This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Online Data Curation for Object Detection via Marginal Contributions to Dataset-level Average Precision

Zitang Sun Masakazu Yoshimura Junji Otsuka Atsushi Irie Takeshi Ohashi
Sony Group Corporation
{Zitang.sun, Masakazu.Yoshimura, Junji.Otsuka, Atsushi.Irie, Takeshi.a.Ohashi}@sony.com

Abstract

High-quality data has become a primary driver of progress under scale laws, with curated datasets often outperforming much larger unfiltered ones at lower cost. Online data curation extends this idea by dynamically selecting training samples based on the model's evolving state. While effective in classification and multimodal learning, existing online sampling strategies rarely extend to object detection because of its structural complexity and domain gaps. We introduce DetGain, an online data curation method specifically for object detection that estimates the marginal perturbation of each image to dataset-level Average Precision (AP) based on its prediction quality. By modeling global score distributions, DetGain efficiently estimates the global AP change and computes teacher-student contribution gaps to select informative samples at each iteration. The method is architecture-agnostic and minimally intrusive, enabling straightforward integration into diverse object detection architectures. Experiments on the COCO dataset with multiple representative detectors show consistent improvements in accuracy. DetGain also demonstrates strong robustness under low-quality data and can be effectively combined with knowledge distillation techniques to further enhance performance, highlighting its potential as a general and complementary strategy for data-efficient object detection.

1. Introduction

High-quality data have increasingly become a primary driver of performance in the scale-law era. Across language [16, 30], vision [38], and multimodal modeling [1, 13, 47], training on well-curated datasets consistently matches or even exceeds the accuracy of much larger unfiltered datasets, while requiring substantially less compute. Online data curation takes this a step further by dynamically deciding, at training time, which samples to learn next [33]. Unlike classic offline dataset pruning or active learning, an online curator adapts to the model's evolving state, steering optimization toward timely, informative, and non-redundant samples [36]. Recent policies in classification and multimodal contrastive learning implement this idea by ranking samples using a *learnability* signal computed from a teacher and a student, typically the loss gap between a pretrained teacher and the current student [14, 36]. Intuitively, samples that the teacher handles well (low teacher loss, high prediction quality) but the student struggles with (high student loss, lower prediction quality) represent rich residual knowledge to be learned, and thus are promising training targets.

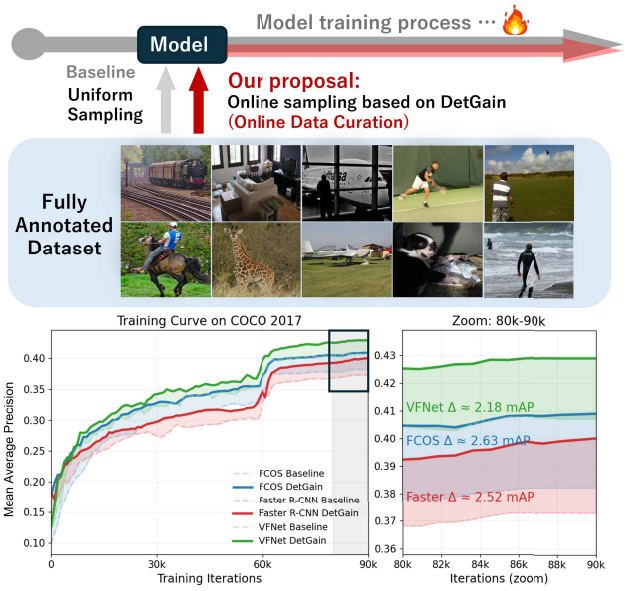


Figure 1. Illustration of our online data curation for object detection, which selects the most informative samples at each training iteration to boost model training performance.

Despite these advances, online data selection for object detection remains underexplored, and few studies have demonstrated tangible performance gains. This gap persists due to two fundamental challenges. First, defining a consistent score for each image in object detection is inherently complex: a single image may contain no instance or multiple instances, some informative and others noisy or ambigu-

ous [25, 53]. Second, the optimization strategy used in detection destabilizes the loss, which is widely used as a component for online data curation [13, 23, 33, 36, 47]. Detection losses are fragmented across tasks (e.g., classification [29, 40], localization [54], centerness [45], etc.) and are affected by stochastic proposal sampling and assignment rules. Modules such as RPN sampling [40] or Hungarian matching [7, 55, 59] dynamically determine which and how many proposals contribute to the loss. These factors cause loss values to fluctuate and drift across iterations, architectures, and even within the same image; thus, loss-based signals such as "learnability" are unreliable for reflecting the residual knowledge to be learned, making direct adaptation of generic data selection metrics prone to domain gaps.

