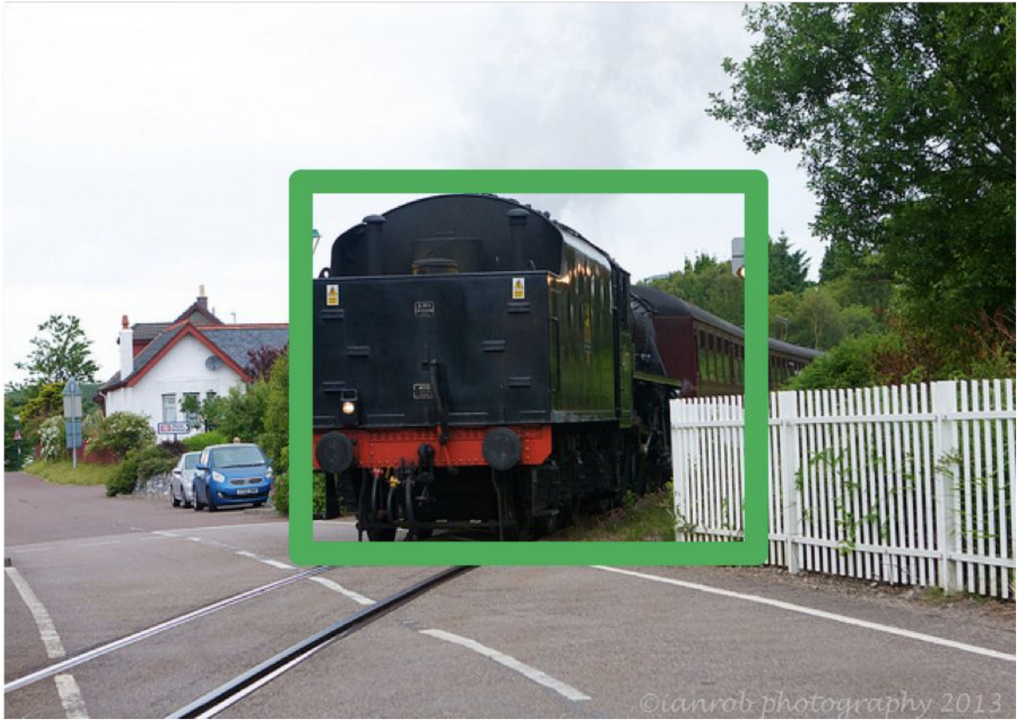
Bounding Box Regression With Uncertainty for Accurate Object Detection

[1]Carnegie Mellon University [2]Megvii
Yihui He[1], Chenchen Zhu[1], Jianren Wang[1], Marios Savvides, [2]Xiangyu Zhang

Ambiguity: inaccurate labelling

- MS-COCO



Ambiguity: inaccurate labelling

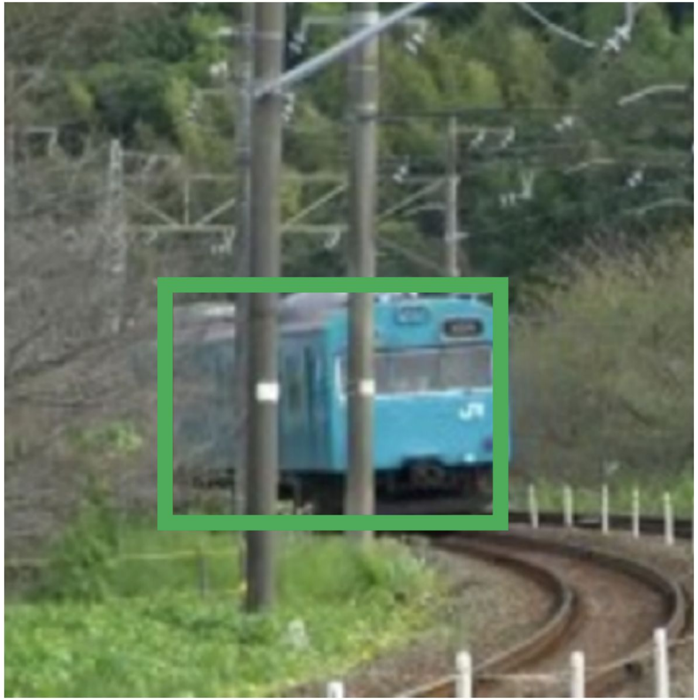
- MS-COCO

Ambiguity: introduced by occlusion

- MS-COCO

Ambiguity: object boundary itself is ambiguous

- YouTube-BoundingBoxes

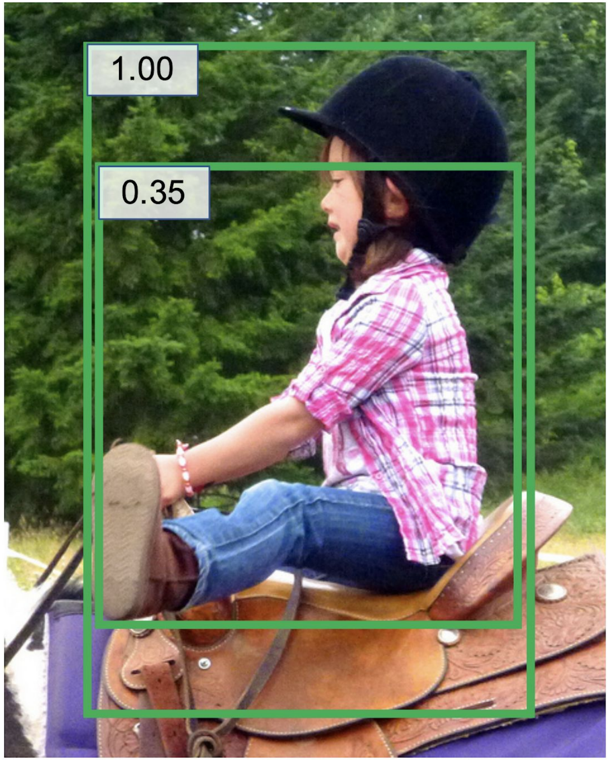


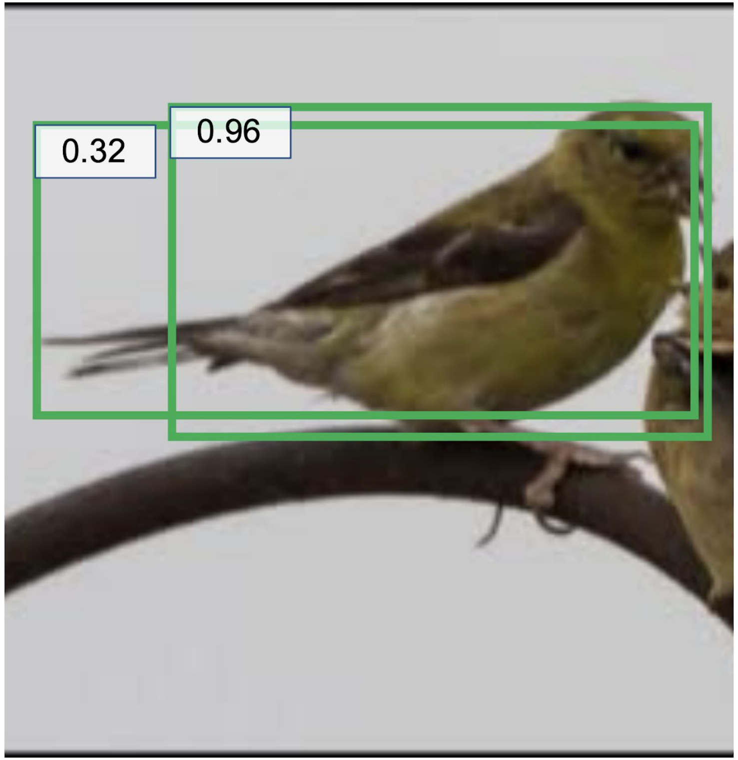
Classification Score & Localization misalignment

