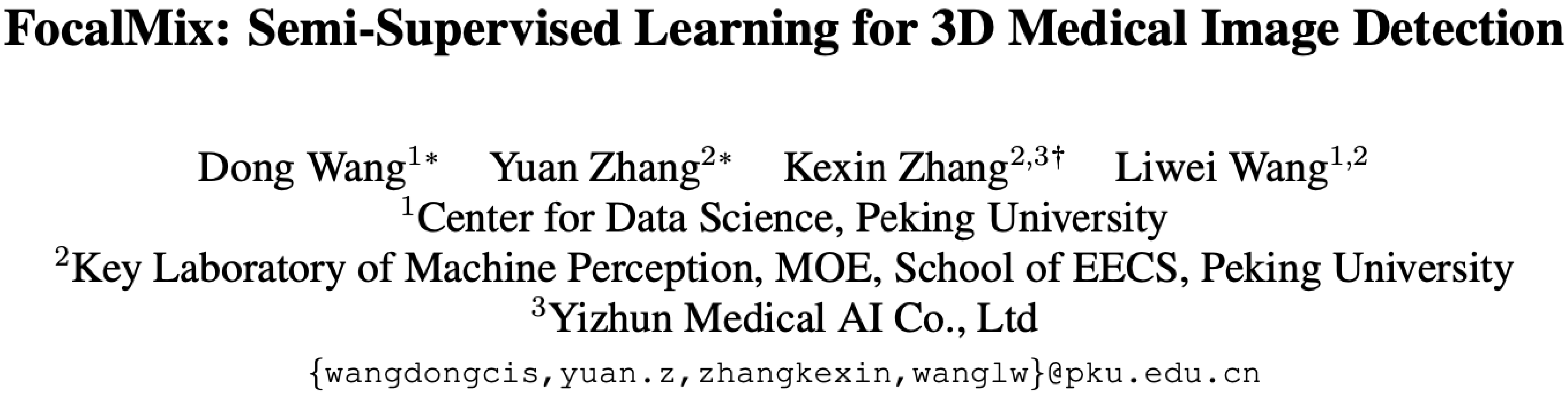
Paper List	
CVPR2020	FocalMix: Semi-Supervised Learning for 3D Medical Image Detection
ICCV2017	Focal Loss for Dense Object Detection

Information



FocalMix: Semi-Supervised Learning for 3D Medical Image Detection

Dong Wang[1*] Yuan Zhang[2*] Kexin Zhang[2,3†] Liwei Wang[1,2]
[1]Center for Data Science, Peking University
[2]Key Laboratory of Machine Perception, MOE, School of EECS, Peking University
[3]Yizhun Medical AI Co., Ltd
{wangdongcis,yuan.z,zhangkexin,wanglw}@pku.edu.cn

Introduction

- Medical Annotation is expensive and time-consuming, especially for 3D images.
- A large number of raw medical images remain unused, while the cost of annotation is high.
- Semi-Supervised Learning (SSL) is capable of utilizing unlabeled data.
- SSL is widely used in medical image processing and focuses more on classification tasks.

Difficulties

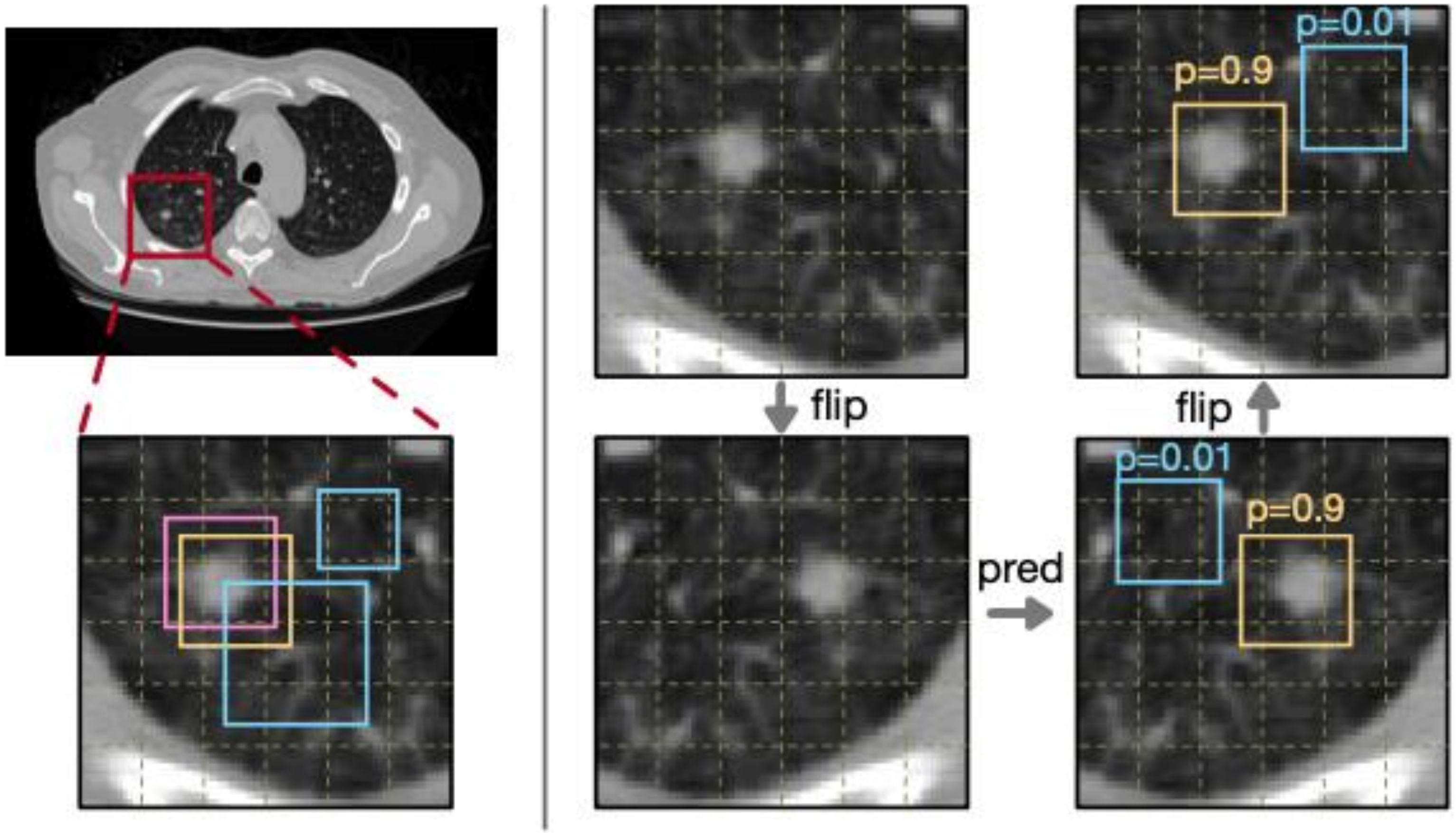
Existing SSL Literature	Unexplored detaks in medical
More focus in classification task	More concern lesion detection
Require the loss function to be able to deal with soft labels	Focal Loss in one-stage detection
Latest SSL methods use average to get pseudo-label	Bounding-box cannot take the average
Data augmentation is indispensable	Data augmentation few touched