Our key idea is to score learnability using a metric-aware signal aligned with global (dataset-level) Average Precision (AP) rather than raw loss values. We introduce ***DetGain*** (Detection Gain), an image-level estimate of each sample's marginal contribution to the global mAP. For any detector, whether one-stage, two-stage, or transformer-based, the forward pass outputs predicted bounding boxes with confidence scores, labels, and IoUs with ground-truth. By computing DetGain post hoc from these outputs, we avoid reliance on architecture-specific internals and estimate how adding an image perturbs the dataset-level precision–recall curve. However, since mAP is a ranking-based and non-continuous metric, exact per-iteration marginal updates are computationally expensive and noisy. To make DetGain practical for online use, we introduce a fast parametric estimator that models true- and false-positive (TP/FP) score distributions (e.g., two-parameter families) and derives an analytic closed form for efficient computation. Learnability is then defined as the DetGain gap between the teacher and the student for the same image: when the teacher's predicted AP contribution exceeds that of the student, the sample is considered informative and prioritized for training, as it still contains residual knowledge to be learned. The method is fully plug-and-play, modifying only the data pipeline without changing model architectures, loss functions, optimizers, or training schedules.

We evaluate DetGain on several representative detectors, including Faster R-CNN [40], ATSS [56], FCOS [45], VFNet [54], GFL [27], and Deformable DETR [59], on COCO 2017 benchmark [28]. Acting at a meta-level, DetGain adapts to diverse loss functions and architectures, consistently improving validation performance by up to +2.7 mAP under standard schedules. Moreover, on low-quality datasets with noisy or pseudo labels, DetGain yields gains of up to +6.9 mAP. By sharing the teacher, it can also be combined with knowledge distillation (KD) methods [5, 48], further boosting lightweight models. Existing work has rarely explored image-wise online data sampling that produces measurable gains on standard detection benchmarks. DetGain fills this gap by providing extensive comparisons against alternative online metrics. Our results demonstrate that selecting the most informative samples during training consistently improves both detection accuracy and efficiency (i.e., convergence speed), highlighting the strong potential of DetGain for broader applications in data-efficient learning.

2. Related Works

2.1. Online Data Curation

Online data curation ranks and selects training examples on the fly based on the model's evolving state [33, 36, 43, 49]. Earlier work focuses on hard-example mining, where samples are chosen by high loss [23, 33] or prediction uncertainty [12, 26]. Other variants target examples that are easily forgotten during training [46] or reweight losses in a curriculum-learning manner [8, 20]. RAIS [21] introduces importance sampling using normalized gradient magnitudes, while Co-Teaching [17] trains peer networks to mitigate label noise. Xinyue et al. [18] propose a progressive curriculum that drops data portions over time. Recently, RHO-LOSS [36] formalizes "learnability" via a teacher–student signal emphasizing non-noisy and non-redundant samples; JEST [13] extends this to batch-level joint selection for contrastive learning, and ACED [47] interprets this batch selection as implicit KD methods.

Despite these advances, few studies have explored data selection metrics tailored to object detection or shown measurable gains through online sampling. This gap stems from the complexity of detection models, which involve multiple interdependent objectives and image-level sample definitions with variable instance counts [31]. Such properties make generic learnability metrics unreliable because teacher–student loss discrepancies often fail to reflect prediction quality in detection frameworks. *DetGain* instead offers a post-hoc, model-agnostic *image* score aligned with dataset-level mAP, ranking images by a teacher-student gap. Some existing detectors also highlight the importance of online sampling at the proposal/anchor level [6, 32, 44], e.g., assignment or reweighting schemes such as ATSS and Focal Loss [29, 56]. However, most of them are architecture-coupled and operate on per-RoI/per-anchor gradients rather than providing an *image-level* selection metric, so their applicability to image-wise online curation is limited. We instead decouple the data flow from model internals, avoiding unstable loss signals and architectural idiosyncrasies, and enabling transfer across detector families. Moreover, our method operates within the data pipeline, thus complementing existing internal reweighting or sampling approaches; in our experiments, combining *DetGain* with ATSS or Focal Loss yields consistent additional gains.

2.2. Active Learning and Offline Coreset Selection

Early active learning (AL) for object detection mainly relied on uncertainty-based querying, such as entropy or margin scores from detector outputs [11, 41]. Later works highlighted the importance of localization quality [22] and combined instance-level cues: CDAL [2] modeled contextual diversity among co-occurring classes, MI-AOD [53] estimated instance-wise uncertainty via multiple instance learning, and probabilistic models captured both aleatoric and epistemic uncertainty [11]. Recent approaches integrate uncertainty and diversity within iterative query loops—Entropy-based AL [50] combines entropy-guided NMS with prototype diversity, while PPAL [51] ranks samples by calibrated uncertainty and clusters. Semi-supervised methods such as Active Teacher [35] further couple querying with pseudo-labeling for efficient learning.

In parallel, coreset and dataset selection methods aim to construct compact, representative subsets offline [3, 24, 25, 37, 42, 58]. Most AL and coreset pipelines remain offline or semi-offline and recursive, and the main goal is label efficiency—approaching fully supervised accuracy with minimal human annotations. However, such strategies cannot be directly applied when the dataset is already fully labeled. In contrast, DetGain performs online, real-time batch selection during training using fully labeled data. Instead of handcrafted mixtures of uncertainty or diversity, it uses a teacher–student signal, a dynamic, AP-oriented contribution gap, to identify batches that maximize each image's marginal contribution to global AP. Recent study [58] applies AP-based metrics for offline data selection. DetGain generalizes this concept by dynamically estimating global AP contributions rather than local image-level AP.