MS-COCO

VGG-16

Faster RCNN



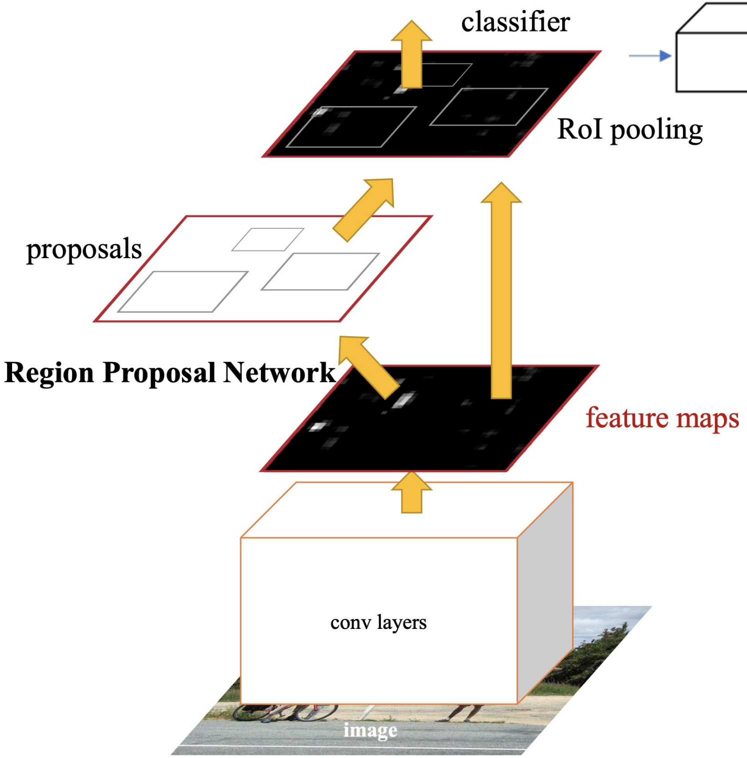


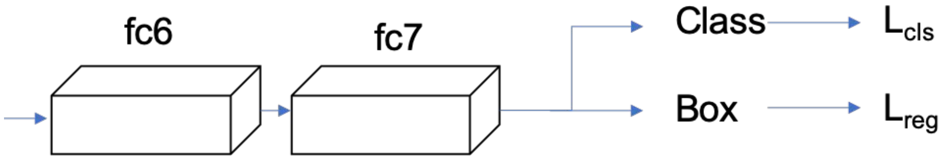
Standard Faster R-CNN Pipeline

Cross entropy/focal loss

1024 x 81

1024 x 81x4



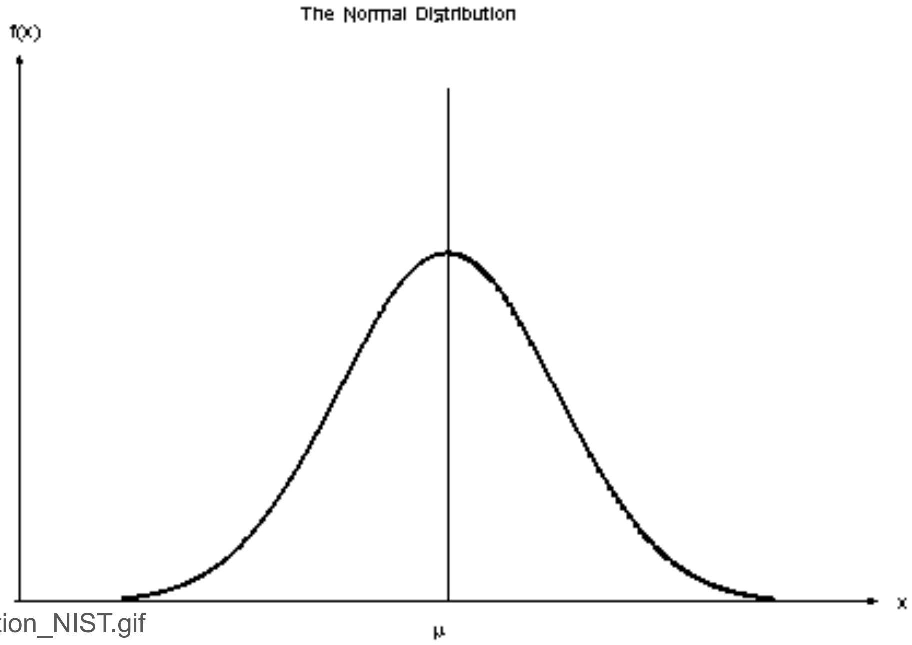


$$\text{smooth}_{L_1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise} \end{cases}$$

Modeling bounding box prediction

- Predict Gaussian distribution instead of a number

$$P_{\Theta}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-x_e)^2}{2\sigma^2}}$$

