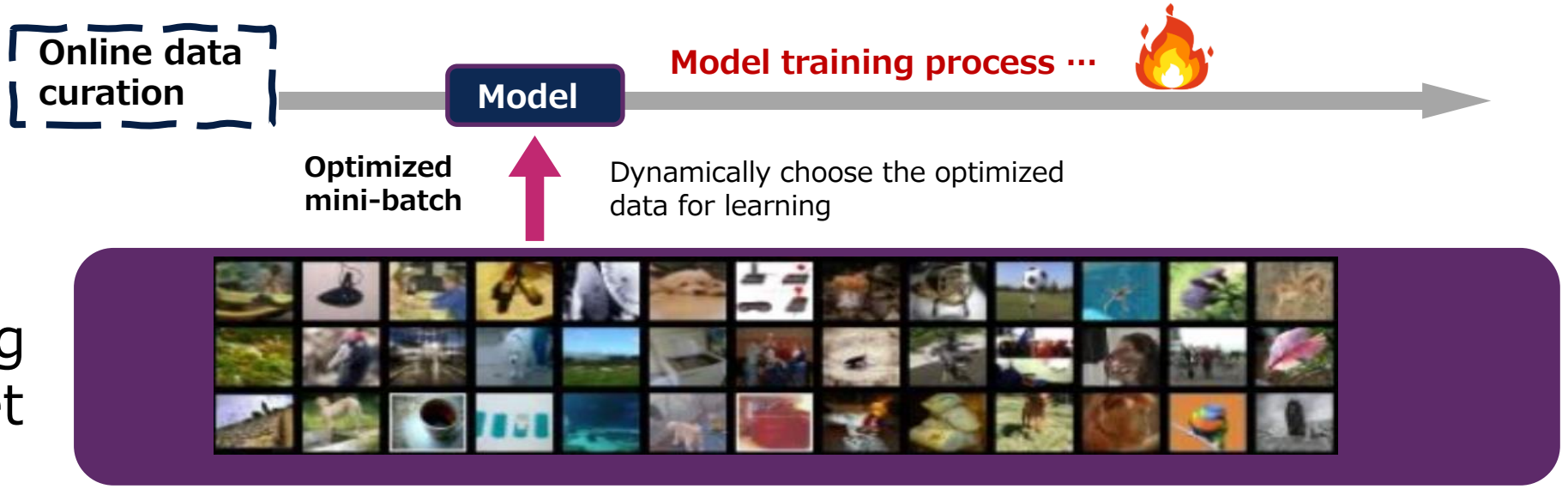
SONY

Online Data curation for Object Detection via Marginal Contribution to Dataset-level Average Precision

Zitang Sun, Masakazu Yoshimura, Junji Otsuka, Atsushi Irie, Takeshi Ohashi
(Sony Group Corporation)