3. Methods

3.1. Batch-selection Criteria

Previous Learnability definition: We denote the model under training as the *student*. At each iteration, a "super-batch" $\mathcal{S}$ of B samples is first loaded, from which we wish to extract a smaller, more informative *sub-batch* $\mathcal{B} = \{x_i\}_{i=1}^{b} \subset \mathcal{S}$ to perform gradient updates. To guide this selection, learnability-based [13, 14, 36, 47] scoring functions leverage loss signals from both the *student* model $f_s(x;\theta)$ and a pretrained *teacher* model $f_t(x;\theta^*)$, defined as: $s_{\text{learn}} = \ell(\mathcal{B} \mid f_s) - \ell(\mathcal{B} \mid f_t)$, where $\ell(\mathcal{B} \mid f)$ denotes the mean loss of model f over batch $\mathcal{B}$. Intuitively, samples with high student loss but low teacher loss are prioritized, as they indicate regions where the student still lags behind an optimized reference model. The first term, $\ell(\mathcal{B} \mid f_s)$, reflects sample difficulty relative to the learner's current state. Conversely, the second term, $\ell(\mathcal{B} \mid f_t)$, favors high-quality examples but ignores the student's current progress. By combining the two, s_{learn} captures the *improvement margin*, namely how much a sample could help bridge the gap between the student and teacher. The ratio between sub-batch and super-batch sizes defines the selection ratio $k = b/B$. Smaller k values lead to more selective sampling but increase the computational cost of scoring, as additional forward passes on the super-batch are required.

DetGain-based learnability: Loss-based learnability is unstable for object detection because the training loss mixes heterogeneous terms (classification, box regression, etc) with architecture- and scale-dependent weights, and it is not directly aligned with the evaluation metric. We therefore use a *metric-driven* formulation that is detector-agnostic and tied to dataset-level AP.

Let $\mathcal{D}$ be the current dataset (or pool) and $f(\cdot)$ a detector. Let $\mathcal{C}$ be the set of classes and $\mathcal{T}$ the set of IoU thresholds used for evaluation (e.g., COCO uses $\mathcal{T} = \{0.50, 0.55, \ldots, 0.95\}$). The dataset-level metric is

$$\text{mAP}(f;\, \mathcal{D}) = \frac{1}{|\mathcal{C}|\,|\mathcal{T}|} \sum_{c \in \mathcal{C}} \sum_{\tau \in \mathcal{T}} \text{AP}_{c,\tau}\big(f;\, \mathcal{D}\big). \tag{1}$$

For a candidate image $x \notin \mathcal{D}$, the *DetGain* of x under model f is the marginal perturbation of dataset-level mAP:

$$\delta_{\text{mAP}}(x;\, f, \mathcal{D}) \triangleq \text{mAP}\big(f;\, \mathcal{D} \cup \{x\}\big) - \text{mAP}\big(f;\, \mathcal{D}\big). \tag{2}$$

This quantity measures how the true/false positives from x, together with their ranks, deform the global precision–recall curve when merged with $\mathcal{D}$. Let f_s denote the student and f_t a fixed, pretrained teacher. We redefine a learnability score that quantifies the remaining improvement margin on x:

$$s_{\text{DG}}(x) = \delta_{\text{mAP}}(x;\, f_t, \mathcal{D}) - \delta_{\text{mAP}}(x;\, f_s, \mathcal{D}). \tag{3}$$

Large values indicate images where the teacher contributes more to global mAP than the student, i.e., where the student still lags in *dataset-level* utility. Given a super-batch, we score each image with (3) and select the top-b to form the sub-batch $\mathcal{B}$. For efficiency, we use the additive approximation $\delta_{\text{mAP}}(\mathcal{B};\, f, \mathcal{D}) \approx \sum_{x \in \mathcal{B}} \delta_{\text{mAP}}(x;\, f, \mathcal{D})$, which is accurate when $b \ll \mathcal{D}$ and enables per-image, online scoring. Our overall pipeline is shown in Fig. 2 A.

Directly maximizing $\text{mAP}(f; \mathcal{D})$ during training is notoriously difficult due to the non-smooth and non-decomposable nature of AP and the heavy computational cost across the whole dataset [9, 39]. DetGain preserves the metric target but relocates it to the data selection step, avoiding gradient issues while aligning sampling with evaluation. In the next subsection we introduce a fast estimator for Eq. (2) that supports real-time scoring.

3.2. Fast Calculation of DetGain

We aim to approximate the marginal change of dataset-level mAP contributed by inserting the detections from one image without rerunning the full evaluator over the dataset.