Contributions

1. Propose FocalMix, a novel SSL framework for 3D medical image detection.
2. The first to investigate the problem of SSL for medical image detection.
3. Through extensive experiments, demonstrate that the proposed SSL approach can significantly improve the performance of fully-supervised learning approaches.

Preliminaries

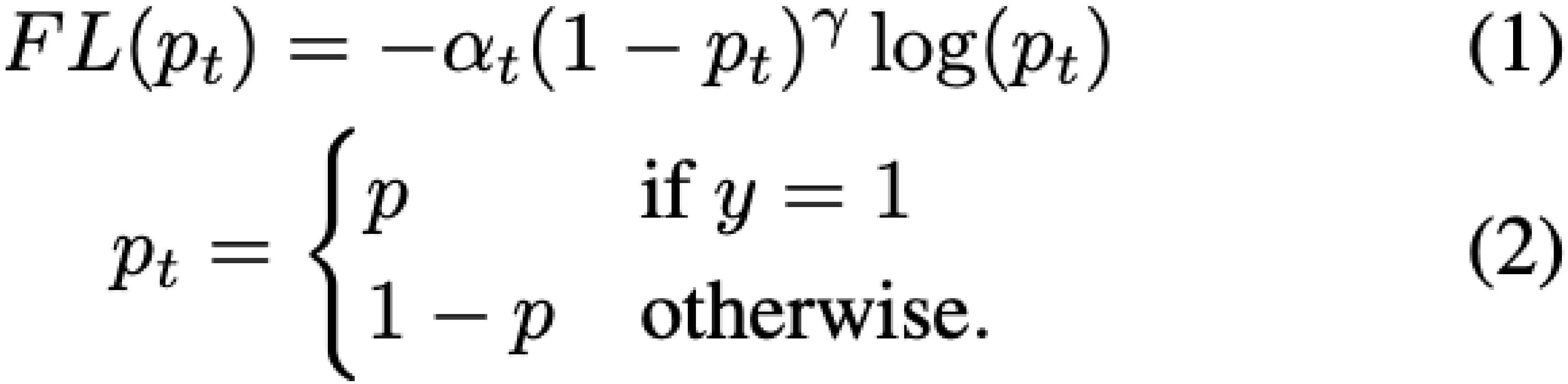
1. Anchor Box
2. Focal Loss
3. MixMatch



(a) (b)

Preliminaries

1. Anchor Box
2. Focal Loss
3. MixMatch



$$FL(p_t) = -\alpha_t (1 - p_t)^\gamma \log(p_t) \tag{1}$$

$$p_t = \begin{cases} p & \text{if } y = 1 \\ 1 - p & \text{otherwise.} \end{cases} \tag{2}$$

Preliminaries

1. Anchor Box
2. Focal Loss
3. MixMatch: State-of-the-art unified SSL framework that integrates the spirits of most successful attempts in this line of research (Left: Target prediction for unlabeled data; Right: MixUp Augmentation)

$$\bar{y} = \frac{1}{K} \sum_{k=1}^{K} \mathrm{p}_{\mathrm{Model}}(\hat{u}_k; \theta). \tag{3}$$

$$\mathrm{Sharpen}(\bar{y}, T)_i = \bar{y}_i^{\frac{1}{T}} \bigg/ \sum_{j=1}^{L} \bar{y}_j^{\frac{1}{T}}, \tag{4}$$