Background & Motivation

●training pipeline in vision model



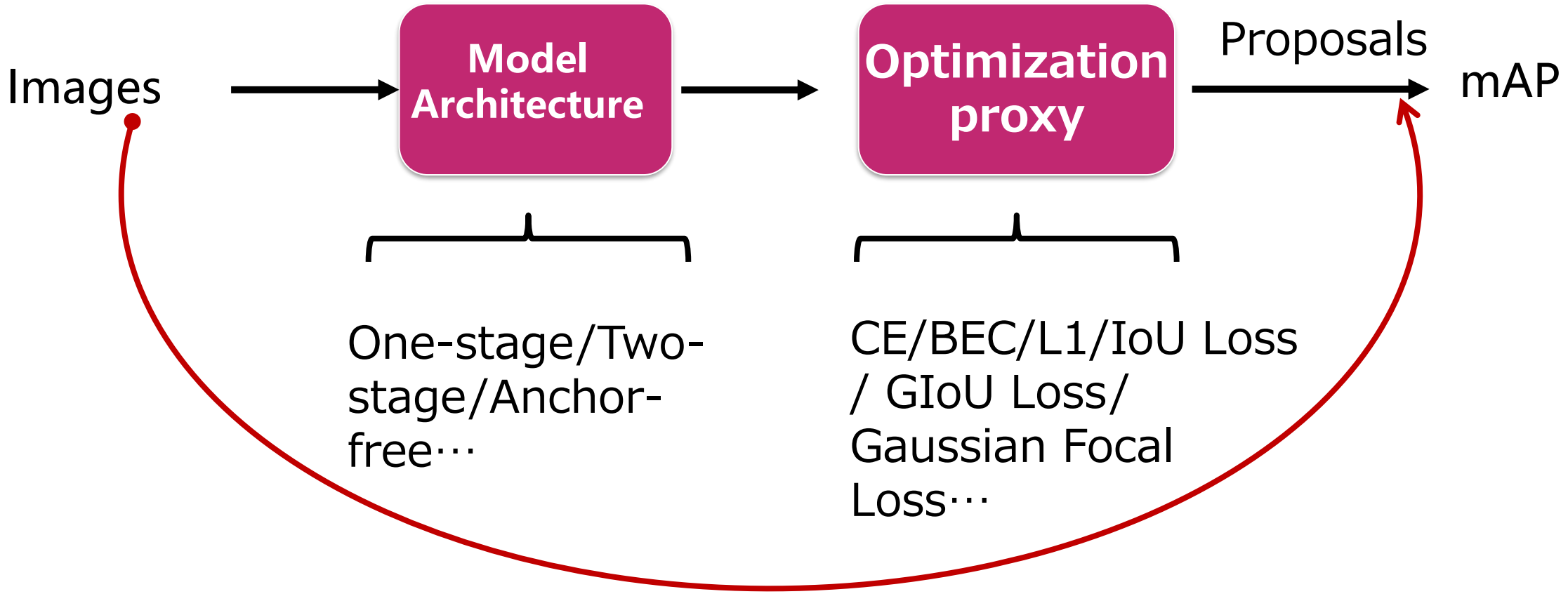
Data curation offline: Make sure the data is balanced, diverse, (Active learning) etc。

Data curation online: Beyond that, also track the model's learning progress and dynamically decide the best data based on model's current state of each iteration.

- We focus **online data curation**
- We focus **object detection task**

◆ Method & Implementation

- Marginal Contribution to the overall mean Average Precision (mAP)

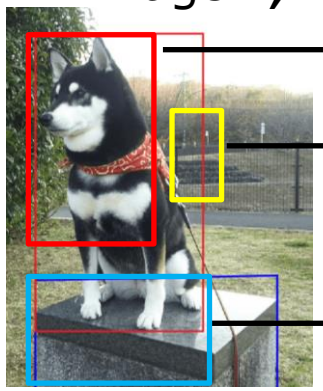


- We avoid the middle part and direct consider the Marginal Contribution to mAP of each sample. (Theoretically apply to any det models)

◆ Method & Implementation

● Learnability based on Marginal Contribution to global mAP over whole dataset (DetGain of each image)

Image ***x***, whole training set *D*, *detection model f*



True Positive (Score = 0.8)

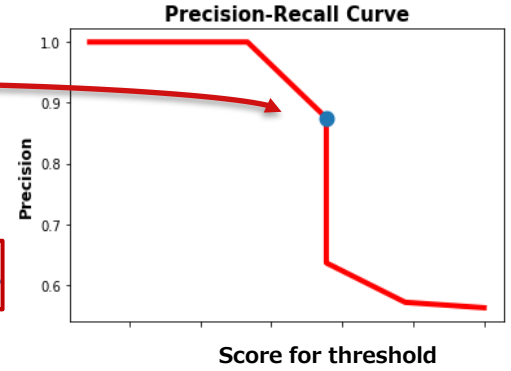
False Positive (Score = 0.3)