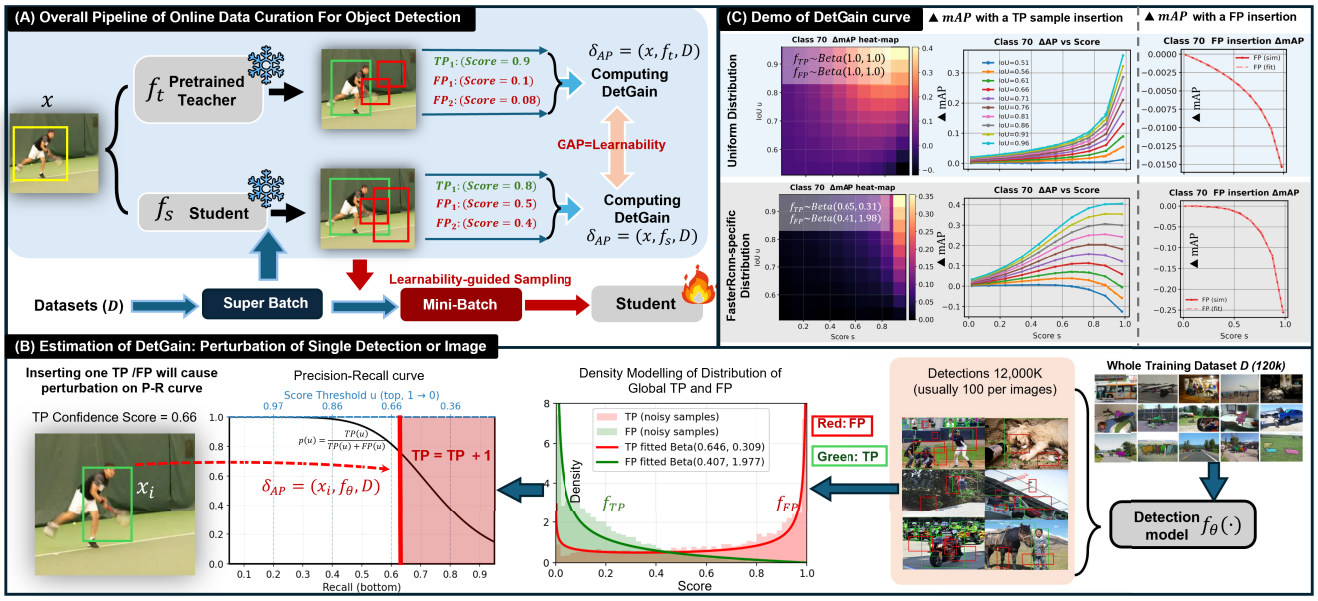


Figure 2. **Overview of the proposed online data curation framework for object detection.** (A) *Overall pipeline:* a pretrained teacher and a student model compute their respective marginal mAP contributions (δ_{AP}) for each image. The difference between them defines the *learnability* score, which guides mini-batch sampling during training. (B) *Estimation of DetGain:* inserting a single TP or FP perturbs the precision–recall curve. We model global TP/FP score distributions and analytically estimate each detection's contribution to mAP under these densities. (C) *Demonstration of DetGain behavior.* Left: ΔmAP when inserting a class-specific TP detection—determined jointly by the confidence score and bounding-box quality (IoU). Right: ΔmAP when inserting a class-specific FP detection—determined primarily by the prediction score. The bottom row uses the true TP/FP score distributions measured from Faster R-CNN predictions; recomputing these at every iteration is computationally expensive. We adopt the top row (uniform distribution) for a model-agnostic faster computation.

Each detection (predicted bbox) from one image can be viewed as a small perturbation to the global counts of true and false positives (T, F) already accumulated from $\mathcal{D}$. The AP integrates precision along the PR curve as the detection score threshold decreases, making it rank-dependent. To approximate dataset-level Δ mAP, we therefore model the score distribution of TPs and FPs in a continuous form.

Formulation in the score–threshold domain: Let $u \in (0, 1)$ be the confidence threshold (scanned from high to low). Let T_{GT} denote the total number of ground truths, and (T, F) the current numbers of true and false positives. Let $F_{\mathrm{TP}}(u), F_{\mathrm{FP}}(u)$ be the cumulative distribution functions (CDFs) and $f_{\mathrm{TP}}(u), f_{\mathrm{FP}}(u)$ their probability densities (PDFs). The cumulative TP/FP counts above threshold u are:

$$C_{\mathrm{TP}}(u) = T\left[1 - F_{\mathrm{TP}}(u)\right], C_{\mathrm{FP}}(u) = F\left[1 - F_{\mathrm{FP}}(u)\right], \quad (4)$$

and the total predictions above u are $N(u) = C_{\mathrm{TP}}(u) + C_{\mathrm{FP}}(u)$. Precision and recall at threshold u are

$$p(u) = \frac{C_{\mathrm{TP}}(u)}{N(u)}, r(u) = \frac{C_{\mathrm{TP}}(u)}{T_{\mathrm{GT}}}. \quad (5)$$

Hence, the (non-interpolated) AP can be written as the Stieltjes integral:

$$\mathrm{AP} = \int_0^1 p(u)\, dr(u) = -\frac{T}{T_{\mathrm{GT}}} \int_0^1 p(u)\, f_{\mathrm{TP}}(u)\, du. \quad (6)$$

Marginal Δ mAP for inserting a detection. Consider adding one detection with score s. There are two cases.