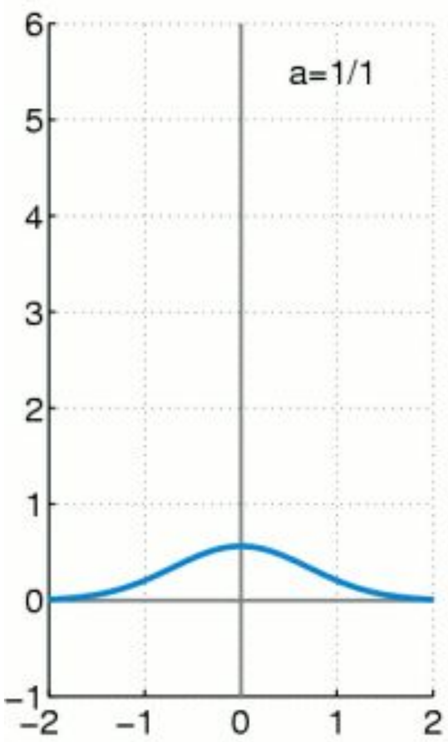


https://upload.wikimedia.org/wikipedia/commons/9/9e/Normal_Distribution_NIST.gif

Modeling ground truth bounding box

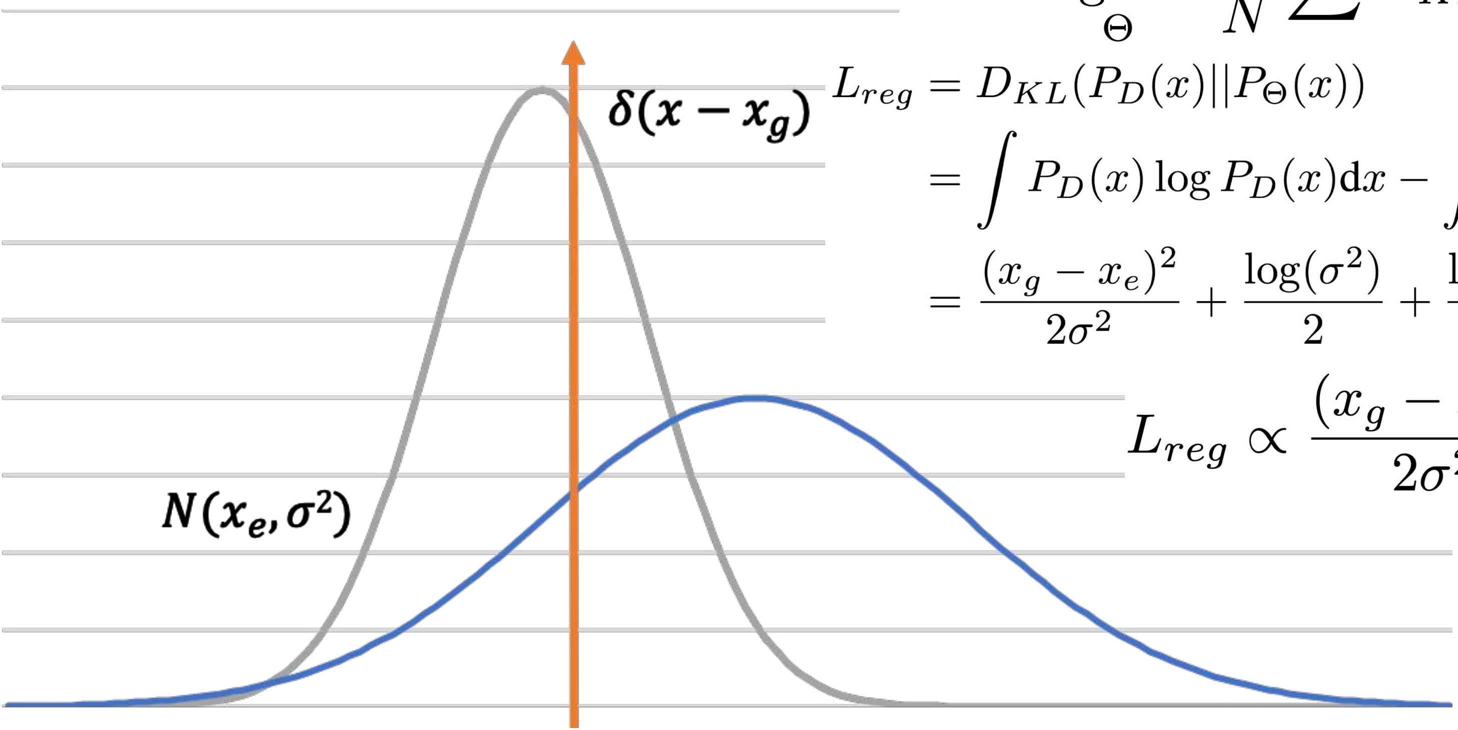
- Dirac delta function

$$P_D(x) = \delta(x - x_g)$$



https://upload.wikimedia.org/wikipedia/commons/b/b4/Dirac_function_approximation.gif

KL Loss: Gaussian meets delta function



$$\hat{\Theta} = \arg\min_{\Theta} \frac{1}{N} \sum D_{KL}(P_D(x) || P_\Theta(x))$$

$$L_{reg} = D_{KL}(P_D(x) || P_\Theta(x))$$

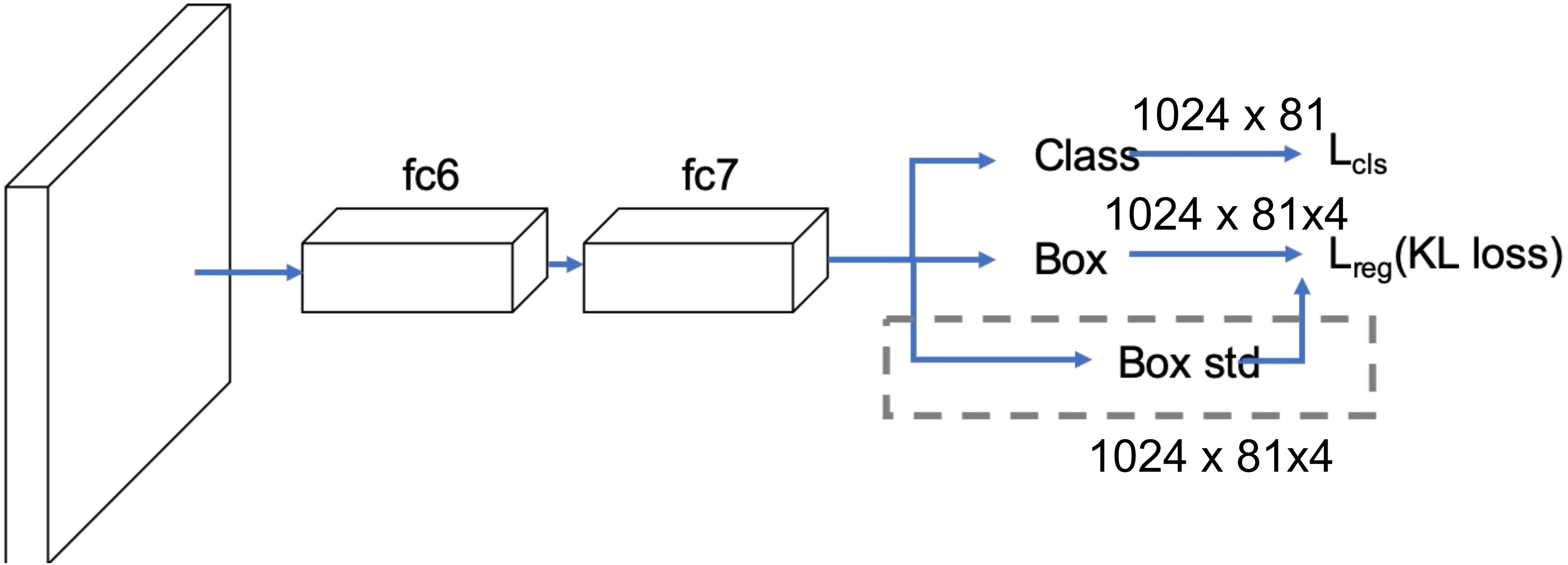
$$= \int P_D(x) \log P_D(x) \mathrm{d}x - \int P_D(x) \log P_\Theta(x) \mathrm{d}x$$

$$= \frac{(x_g - x_e)^2}{2\sigma^2} + \frac{\log(\sigma^2)}{2} + \frac{\log(2\pi)}{2} - H(P_D(x))$$

$$L_{reg} \propto \frac{(x_g - x_e)^2}{2\sigma^2} + \frac{1}{2}\log(\sigma^2)$$

Architecture

An additional fully-connected layer for prediction variance (1024 x 81 x 4)