$$\lambda \sim \mathrm{Beta}(\eta, \eta), \tag{5}$$

$$\tilde{\lambda} = \max(\lambda, 1 - \lambda), \tag{6}$$

$$\hat{x} = \tilde{\lambda} x + (1 - \tilde{\lambda}) x', \tag{7}$$

$$\hat{y} = \tilde{\lambda} y + (1 - \tilde{\lambda}) y'. \tag{8}$$

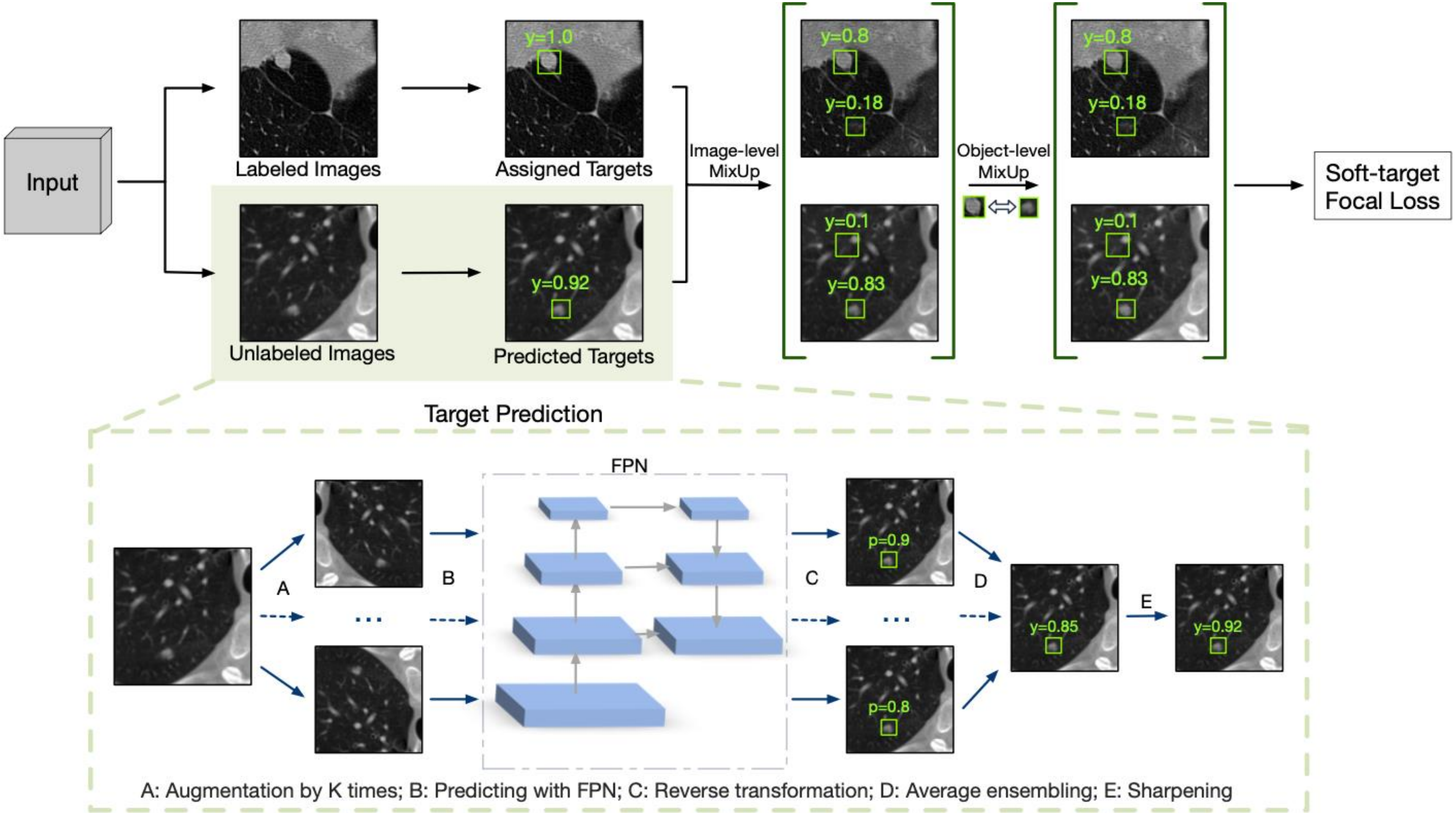


Figure 2: **Overview of our proposed method FocalMix.** For an input batch, the training targets of anchors in labeled images are assigned according to annotated boxes, while the unlabeled are predicted with the current model as shown in the lower part of the figure. After applying two levels of MixUp to the entire batch, we use the proposed soft-target focal loss to train the model. Throughout this paper, we only show a slice of each 3D CT scan with 3D anchors on it for ease of presentation.

Methodology

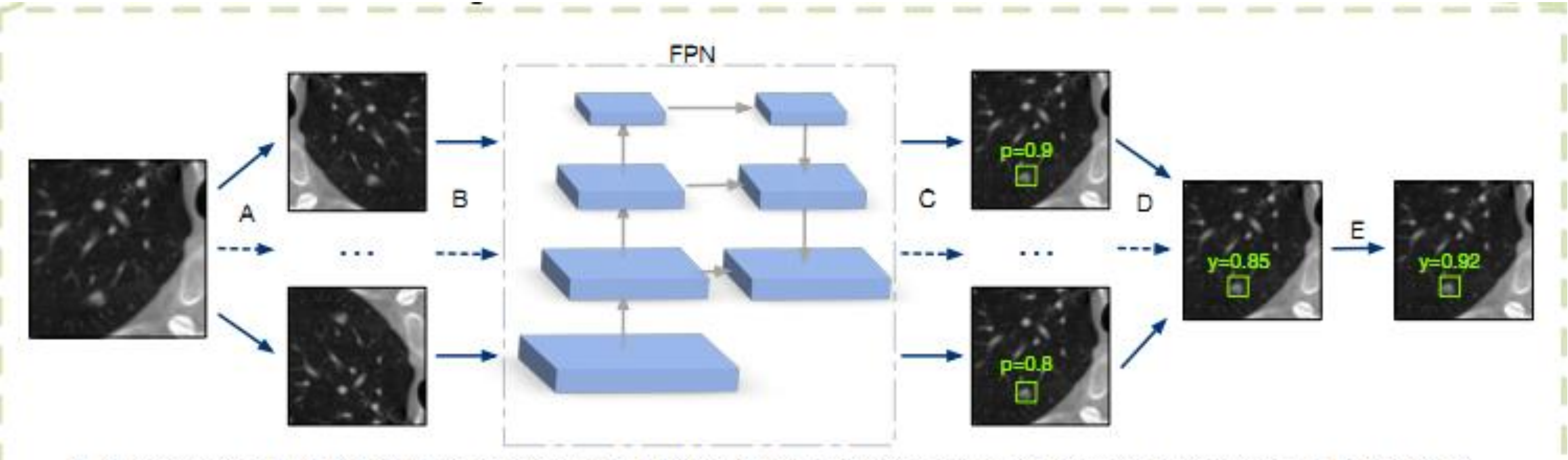
- Soft-target Focal Loss

$$SFL(p) = [\alpha_0 + y(\alpha_1 - \alpha_0)] \cdot |y - p|^\gamma \cdot CE(y, p), \quad (9)$$

, where CE denotes the Cross-entropy loss and y belongs to [0,1].