(i) when Insert a TP at score s. For $u \le s$, $C'_{\mathrm{TP}}(u) = C_{\mathrm{TP}}(u) + 1$ and $N'(u) = N(u) + 1$, while $r(u)$ gains a point mass $1/T_{\mathrm{GT}}$ at $u = s$. Using Eq. (6) one obtains

$$\delta^{\mathrm{TP}}_{\mathrm{AP}}(s) = \frac{1}{T_{\mathrm{GT}}} \frac{C_{\mathrm{TP}}(s) + 1}{N(s) + 1} + \frac{T}{T_{\mathrm{GT}}} \int_0^s \frac{C_{\mathrm{FP}}(u)}{N(u)\,[N(u) + 1]}\, f_{\mathrm{TP}}(u)\, du. \quad (7)$$

The first line is the self-contribution at the insertion point; the second line accounts for precision changes on all existing TPs with scores $\ge s$.

(ii) when Insert an FP at score s. Recall does not change; for $u \le s$ we have $C'_{\mathrm{FP}}(u) = C_{\mathrm{FP}}(u) + 1$ and $N'(u) = N(u) + 1$. Thus

$$\delta^{\mathrm{FP}}_{\mathrm{AP}}(s) = -\frac{T}{T_{\mathrm{GT}}} \int_0^s \frac{C_{\mathrm{TP}}(u)}{N(u)\,[N(u) + 1]}\, f_{\mathrm{TP}}(u)\, du. \quad (8)$$

The computational pipeline is illustrated by Fig 2 (B).

Image-level aggregation across classes and IoU thresholds: For a single image x, under a first-order additive assumption, its marginal mAP change is approximated by summing the per-detection effects after standard one-to-one matching at each IoU threshold τ and then averaging over classes and thresholds. Generally, COCO-style mAP

averages AP over object classes $\mathcal{C}$ and IoU thresholds $\mathcal{T}$ (Eq. 1). Accordingly, we account for both the IoU- and class-wise effects in $\delta_{\mathrm{mAP}}(x)$. Let $T_{c,\tau}$ and $F_{c,\tau}$ be the dataset-level totals of TPs and FPs used to compute $\mathrm{AP}_{c,\tau}$, and let T_c^{GT} be the total number of ground-truth instances of class c. For detections on x, $\{(s_j, c_j, \mathrm{IoU}_j)\}_{j=1}^{m}$ (score s_j, predicted class c_j, IoU to the best-matched ground truth), define $y_{j,\tau} \in \{0,1\}$ as the indicator that detection j is a TP at threshold τ.

Given class/IoU-specific score densities $f_{\mathrm{TP}}^{(c,\tau)}(u)$ and $f_{\mathrm{FP}}^{(c,\tau)}(u)$ (used inside Eqs. 7–8), the image-level marginal is approximated by

$$\delta_{\mathrm{mAP}}(x) \approx \frac{1}{|\mathcal{C}|\,|\mathcal{T}|} \sum_{c\in\mathcal{C}} \sum_{\tau\in\mathcal{T}} \sum_{j:\, c_j=c} \Big[y_{j,\tau}\, \delta_{\mathrm{AP}}^{\mathrm{TP},(c,\tau)}(s_j) + \big(1 - y_{j,\tau}\big)\, \delta_{\mathrm{AP}}^{\mathrm{FP},(c,\tau)}(s_j) \Big], \tag{9}$$

where $\delta_{\mathrm{AP}}^{\mathrm{TP},(c,\tau)}(\cdot)$ and $\delta_{\mathrm{AP}}^{\mathrm{FP},(c,\tau)}(\cdot)$ are the single-insertion effects from Eqs. 7–8. In practice, we observe that the IoU of matched TP shows an approximately linear relationship with their DetGain. Moreover, since $\delta_{\mathrm{AP}}(\cdot)$ includes the normalization factor $1/T_c^{\mathrm{GT}}$ (Eq. 7–8), a TP or FP from a rarer class induces a larger per-class AP change than one from a frequent class. The outer average over classes in mAP further assigns equal weight to each category, thereby naturally mitigating long-tail imbalance.

Parametric modeling of score and IoU effects: The per-threshold terms in Eq. (7)–(8) depend on TP/FP score densities for each class and IoU threshold. A practical choice is to fit lightweight parametric models from the current student's predictions, for example using a Beta distribution with method-of-moments fitting, i.e., $u \sim \mathrm{Beta}\big(\alpha_{\mathrm{TP}}, \beta_{\mathrm{TP}}\big)$ for TP scores (and analogously for FP), after which the image-level $\Delta\mathrm{mAP}$ is obtained by the one-dimensional numerical integrals in Eq. (7)–(8).