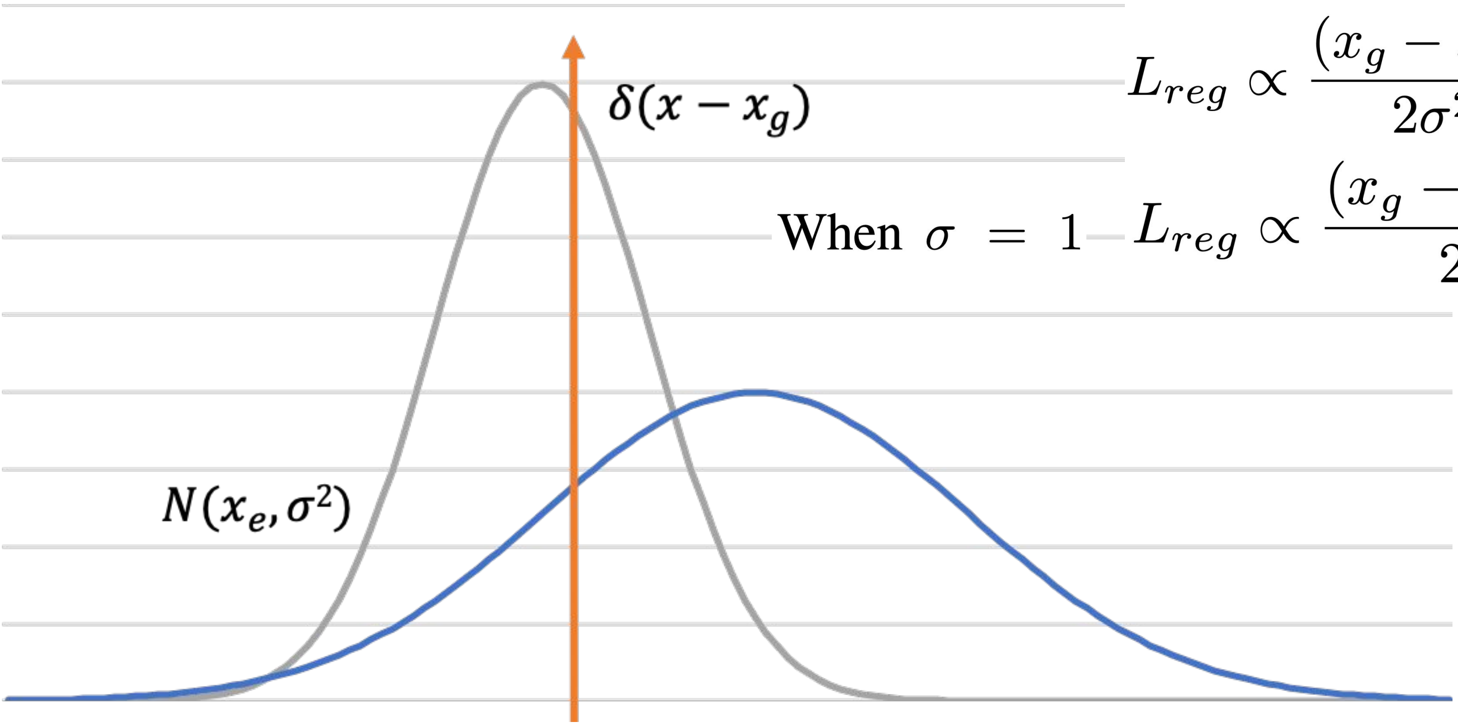
Why KL Loss

(1) The ambiguities in a dataset can be successfully captured. The bounding box regressor gets smaller loss from ambiguous bounding boxes.

(2) The learned variance is useful during post-processing. We propose var voting (variance voting) to vote the location of a candidate box using its neighbors' locations weighted by the predicted variances during nonmaximum suppression (NMS).

(3) The learned probability distribution is interpretable. Since it reflects the level of uncertainty of the bounding box prediction, it can potentially be helpful in down-stream applications like self-driving cars and robotics