Methodology

- Anchor-level Target Prediction



A: Augmentation by K times; B: Predicting with FPN; C: Reverse transformation; D: Average ensembling; E: Sharpening

Methodology

- MixUp Augmentation for Detection

1. Image-level: require the capability to detect lesions that are mixed with stronger background noises than usual; y_i for the prediction confidence for each target

$$\lambda \sim \mathrm{Beta}(\eta, \eta), \tag{10}$$

$$\tilde{\lambda} = \max(\lambda, 1 - \lambda), \tag{11}$$

$$\hat{x} = \tilde{\lambda} x + (1 - \tilde{\lambda}) x', \tag{12}$$

$$\hat{y}_i = \tilde{\lambda} y_i + (1 - \tilde{\lambda}) y'_i, \forall i. \tag{13}$$

2. Objective-level (lesion): the annotated boxes and predicted boxes with high prediction confidence; aim to generate extra object instances

Methodology

3. MixUp Augmentation for Detection

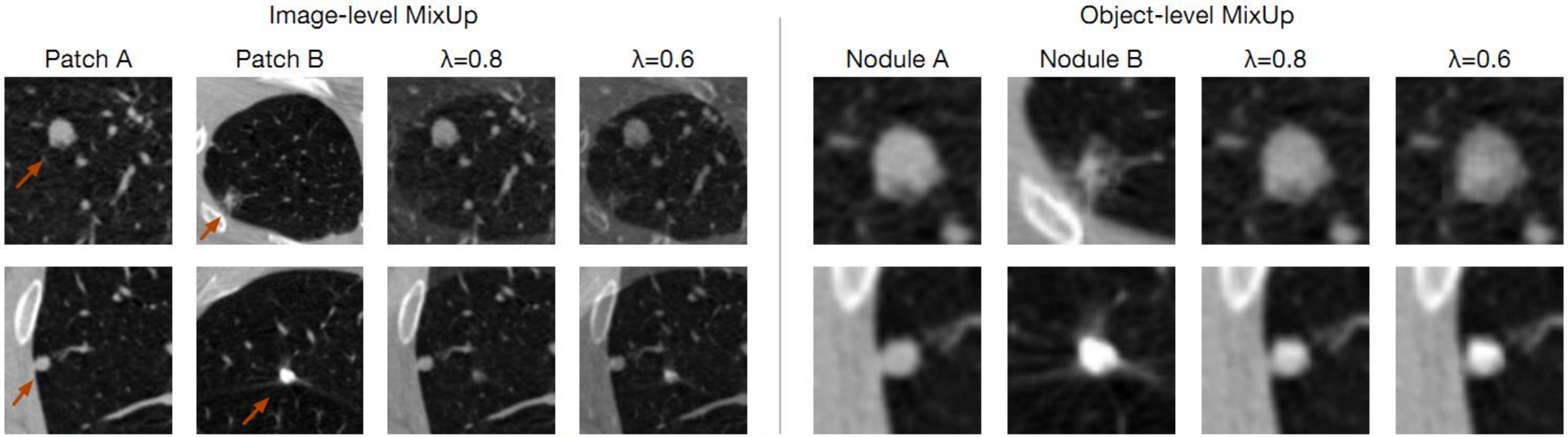


Figure 4: **Illustrative examples for two MixUp methods.** The left figure shows the image-level MixUp, where red arrows point to nodules in the original image. The right figure demonstrates the object-level MixUp, where we zoom in on the nodules and locate them in the center of each image patch for better visualization.

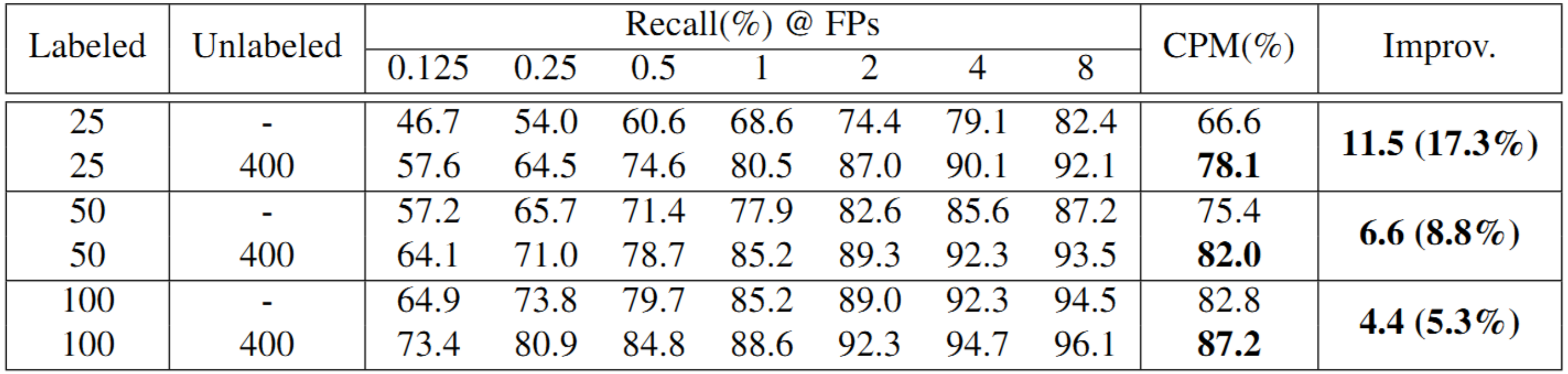
Dataset

1. LUNA16:
   - a high quality subset of the LIDC-IDRI dataset
   - 888 thoracic CT scans in total, with 1186 annotated nodules larger than 3mm.
2. NLST: National Lung Screening Trial:
   - only used as an extra unlabeled dataset after a selection process