To reduce the burden of online distribution updates and keep computation real-time, we adopt a universal simplification where all TP/FP scores follow $\mathrm{Beta}(1,1)$, i.e., a uniform distribution with $f_{\mathrm{TP}}^{(c,\tau)}(u) \equiv f_{\mathrm{FP}}^{(c,\tau)}(u) \equiv 1$ on $(0,1)$. This eliminates the need for per-model estimation of the prior-distribution and allows the method to be directly applied to any detector without parameter calibration. More importantly, under the above uniform prior, Eq. (7)–(8) admit compact closed forms ($s \in (0,1)$):

$$\delta_{\mathrm{AP}}^{\mathrm{TP}}(s) = \frac{1}{T_c^{\mathrm{GT}}} \left[\frac{T(1-s)+1}{A(1-s)+1} + \frac{TF}{A^2} \ln \frac{A+1}{A(1-s)+1} \right],$$
$$\delta_{\mathrm{AP}}^{\mathrm{FP}}(s) = -\frac{T^2}{T_c^{\mathrm{GT}} A^2} \ln \frac{A+1}{A(1-s)+1}, \tag{10}$$

where $T = T_{c,\tau}$, $F = F_{c,\tau}$, and $A = T + F$. Both functions are monotonic with respect to s—increasing for TPs and decreasing for FPs. In deployment, these terms are integrated into the image-level aggregator (Eq. 9); the analytical solution also enables an $\mathcal{O}(1)$ online cost per detection, i.e., $\mathcal{O}(m)$ per image with m detections.

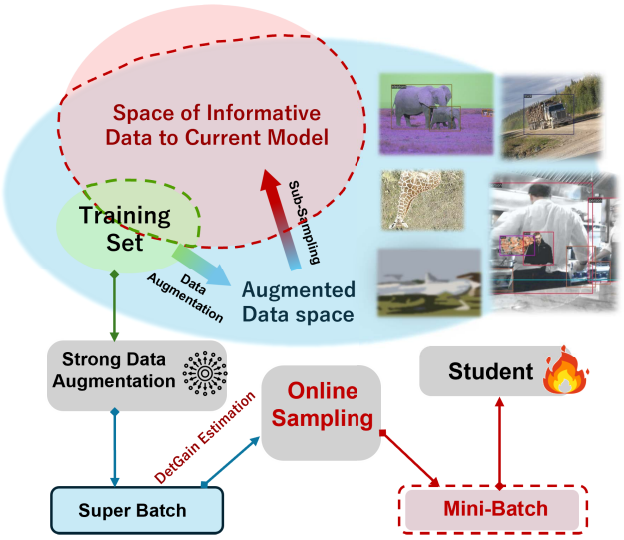


Figure 3. **Joint DetGain–augmentation sampling framework.** Strong augmentation expands the *augmented data space* beyond the original training set. The subsampling space extends from the small overlap between the training set and informative data (green–red) to a larger overlap (blue–red), reducing overfitting and improving data diversity for the student model.

The uniform prior reflects the maximum-entropy distribution over confidence scores, a random-guess baseline. Under this baseline, DetGain quantifies how much the current detector improves over a naive predictor. We adopt this simplification because (i) it admits a closed-form, $\mathcal{O}(1)$ analytic computation of TP/FP contributions; (ii) it avoids brittle, model-specific density fitting; and (iii) in practice we rank images by the teacher–student DetGain gap, so both models are compared against the same neutral baseline. Although the absolute values are approximate, the induced ordering is stable and informative for sample selection. We further visualize the numerical solution of Eq. (9) as DetGain surfaces over IoU and score (Fig. 2 C). We compare two instantiations: (i) the uniform-prior closed forms (Eq. (10)), and (ii) class-/threshold-specific Beta fits estimated from Faster R-CNN. Both exhibit the same trend that higher-score TPs produce larger positive gains while negative for FPs—and empirically they deliver nearly identical downstream performance (see Appendix 6 for details).

3.3. Combine with Online Data Augmentation

Pure online sampling can easily overfit. Intuitively, DetGain-based subsampling can be viewed as a pointer oriented toward a hidden subspace containing the most informative data to the current model. However, repeatedly sampling high-learnability samples may cause the training

Table 1. **Overall results on COCO `val2017`.** We report AP@50 and standard AP (AP@[50:95]) "+Data Aug." applies strong online augmentation. "+ DetGain" denotes combination of strong augmentation with DetGain-based online sampling (subsampling ratio = 20%). All student models use ResNet-50 backbones; teachers are same-family models pretrained on the same training set. (Res152 for CNN-based detectors and Res101 for Deform. DETR). Results are repeated three times with different seeds (mean ± std).

Model	Type	Baseline		+Data Aug.		+DetGain	
		AP@50	AP	AP@50	AP	AP@50	AP
Faster RCNN [40]	2-stage, anchor-based	58.3±.094	37.5±.047	58.3±.125 (+0.0)	37.5±.082 (+0.1)	**61.0±.125** (+2.7)	**40.0±.125** (+2.5)
ATSS [56]	1-stage, anchor-based	57.3±.082	39.2±.094	56.0±.141 (-1.3)	38.6±.125 (-0.8)	**59.8±.094** (+2.5)	**41.5±.047** (+2.3)
FCOS [45]	1-stage, anchor-free	56.7±.125	38.2±.125	56.6±.094 (-0.1)	38.2±.047 (+0.0)	**59.8±.125** (+3.1)	**41.0±.047** (+2.8)
GFL [27]	1-stage, anchor-based	58.3±.094	40.2±.082	57.9±.125 (-0.4)	40.3±.094 (+0.1)	**59.9±.082** (+1.6)	**42.0±.094** (+1.8)
VFNet [54]	1-stage, anchor-based	44.3±.125	40.7±.094	44.3±.082 (+0.0)	40.3±.125 (-0.4)	**46.5±.094** (+2.2)	**42.9±.082** (+2.2)
Def. DETR [59]	DETR., anchor-free	65.8±.141	46.6±.125	66.5±.245 (+0.7)	47.5±.189 (+0.9)	**68.2±.163** (+2.4)	**48.9±.125** (+2.3)