True Positive (Score = 0.6)

$$Precision = \frac{TP}{TP + FP}$$

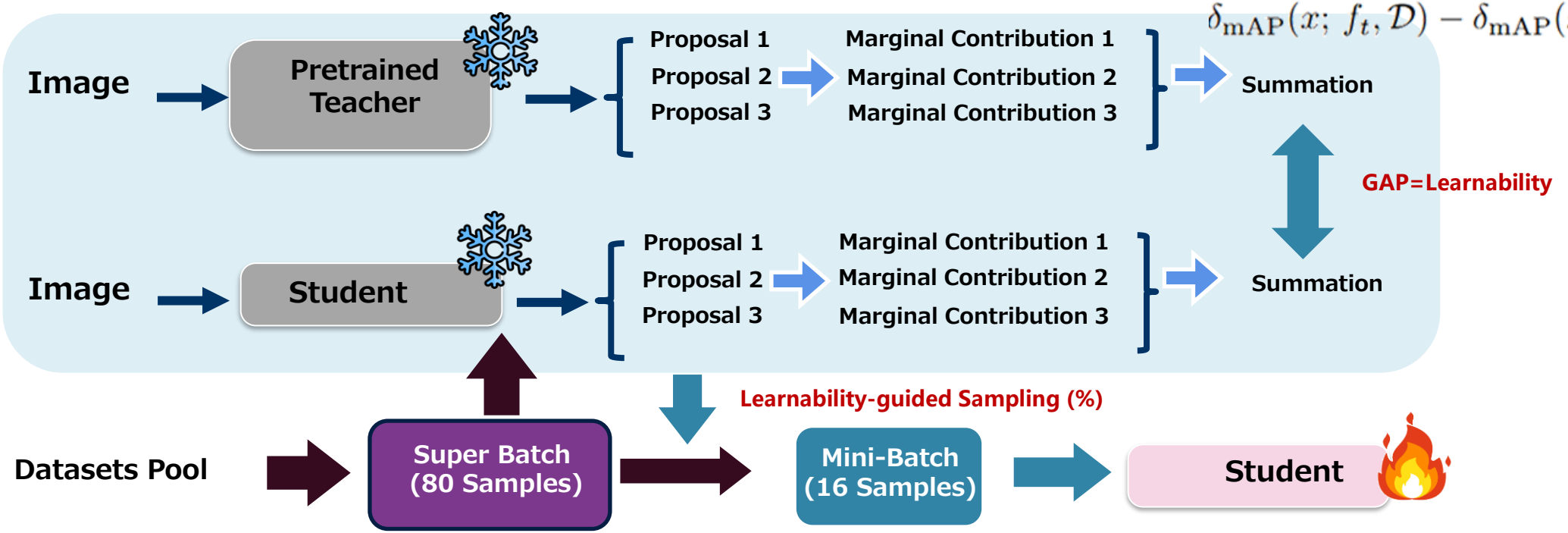
Higher score/ IoU → Higher Influence on mAP

$$\delta_{\mathrm{mAP}}(x;\, f, \mathcal{D}) \triangleq \mathrm{mAP}\big(f;\, \mathcal{D} \cup \{x\}\big) - \mathrm{mAP}\big(f;\, \mathcal{D}\big).$$



● We develop a fast method to estimate the DetGain, *learnability = teacher's DetGain – student's DetGain*

$$\delta_{\mathrm{mAP}}(x;\, f_t, \mathcal{D}) - \delta_{\mathrm{mAP}}(x;\, f_s, \mathcal{D}).$$