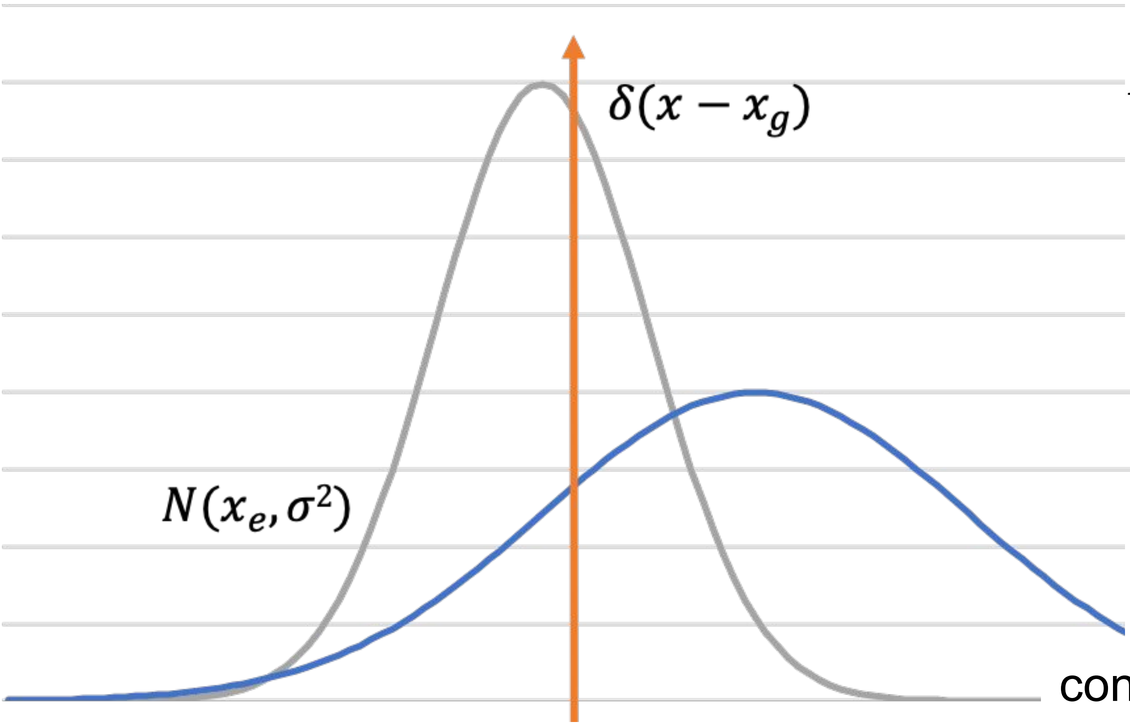
KL Loss: Degradation Case



$$L_{reg} \propto \frac{(x_g - x_e)^2}{2\sigma^2} + \frac{1}{2}\log(\sigma^2)$$

When $\sigma = 1$ $$L_{reg} \propto \frac{(x_g - x_e)^2}{2}$$

KL Loss: Reparameterization trick



$$L_{reg} \propto \frac{(x_g - x_e)^2}{2\sigma^2} + \frac{1}{2}\log(\sigma^2)$$

$$\frac{\mathrm{d}}{\mathrm{d}x_e} L_{reg} = \frac{x_e - x_g}{\sigma^2}$$

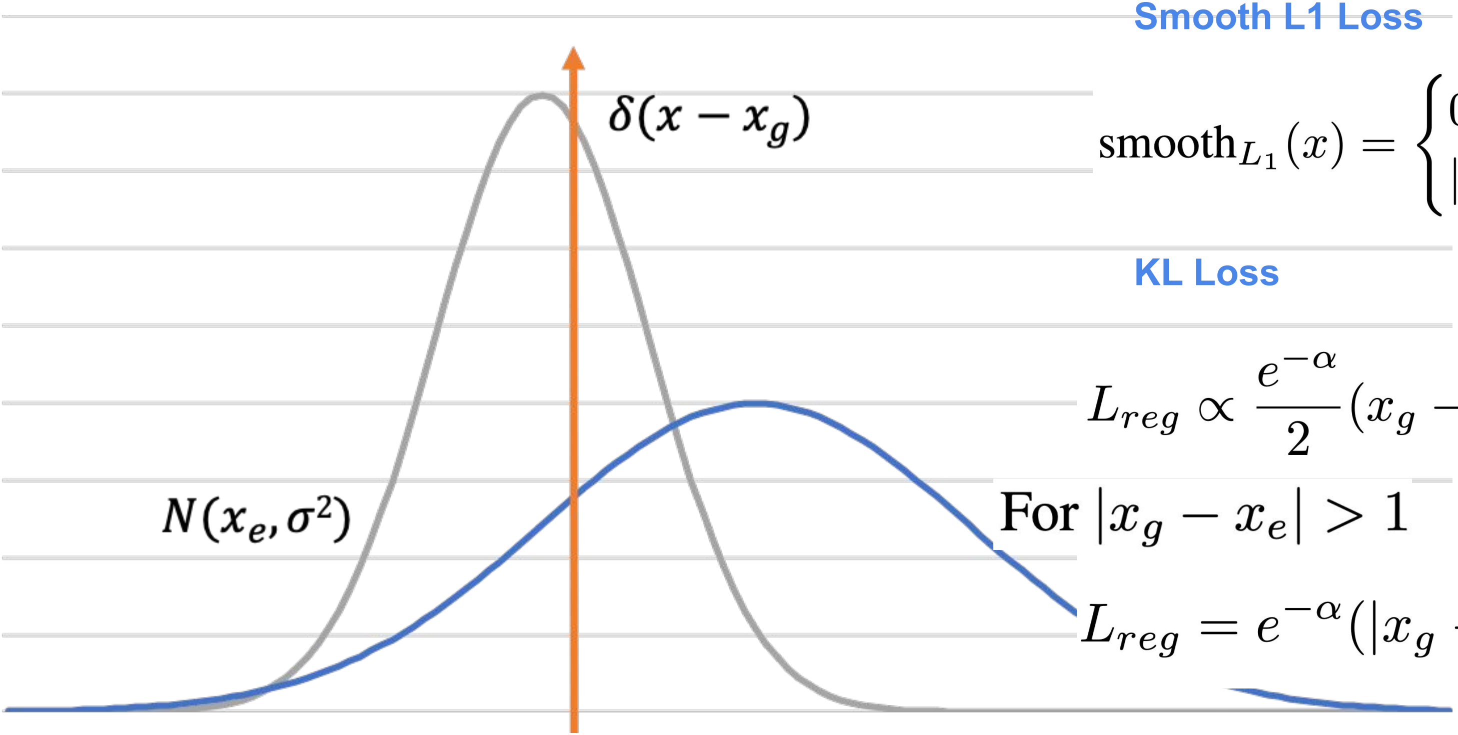
$$\frac{\mathrm{d}}{\mathrm{d}\sigma} L_{reg} = -\frac{(x_e - x_g)^2}{\sigma^3} + \frac{1}{\sigma}$$

predicts $\alpha = \log(\sigma^2)$

$$L_{reg} \propto \frac{e^{-\alpha}}{2}(x_g - x_e)^2 + \frac{1}{2}\alpha$$

convert α back to σ during testing

KL Loss: Rubust L1 Loss (Smooth L1 Loss)



Smooth L1 Loss

$$\text{smooth}_{L_1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise} \end{cases}$$

KL Loss

$$L_{reg} \propto \frac{e^{-\alpha}}{2}(x_g - x_e)^2 + \frac{1}{2}\alpha$$

For $|x_g - x_e| > 1$

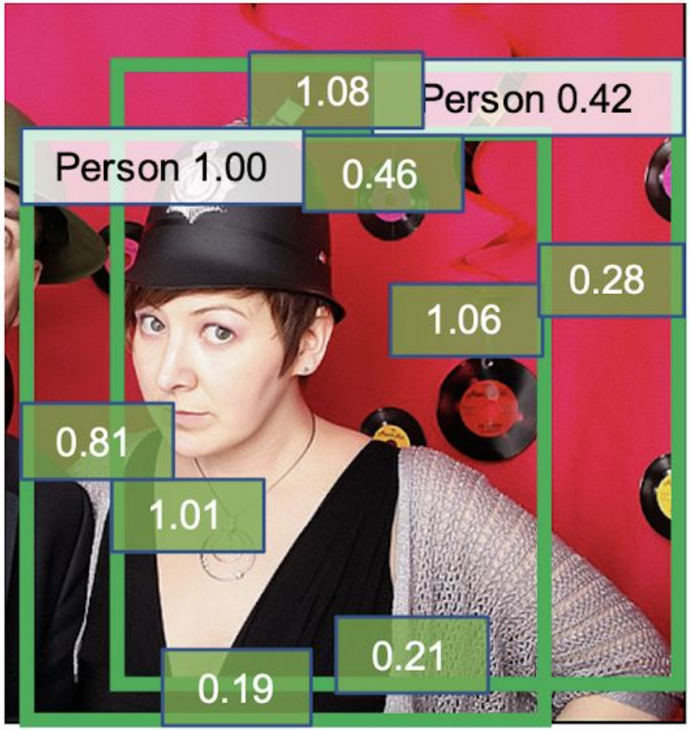
$$L_{reg} = e^{-\alpha}(|x_g - x_e| - \frac{1}{2}) + \frac{1}{2}\alpha$$

