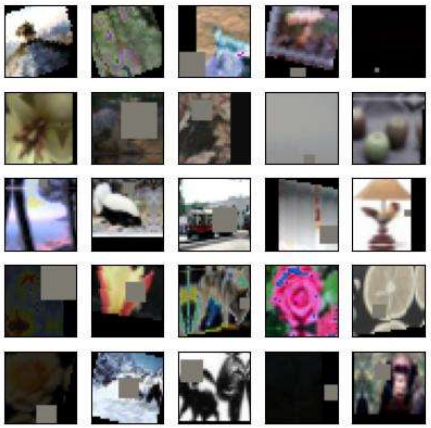
FlexMatch (Zhang et al., 2021): CIFAR100, 2500



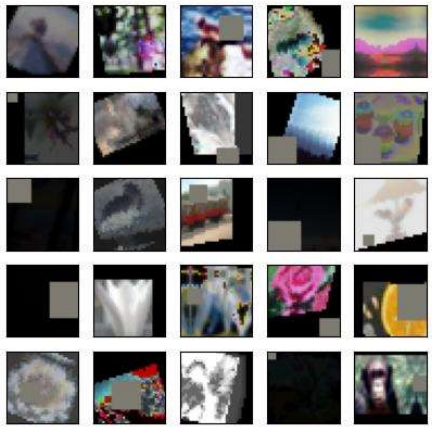
Original image



Weak Augmentation
(65.13%)



Strong Augmentation
(71.56%, baseline)

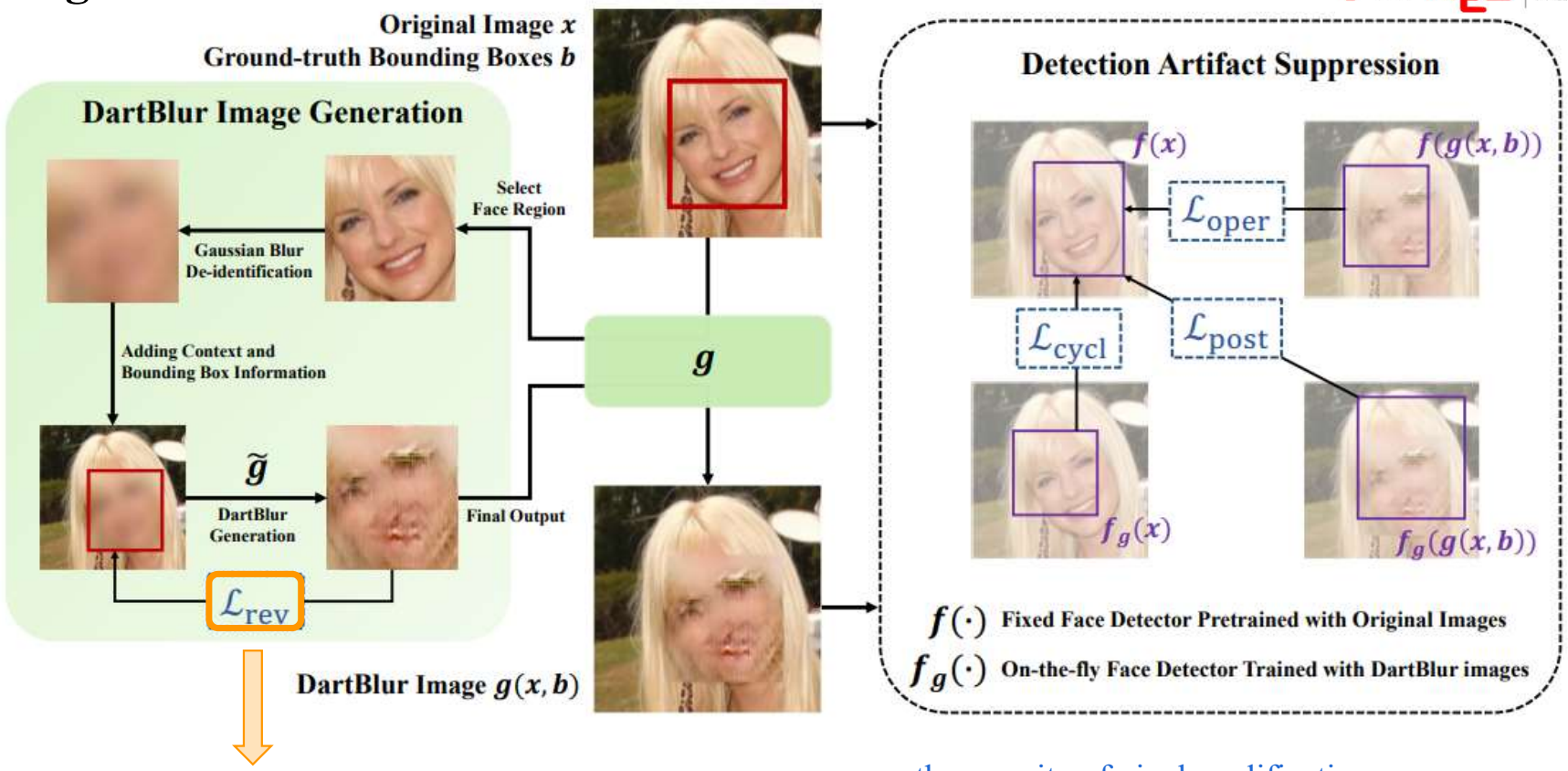


Strong Augmentation+
(72.26%)

Higher accuracy, but not stable

FlexMatch ✓
FixMatch ✓
FreeMatch ×

Anything else



encourage the sparsity of pixel modification

$$\begin{aligned}\mathcal{L}_{\text{rev}} &= \mathcal{L}_{\text{rev}}\left(g, \boldsymbol{x}, \boldsymbol{b}, \epsilon_{\text{rev}}\right) \\ &= \max\left(\left\|g(\boldsymbol{x}, \boldsymbol{b}) - \mathcal{G}(\boldsymbol{x}, \boldsymbol{b})\right\|_1 - \epsilon_{\text{rev}}, 0\right)\end{aligned}$$

a budget for DartBlur to be different from Gaussian blur

DartBlur: Privacy Preservation With Detection Artifact Suppression(CVPR2023)