Evaluation:

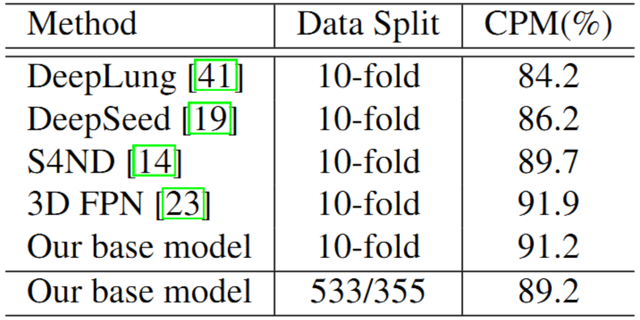
- Free-Response Receiver Operating Characteristic (FROC)
- Competition Performance Metric (CPM): the average recalls when false positive rates are 1/8, 1/4, 1/2, 1, 2, 4, and 8 FPs per scan

Results



Labeled	Unlabeled	Recall(%) @ FPs							CPM(%)	Improv.
		0.125	0.25	0.5	1	2	4	8		
25	-	46.7	54.0	60.6	68.6	74.4	79.1	82.4	66.6	**11.5 (17.3%)**
25	400	57.6	64.5	74.6	80.5	87.0	90.1	92.1	**78.1**	
50	-	57.2	65.7	71.4	77.9	82.6	85.6	87.2	75.4	**6.6 (8.8%)**
50	400	64.1	71.0	78.7	85.2	89.3	92.3	93.5	**82.0**	
100	-	64.9	73.8	79.7	85.2	89.0	92.3	94.5	82.8	**4.4 (5.3%)**
100	400	73.4	80.9	84.8	88.6	92.3	94.7	96.1	**87.2**	

Table 1: **Main results on the LUNA16 dataset.** We evaluate FocalMix with {25, 50, 100} labeled CT scans, respectively. *Improv.* denotes the improvements in CPM over the fully-supervised baseline (relative improvements shown in parentheses).



Method	Data Split	CPM(%)
DeepLung [41]	10-fold	84.2
DeepSeed [19]	10-fold	86.2
S4ND [14]	10-fold	89.7
3D FPN [23]	10-fold	91.9
Our base model	10-fold	91.2
Our base model	533/355	89.2

Table 2: **Performance of the base model used in our experiments.** Our re-implemented 3D FPN is comparable with state-of-the-art single-stage nodule detection models.

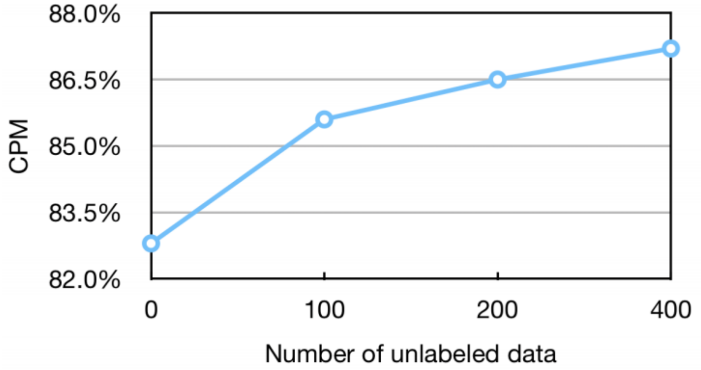
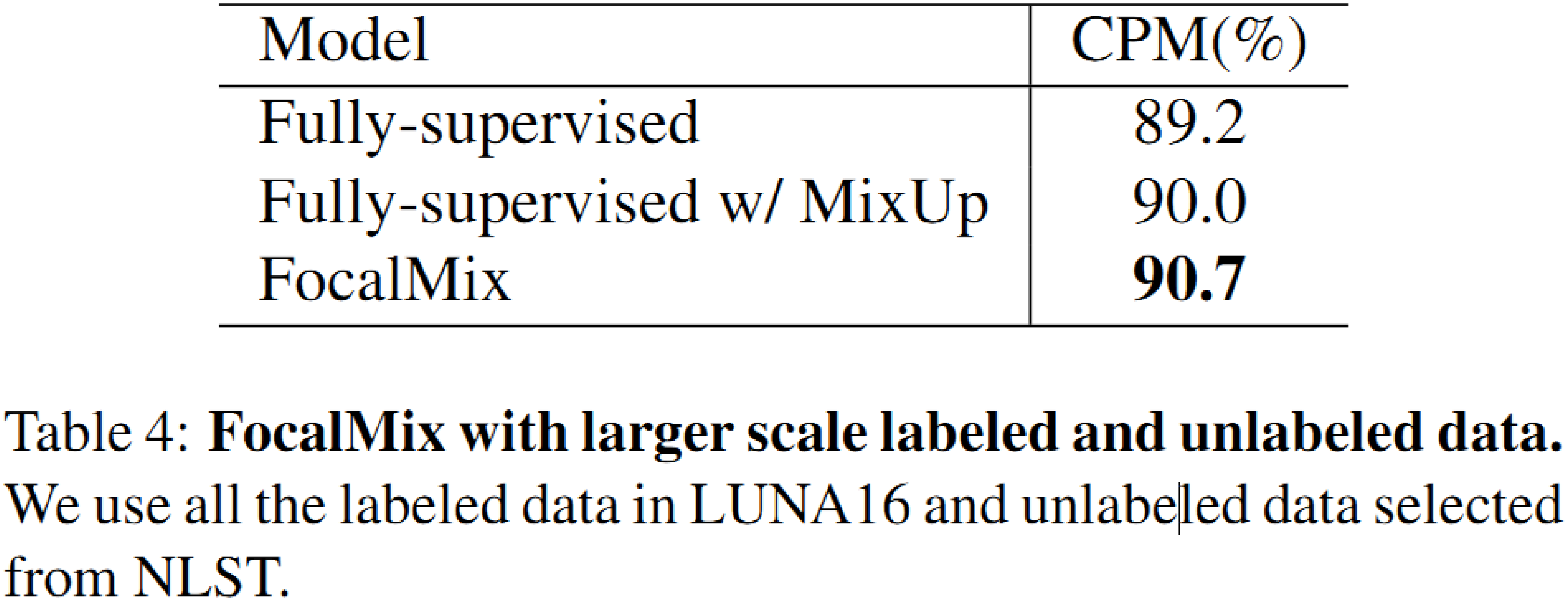


Figure 3: **Performance with different amounts of unlabeled data on LUNA16.** We use 100 labeled images.