KL Loss: Uncertainty Prediction

Sigma in Green box

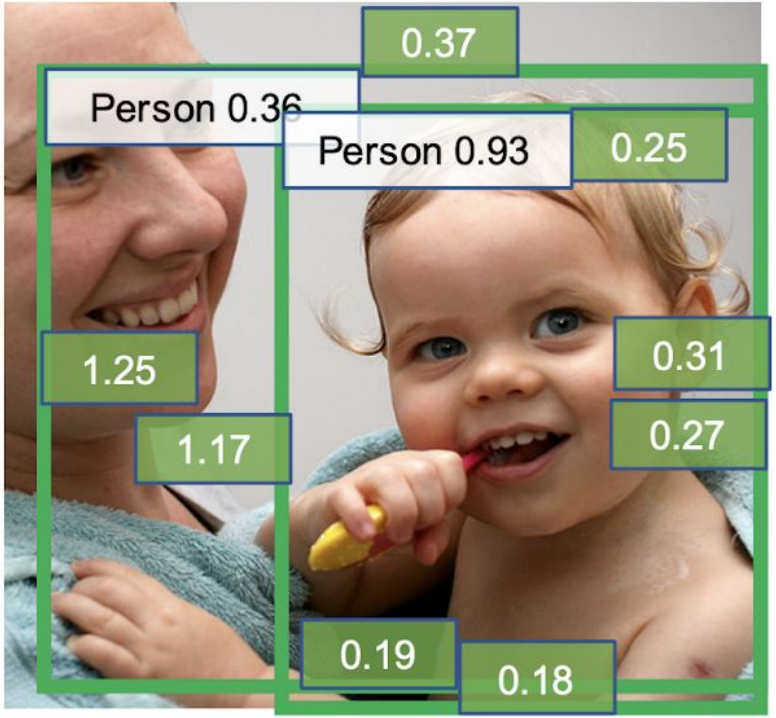
KL Loss: Uncertainty Prediction

Sigma in Green box



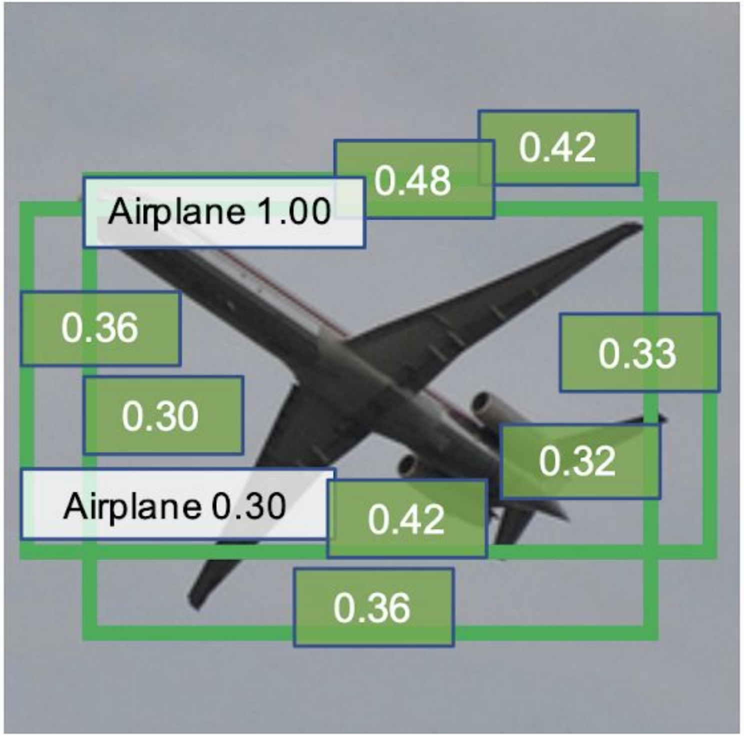
KL Loss: Uncertainty Prediction

Sigma in Green box



KL Loss: Uncertainty Prediction

Sigma in Green box



Variance Voting

- **Larger IoU** gets higher score
- **Lower variance** gets higher score
- **Classification score invariance**

$$p_i = e^{-(1-IoU(b_i,b))^2/\sigma_t}$$

$$x = \frac{\sum_i p_i x_i/\sigma^2_{x,i}}{\sum_i p_i/\sigma^2_{x,i}}$$

subject to $IoU(b_i, b) > 0$

Algorithm 1 var voting

$\mathcal{B}$ is $N \times 4$ matrix of initial detection boxes. $\mathcal{S}$ contains corresponding detection scores. $\mathcal{C}$ is $N \times 4$ matrix of corresponding variances. $\mathcal{D}$ is the final set of detections. σ_t is a tunable parameter of var voting. The lines in blue and in green are soft-NMS and var voting respectively.

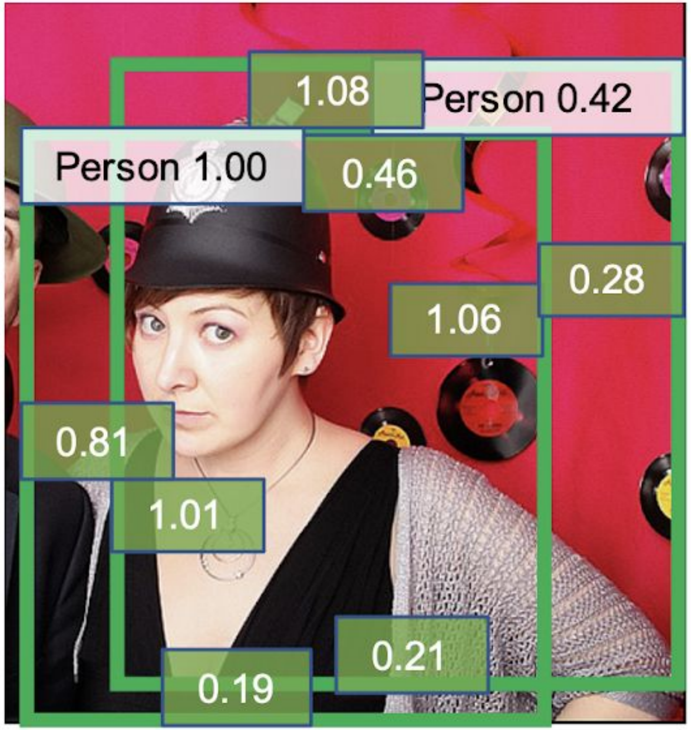
$\mathcal{B} = \{b_1, .., b_N\}, \mathcal{S} = \{s_1, .., s_N\}, \mathcal{C} = \{\sigma_1^2, .., \sigma_N^2\}$
$\mathcal{D} \leftarrow \{\}$
$\mathcal{T} \leftarrow \mathcal{B}$
while $\mathcal{T} \neq$ empty **do**
  $m \leftarrow$ argmax $\mathcal{S}$
  $\mathcal{T} \leftarrow \mathcal{T} - b_m$
  $\mathcal{S} \leftarrow \mathcal{S} f(IoU(b_m, \mathcal{T}))$ ▷ soft-NMS
  $idx \leftarrow IoU(b_m, B) > 0$ ▷ var voting
  $p \leftarrow exp(-(1 - IoU(b_m, \mathcal{B}[idx]))^2/\sigma_t)$
  $b_m \leftarrow p(\mathcal{B}[idx]/\mathcal{C}[idx])/p(1/\mathcal{C}[idx])$
  $\mathcal{D} \leftarrow \mathcal{D} \bigcup b_m$
end while
return $\mathcal{D}, \mathcal{S}$