to collapse onto a narrow subspace within the training set. Our goal is to improve the global mAP, but such collapse often leads to a gap between training and validation performance (see Table 2). Previous methods address this by maintaining a hold-out teacher [36] or expanding the dataset for student [13, 47], both requiring extra data or data partitions. We propose a simpler alternative: combining *strong online augmentation* with our image-level DetGain scoring. An augmentation operator $\mathcal{A}(\cdot) \sim \lambda$ (e.g., color jitter, affine transform, noise, copy-paste) is applied to each sample x before DetGain computation to expand the data space before subsampling. This paired design simplifies the RHO-style setup [36]: the teacher is trained on clean (no augmented) data, while the student learns from augmented views without requiring a hold-out split. Their combination significantly enlarges the sampling space and allows the sampler to filter out low-quality augmentations while focusing on informative regions. As shown in Fig. 3, by first transferring samples into the augmented space and then subsampling toward informative regions, the effective sampling area expands from the small overlap between the training set and informative data (green–red) to a larger overlap with the augmented data space (blue–red).

Table 2. **Effect of online sampling and data augmentation on Faster R-CNN (COCO val).** Sampling by learnability alone easily overfits, while strong augmentation alone lacks focus. Combining both restores diversity and achieves the best generalization.

Setting		AP	
Data aug.	Online sampling	Train	Val
✗	✗	44.6	37.4
✗	✓	**50.3** (+5.7)	37.3 (-0.1)
✓	✗	40.4 (-4.2)	37.5 (+0.1)
✓	✓	45.3 (+0.7)	**39.9** (+2.5)

4. Experiments

4.1. Implementation Details

We evaluate on COCO 2017 [28] (118k train / 5k val), a widely used and challenging detection benchmark with a long-tailed class distribution, diverse object scales, and dense annotations. To probe robustness, we also study (i) label-noise scenarios, (ii) a pseudo-labeling regime using the *unlabeled* split, and (iii) Extra validation on PASCAL VOC [15] and BDD100k [52] dataset in appendix. All experiments are implemented in MMDetection [10] (PyTorch) and run on a high-performance computing cluster with NVIDIA A100/H100 GPUs. The effective batch size is 16 for both the baseline and DetGain. For online sampling, each iteration forms an 80-image *super-batch* and backpropagates the top 20% ranked by the teacher–student DetGain gap (i.e., 16 images). Unless otherwise noted, we follow MMDetection defaults: a $1\times$ schedule for CNN-based detectors ($\approx$ 90,000 iterations, SGD) and the default 50-epoch schedule for Deformable DETR ($\approx 3.7\times10^5$ iterations, AdamW [34]). We also evaluate longer schedules and compare with knowledge-distillation baselines in Sec. 4.5. Student models use ResNet-50 [19] backbones by default; teachers use the larger backbones (Res101/Res152) pretrained on the same training set by default. All other settings (data preprocessing, optimizer, and learning policy) follow each detector's default configuration.

4.2. Main Results

As shown in Tab. 1, DetGain-based online curation consistently improves performance across diverse detector architectures and optimization objectives. For most models, strong data augmentation alone brings limited mAP gains, as excessive perturbations often degrade sample quality. However, when combined with our sampling pipeline, the model effectively filters out low-quality augmentations, selects more diverse and learnable samples, and achieves substantial improvements. Overall, across six representative detectors, our method yields an average gain of approximately $\sim +2.0$ mAP. Notably, these improvements

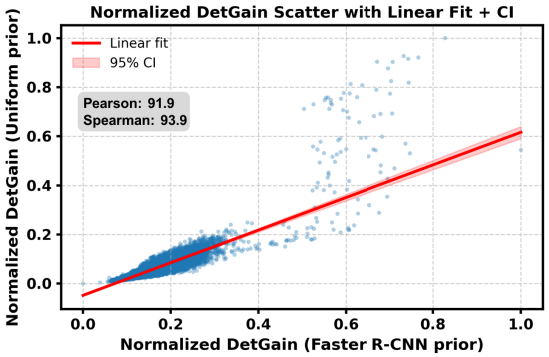


Figure 4. **Uniform prior v.s. model-fit ranking**. Comparison between DetGain value computed from statistics estimated from Faster R-CNN and uniform prior.

Table 3. **Teacher backbone used for DetGain scoring.** All results are mAP on COCO val2017 with ResNet-50 as students' backbone. Larger teacher backbones slightly yield stronger student performance.