Results

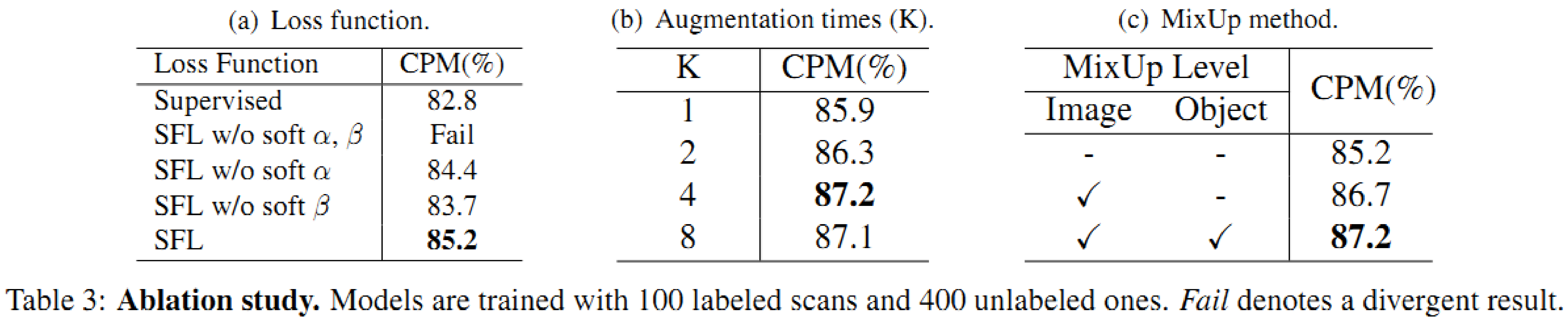
- by leveraging around 3,000 images without annotation



Model	CPM(%)
Fully-supervised	89.2
Fully-supervised w/ MixUp	90.0
FocalMix	**90.7**

Table 4: **FocalMix with larger scale labeled and unlabeled data.** We use all the labeled data in LUNA16 and unlabeled data selected from NLST.

Ablation Study



(a) Loss function.

Loss Function	CPM(%)
Supervised	82.8
SFL w/o soft α, β	Fail
SFL w/o soft α	84.4
SFL w/o soft β	83.7
SFL	**85.2**

(b) Augmentation times (K).

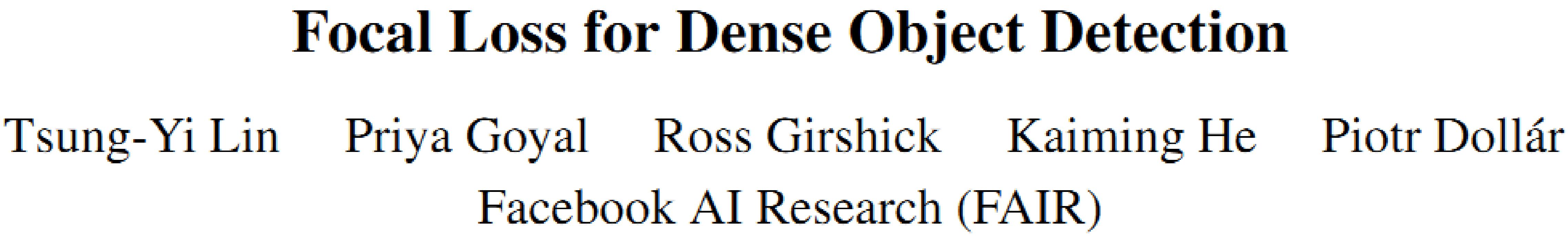
K	CPM(%)
1	85.9
2	86.3
4	**87.2**
8	87.1

(c) MixUp method.

MixUp Level		CPM(%)
Image	Object	
-	-	85.2
✓	-	86.7
✓	✓	**87.2**

Table 3: **Ablation study.** Models are trained with 100 labeled scans and 400 unlabeled ones. *Fail* denotes a divergent result.

Information



Focal Loss for Dense Object Detection

Tsung-Yi Lin Priya Goyal Ross Girshick Kaiming He Piotr Dollár
Facebook AI Research (FAIR)

Motivation

"Despite the success of two-stage detectors, a natural question to ask is: could a simple one-stage detector achieve similar accuracy?"

Introduction

- One stage detectors are applied over a regular, dense sampling of object locations, scales, and aspect ratios.
- Situation: One-stage detectors, such as YOLO and SSD, are faster with accuracy within 10-40% relative to state-of-the-art two-stage methods.
- Main obstacle: class imbalance
- Target: matches the state-of-the-art COCO AP of more complex two-stage detectors

Contributions

1. Focal Loss: address class imbalance.
2. RetinaNet: a simple one-stage object detector
3. Achieves a COCO test-dev AP of 39.1 with 5 fps, surpassing the previously best published single-model results from both one and two-stage detectors