Variance Voting

Before

after

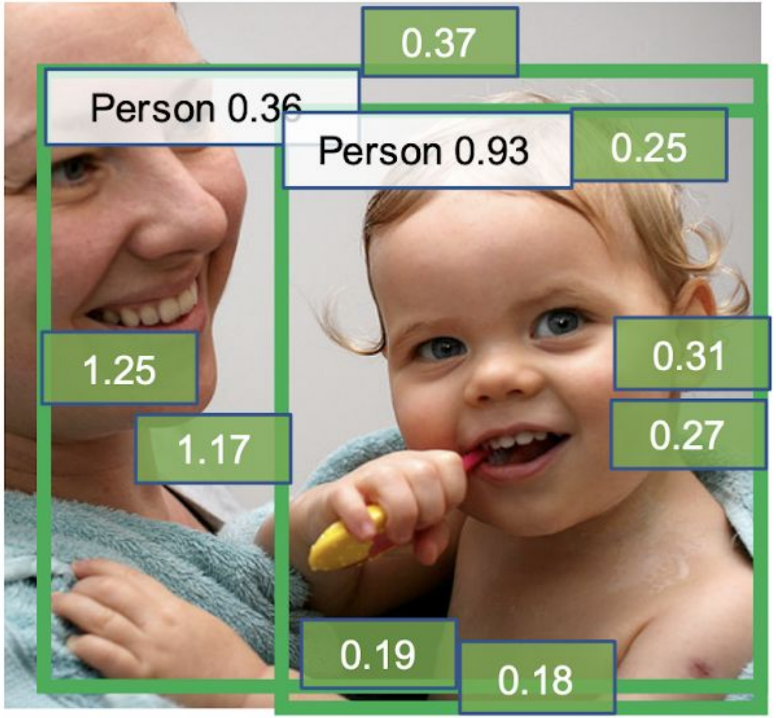
Variance Voting



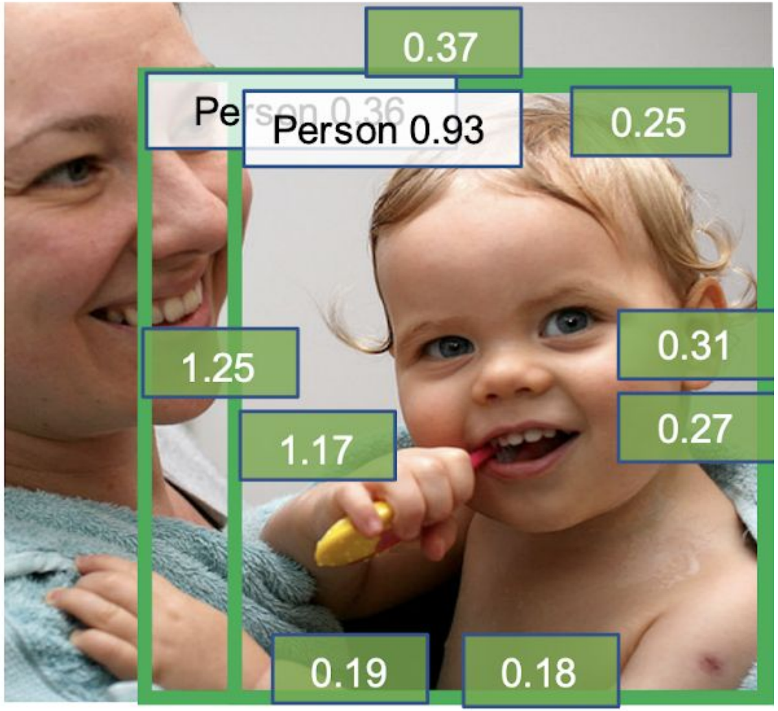
Before

after

Variance Voting

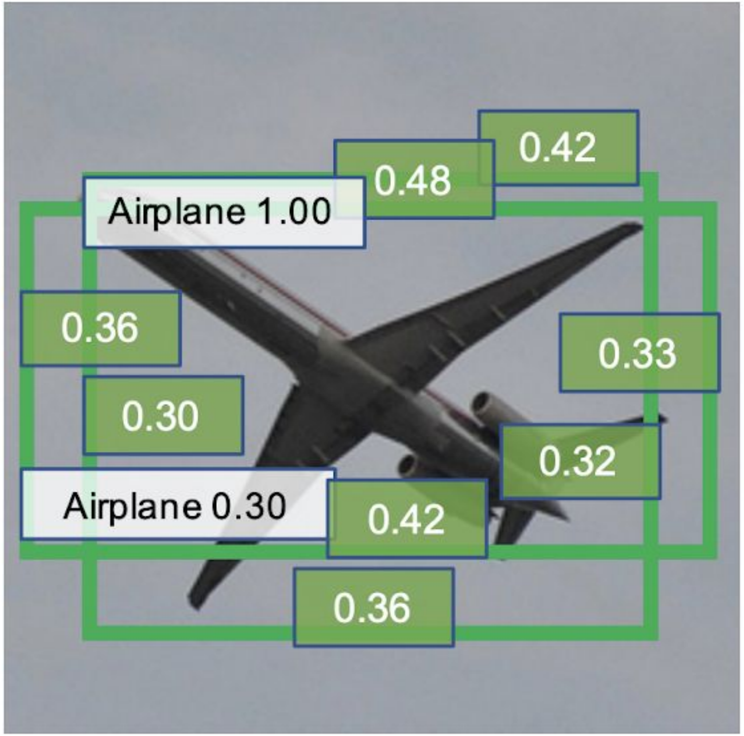


Before

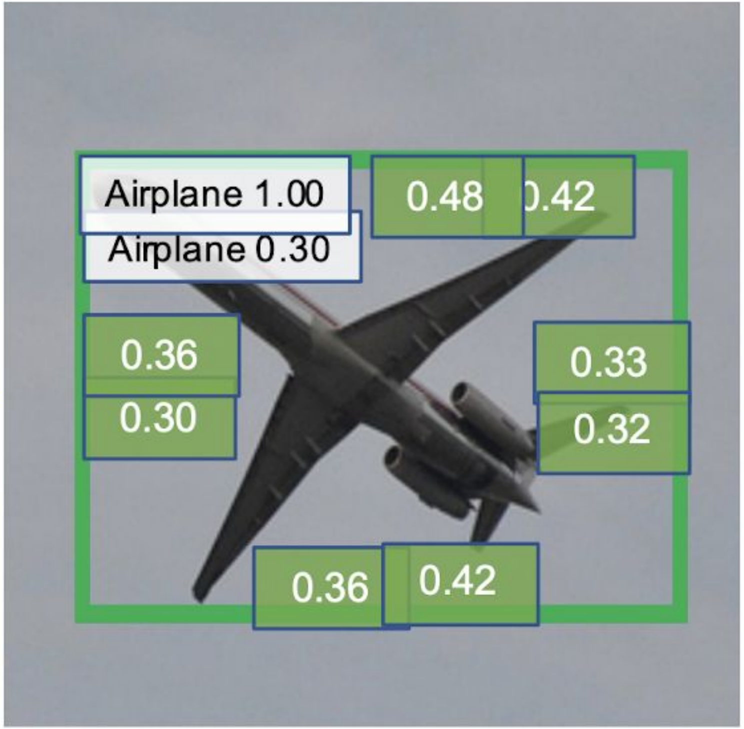


after

Variance Voting



Before



after

Ablation Study: KL Loss, soft-NMS, Variance Voting

- VGG-16
- MS-COCO

KL Loss	soft-NMS	var voting	AP	AP^{50}	AP^{75}	AP^{S}	AP^{M}	AP^{L}	AR^{1}	AR^{10}	AR^{100}
			23.6	44.6	22.8	6.7	25.9	36.3	23.3	33.6	34.3
	✓		24.8	45.6	24.6	7.6	27.2	37.6	23.4	39.2	42.2
✓			26.4	47.9	26.4	7.4	29.3	41.2	25.2	36.1	36.9
✓		✓	27.8	48.0	28.9	8.1	31.4	42.6	26.2	37.5	38.3
✓	✓		27.8	49.0	28.5	8.4	30.9	42.7	25.3	41.7	44.9
✓	✓	✓	**29.1**	**49.1**	**30.4**	**8.7**	**32.7**	**44.3**	**26.2**	**42.5**	**45.5**

Ablation Study: does #params in head matter?