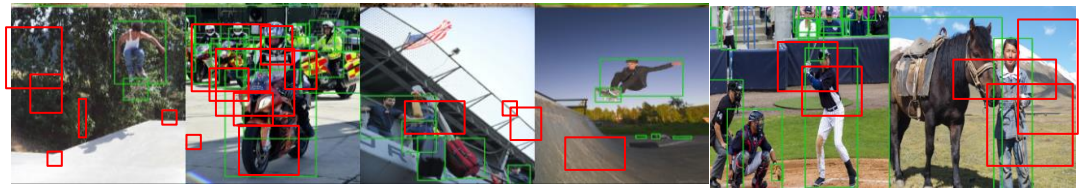


SONY

◆ Method & Implementation

Whole Training Dataset ***D*** **:** 120K COCO 2017

Current student model **f**

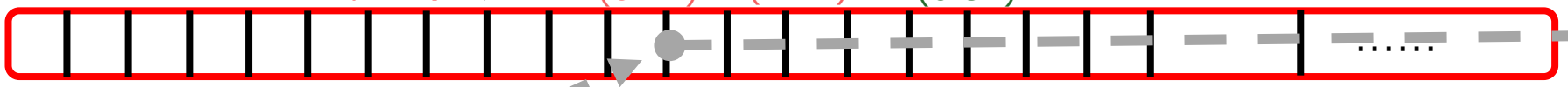


Red: FP
Green TP

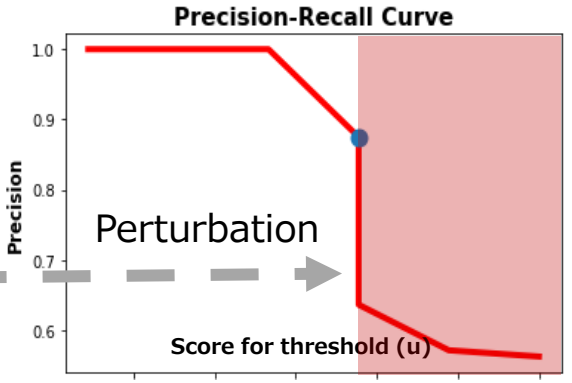
Detections 12,000K (usually 100 per images)

Ranking by confidence score (0,1)

TP(0.99) TP(0.95) FP(0.94) ····· FP(0.44) FP(0.32) TP (0.31) ·····



DetGain does not measure the per-image Average Precision (AP) directly, but instead quantifies how much a single image contributes to the **global mAP** of the entire dataset



Inserting One TP detection x with score = 0.6

DetGain: $\delta_{\mathrm{mAP}}(x;\, f, \mathcal{D}) \triangleq \mathrm{mAP}(f;\, \mathcal{D} \cup \{x\}) - \mathrm{mAP}(f;\, \mathcal{D}).$

learnability: $\delta_{\mathrm{mAP}}(x;\, f_t, \mathcal{D}) - \delta_{\mathrm{mAP}}(x;\, f_s, \mathcal{D}).$

$$\mathrm{AP} = \int_0^1 p(u)\, dr(u)$$

P(u)= $\dfrac{TP}{TP + FP}$

◆ Experimental Result

Result on Five Different Models

COCO 2017 val Benchmark

Student Backbone: Res50

Teacher Backbone: Res101/Res152

Model	Type	Learning Proxy	Baseline		+Data Aug.		+DetGain	
			AP50	AP50:95	AP50	AP50:95	AP50	AP50:95
Faster R-CNN [36]	2-stage, anchor-based	CE + SmoothL1	58.3	37.4	58.3 (+0.0)	37.5 (+0.1)	**61.0** (+2.7)	**40.0** (+2.6)
ATSS [54]	1-stage, anchor-based	Focal + GIoU	57.3	39.4	56.0 (-1.3)	38.6 (-0.8)	**59.8** (+2.5)	**41.6** (+2.2)
FCOS [41]	1-stage, anchor-free	Focal + IoU + BC	56.8	38.2	56.6 (-0.2)	38.2 (+0.0)	**59.8** (+3.0)	**40.9** (+2.7)
GFL [24]	1-stage, anchor-based	QFL + DFL	58.3	40.2	57.9 (-0.4)	40.3 (+0.1)	**59.9** (+1.6)	**42.0** (+1.8)
VFNet [52]	1-stage, anchor-based	Varifocal + GIoU	44.3	40.7	44.3 (+0.0)	40.3 (-0.4)	**46.6** (+2.3)	**42.9** (+2.2)
Deform. DETR [56]	Transformer, anchor-free	Focal + L1 + GIoU	65.7	46.6	66.4 (+0.7)	47.4 (+0.8)	**68.1** (+2.4)	**48.9** (+2.3)

DetGain can improve majority of representative detectors with around +2.0 mAP gain

Dataset	Backbone	Baseline	+DetGain	Δ
VOC2007	R50	51.3	54.3	+3.0
BDD100K	R50	30.3	32.1	+1.8

- **Note: We only change the data pipeline, no changes on model/ training loss/scheduler**

SONY

Conclusion

- Online data-curation technique for general object-detection models.
- Based on each image’ s marginal contribution to the dataset level mAP.
- The result is faster convergence, higher final accuracy, and strong robustness when labels are noisy.

Preprint version:
https://arxiv.org/pdf/2511.14197