RetinaNet

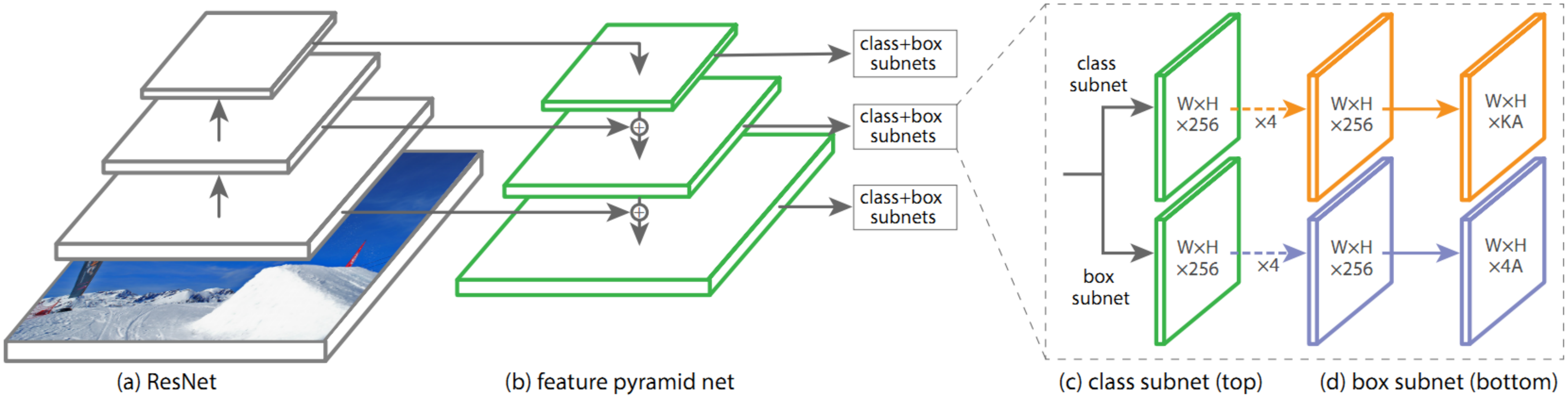
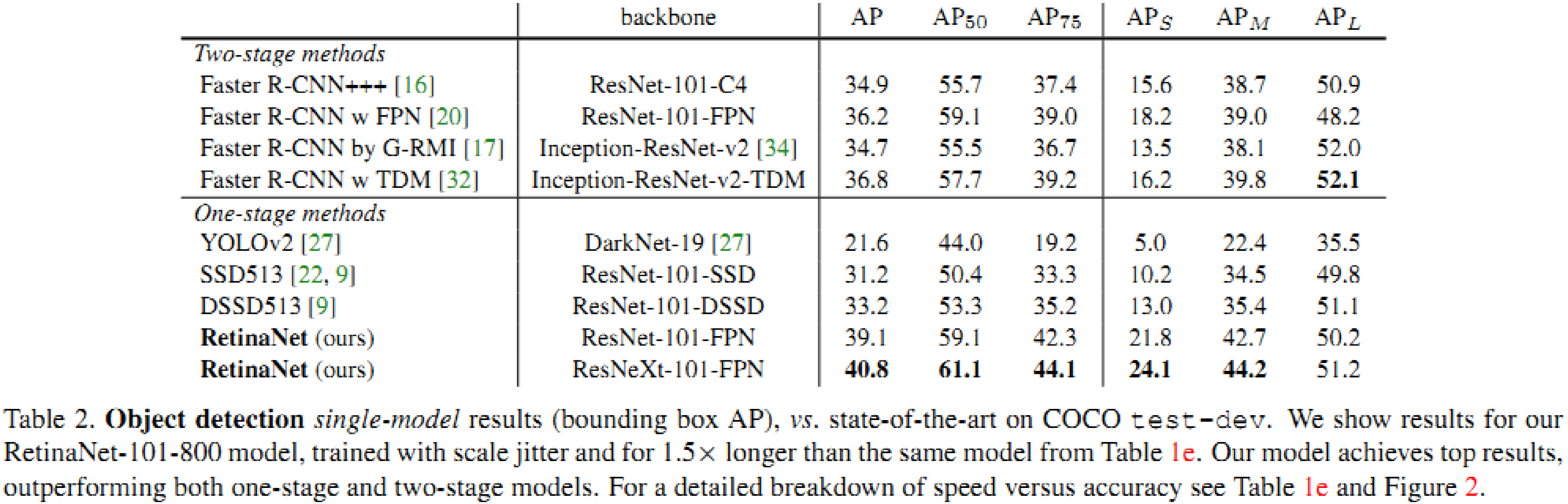


Figure 3. The one-stage **RetinaNet** network architecture uses a Feature Pyramid Network (FPN) [20] backbone on top of a feedforward ResNet architecture [16] (a) to generate a rich, multi-scale convolutional feature pyramid (b). To this backbone RetinaNet attaches two subnetworks, one for classifying anchor boxes (c) and one for regressing from anchor boxes to ground-truth object boxes (d). The network design is intentionally simple, which enables this work to focus on a novel focal loss function that eliminates the accuracy gap between our one-stage detector and state-of-the-art two-stage detectors like Faster R-CNN with FPN [20] while running at faster speeds.

Basic Results



	backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Two-stage methods							
Faster R-CNN+++ [16]	ResNet-101-C4	34.9	55.7	37.4	15.6	38.7	50.9
Faster R-CNN w FPN [20]	ResNet-101-FPN	36.2	59.1	39.0	18.2	39.0	48.2
Faster R-CNN by G-RMI [17]	Inception-ResNet-v2 [34]	34.7	55.5	36.7	13.5	38.1	52.0
Faster R-CNN w TDM [32]	Inception-ResNet-v2-TDM	36.8	57.7	39.2	16.2	39.8	**52.1**
One-stage methods							
YOLOv2 [27]	DarkNet-19 [27]	21.6	44.0	19.2	5.0	22.4	35.5
SSD513 [22, 9]	ResNet-101-SSD	31.2	50.4	33.3	10.2	34.5	49.8
DSSD513 [9]	ResNet-101-DSSD	33.2	53.3	35.2	13.0	35.4	51.1
RetinaNet (ours)	ResNet-101-FPN	39.1	59.1	42.3	21.8	42.7	50.2
RetinaNet (ours)	ResNeXt-101-FPN	**40.8**	**61.1**	**44.1**	**24.1**	**44.2**	51.2

Table 2. **Object detection** *single-model* results (bounding box AP), *vs.* state-of-the-art on COCO `test-dev`. We show results for our RetinaNet-101-800 model, trained with scale jitter and for 1.5× longer than the same model from Table 1e. Our model achieves top results, outperforming both one-stage and two-stage models. For a detailed breakdown of speed versus accuracy see Table 1e and Figure 2.