The Larger R-CNN head, the better

fast R-CNN head	backbone	KL Loss	AP
2mlp head	FPN		37.9
		✓	$38.5^{+0.6}$
2mlp head + mask	FPN		38.6
		✓	$39.5^{+\mathbf{0.9}}$
conv5 head	RPN		36.5
		✓	$38.0^{+\mathbf{1.5}}$

Ablation Study: Variance Voting Threshold

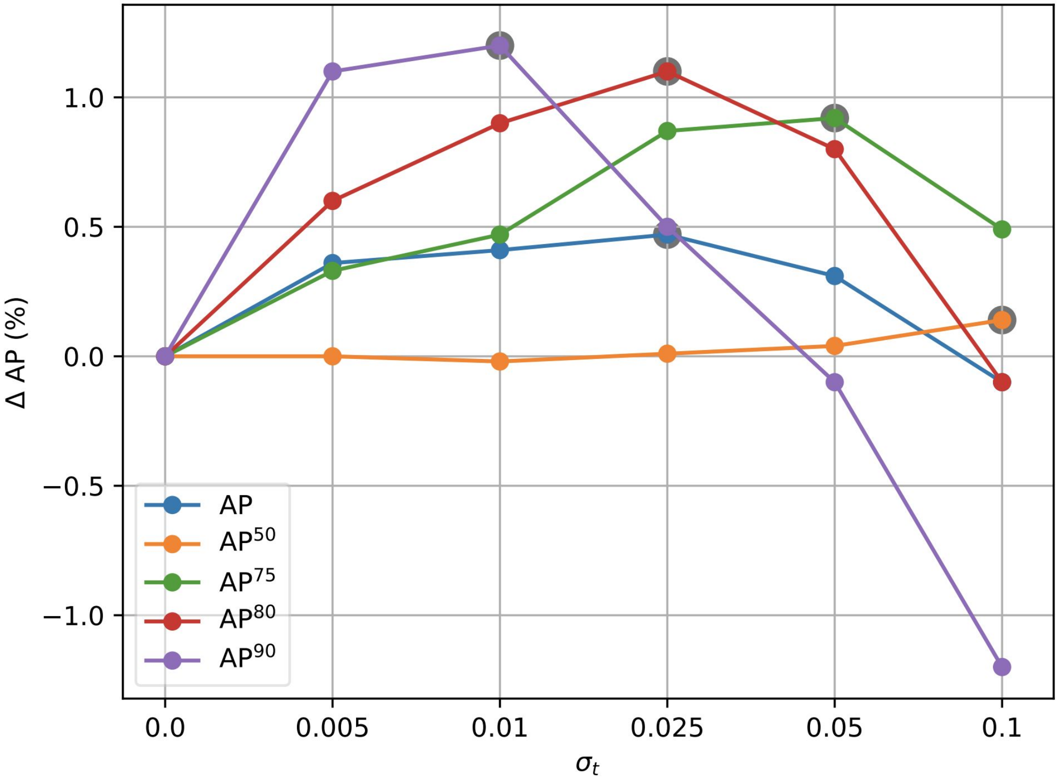
$\sigma_t = 0$, standard NMS

Large σ_t:
farther boxes are considered

$$p_i = e^{-(1-IoU(b_i,b))^2/\sigma_t}$$

$$x = \frac{\sum_i p_i x_i/\sigma_{x,i}^2}{\sum_i p_i/\sigma_{x,i}^2}$$

$$\text{subject to } IoU(b_i,b) > 0$$



Improving State-of-the-Art

- Mask R-CNN
- MS-COCO

	AP	AP^{50}	AP^{60}	AP^{70}	AP^{80}	AP^{90}
baseline [14]	38.6	**59.8**	55.3	47.7	34.4	11.3
MR-CNN [11]	38.9	**59.8**	55.5	48.1	$34.8^{+0.4}$	$11.9^{+0.6}$
soft-NMS [1]	39.3	59.7	**55.6**	**48.9**	$35.9^{+1.5}$	$12.0^{+0.7}$
IoU-NMS+Refine [27]	39.2	57.9	53.6	47.4	$36.5^{+2.1}$	$16.4^{+5.1}$
KL Loss	$39.5^{+0.9}$	58.9	54.4	47.6	$36.0^{+1.6}$	$15.8^{+4.5}$
KL Loss+var voting	$39.9^{+1.3}$	58.9	54.4	47.7	$36.4^{+2.0}$	$17.0^{+5.7}$
KL Loss+var voting+soft-NMS	$\mathbf{40.4}^{+1.8}$	58.7	54.6	48.5	$\mathbf{37.5}^{+3.3}$	$\mathbf{17.5}^{+6.2}$

Inference Latency

- VGG-16
- single image
- single GTX 1080 Ti GPU

method	latency (ms)
baseline	99
ours	101

2ms

Other models on MS-COCO

type	method	AP	AP^{50}	AP^{75}	AP^{S}	AP^{M}	AP^{L}
fast R-CNN	baseline (1x schedule) [14]	36.4	**58.4**	39.3	**20.3**	39.8	48.1
	baseline (2x schedule) [14]	36.8	**58.4**	39.5	19.8	39.5	49.5
	IoU-NMS [27]	37.3	56.0	-	-	-	-
	soft-NMS [1]	37.4	58.2	41.0	**20.3**	40.2	50.1
	KL Loss	37.2	57.2	39.9	19.8	39.7	50.1
	KL Loss+var voting	37.5	56.5	40.1	19.4	40.2	51.6
	KL Loss+var voting+soft-NMS	**38.0**	56.4	**41.2**	19.8	**40.6**	**52.3**
Faster R-CNN	baseline (1x schedule) [14]	36.7	58.4	39.6	21.1	39.8	48.1
	IoU-Net [27]	37.0	58.3	-	-	-	-
	IoU-Net+IoU-NMS [27]	37.6	56.2	-	-	-	-
	baseline (2x schedule) [14]	37.9	59.2	41.1	21.5	41.1	49.9
	IoU-Net+IoU-NMS+Refine [27]	38.1	56.3	-	-	-	-
	soft-NMS[1]	38.6	**59.3**	42.4	21.9	**41.9**	50.7
	KL Loss	38.5	57.8	41.2	20.9	41.2	51.5
	KL Loss+var voting	38.8	57.8	41.6	21.0	41.5	52.0
	KL Loss+var voting+soft-NMS	**39.2**	57.6	**42.5**	21.2	41.8	**52.5**

VGG on PASCAL VOC

backbone	method	mAP
VGG-CNN-M-1024	baseline	60.4
	KL Loss	62.0
	KL Loss+var voting	62.8
	KL Loss+var voting+soft-NMS	**63.6**
VGG-16	baseline	68.7
	QUBO (tabu) [46]	60.6
	QUBO (greedy) [46]	61.9
	soft-NMS [1]	70.1
	KL Loss	69.7
	KL Loss+var voting	70.2
	KL Loss+var voting+soft-NMS	**71.6**

Join us at Tuesday Afternoon Poster Session #41

Bounding Box Regression with Uncertainty for Accurate Object Detection