Basic Results

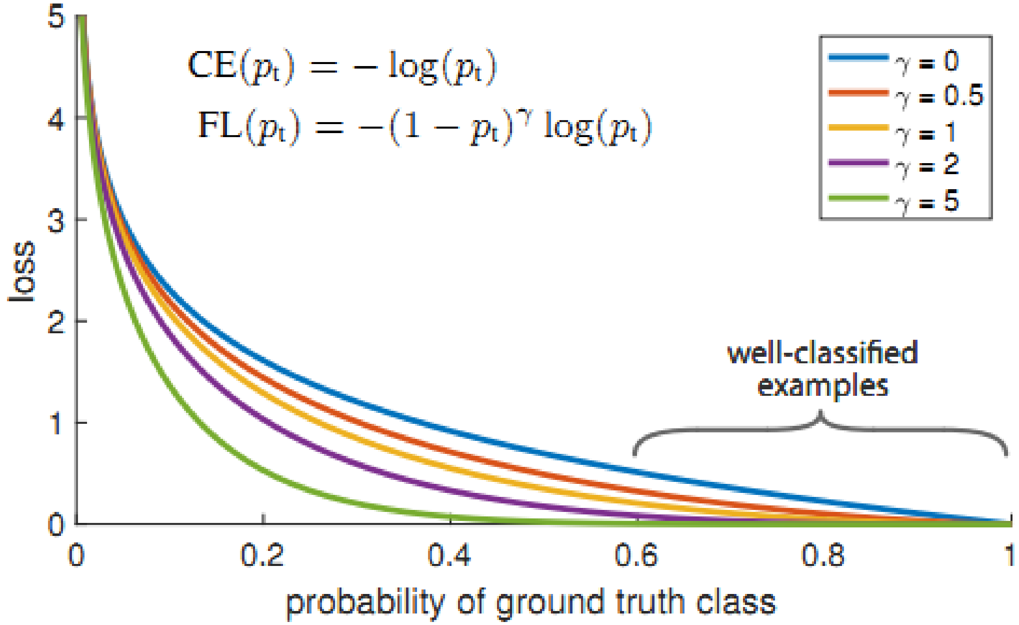
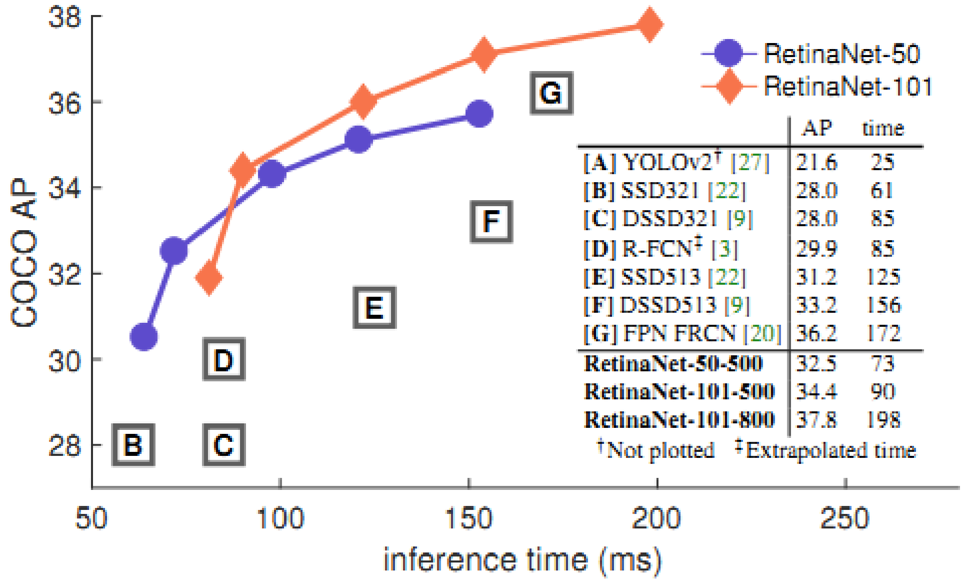


Figure 1. We propose a novel loss we term the *Focal Loss* that adds a factor $(1 - p_t)^\gamma$ to the standard cross entropy criterion. Setting $\gamma > 0$ reduces the relative loss for well-classified examples ($p_t > .5$), putting more focus on hard, misclassified examples. As our experiments will demonstrate, the proposed focal loss enables training highly accurate dense object detectors in the presence of vast numbers of easy background examples.



	AP	time
[A] YOLOv2† [27]	21.6	25
[B] SSD321 [22]	28.0	61
[C] DSSD321 [9]	28.0	85
[D] R-FCN‡ [3]	29.9	85
[E] SSD513 [22]	31.2	125
[F] DSSD513 [9]	33.2	156
[G] FPN FRCN [20]	36.2	172
RetinaNet-50-500	32.5	73
RetinaNet-101-500	34.4	90
RetinaNet-101-800	37.8	198

†Not plotted ‡Extrapolated time

Figure 2. Speed (ms) versus accuracy (AP) on COCO `test-dev`. Enabled by the focal loss, our simple one-stage *RetinaNet* detector outperforms all previous one-stage and two-stage detectors, including the best reported Faster R-CNN [28] system from [20]. We show variants of RetinaNet with ResNet-50-FPN (blue circles) and ResNet-101-FPN (orange diamonds) at five scales (400-800 pixels). Ignoring the low-accuracy regime (AP<25), RetinaNet forms an upper envelope of all current detectors, and an improved variant (not shown) achieves 40.8 AP. Details are given in §5.