Student	Baseline	+DetGain (Teacher Backbone)		
		Res50	Res101	Res152
Faster R-CNN	37.4	39.6 (+2.2)	39.9 (+2.5)	**40.0** (+2.6)
FCOS	38.2	40.2 (+2.0)	40.7 (+2.5)	**40.9** (+2.7)
VFNet	40.7	42.6 (+1.9)	42.7 (+2.0)	**42.9** (+2.2)
ATSS	39.4	40.9 (+1.5)	41.3 (+1.9)	**41.6** (+2.2)

are achieved only by modifying the data sampling strategy without any changes to the model architecture, loss function, or training schedule.

4.3. Ablation Studies

Effect of Data Augmentation: Sampling purely by high DetGain-based learnability narrows training to a small subspace, causing overfitting: training mAP rises sharply while evaluation mAP barely moves. Adding strong online augmentation restores diversity and, when combined with online sampling, yields the best generalization (Table 2). The two are complementary: augmentation expands the data space; online curation sub-samples the most informative instances from this enlarged space (Fig. 3).

Teacher Selections. We examine the effect of teacher capacity. Table 3 shows that larger teacher backbones consistently yield stronger students, echoing [47] that teacher-driven curation functions as implicit knowledge distillation by providing richer supervision on data quality. Accordingly, we use a ResNet-152 teacher with the same architecture family as the student by default to maximize capacity-driven benefits. See Section 4.5 for further discussion of relationship to KD methods.

Ranking of Uniform Prior. To quantify whether the uniform prior changes the *relative ordering* used for selection, we directly compare the per-image DetGain rankings produced by (i) the uniform-prior closed form in E.q. (10) with fixed TP:FP = 1:9, and (ii) a detector-specific Beta-

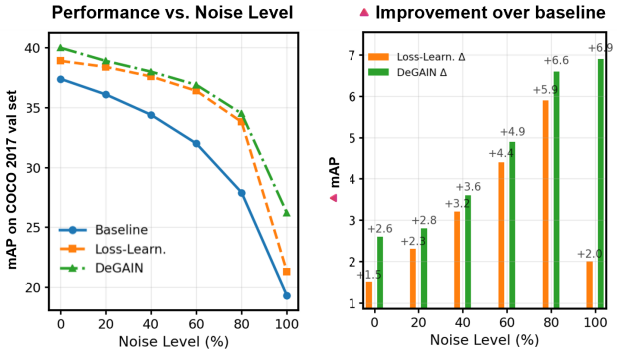


Figure 5. **Robustness under noisy annotations.** Results with *Faster R-CNN–Res50* (stu.) and *Faster R-CNN–Res152* (tea.) on COCO 2017 under controlled annotation-noise ratios. DetGain remains more stable than loss-based learnability baselines [13, 14, 36] and the uniform baseline across varying noise levels.

fit prior with fitted TP/FP score distributions and a fitted TP:FP ratio. We simulate a large number of independently sampled super-batches, matching the typical detection setting: each super-batch contains 64 images, each image produces a fixed budget of 100 post-processed predictions, and the number of ground-truth instances per image is sampled from the same range as in training;

For each super-batch, we compute the Spearman rank correlation ρ between the 64 uniform-prior DetGain values and the 64 Beta+ratio-fit DetGain values, and then aggregate ρ over all trials. Across hundreds of trials, the resulting ranking agreement is consistently high (mean Spearman $\rho \approx 0.94$; Fig. 4), supporting that the uniform prior provides a stable monotonic surrogate for sample selection.

Robustness under Noisy Annotations. We inject annotation noise in the training set by (1) adding fake boxes, (2) removing valid ones, (3) perturbing box locations, and (4) label swaps, controlling the overall noise ratio as the fraction of affected GT boxes. Fig. 5 summarizes the results across noise levels: DetGain outperforms both the uniform baseline and loss-based learnability [13, 14, 36]. This robustness stems from scoring each image by its contribution to global mAP, which promotes cleaner, informative samples while down-weighting corrupted ones. In contrast, loss-based criteria, which directly depend on unstable per-sample losses, are less effective at identifying high-quality data, especially when noise level approaches 100%. See Appendix 2 for extra analysis on pseudo-label scenarios.

4.4. Comparison with Other Methods

Table 4 compares our DetGain-based data curation with various existing sampling metrics on COCO val2017. In addition to loss- and gradient-based indicators commonly used in early studies [21, 33], we include recent loss-learnability approaches [13, 36, 47], as well as entropy-based and hybrid entropy–diversity criteria widely applied

Table 4. **Comparison of online sampling metrics on COCO val 2017.** We report COCO Val AP for three representative students. Each method selects the top 20% samples from a 64-image super-batch with data augmentation. DetGain shows consistent and stable improvements across all architectures, whereas loss- or gradient-based metrics exhibit large variations due to their dependency on optimization design. For a fair comparison, all selection metrics adopt the same online-selection style from the super-batch of each iteration.

	Faster R-CNN – Res50				FCOS – Res50				ATSS – Res50			
Sampling metric	$\mathbf{AP_s}$	$\mathbf{AP_m}$	$\mathbf{AP_l}$	**AP**	$\mathbf{AP_s}$	$\mathbf{AP_m}$	$\mathbf{AP_l}$	**AP**	$\mathbf{AP_s}$	$\mathbf{AP_m}$	$\mathbf{AP_l}$	**AP**
Uniform (baseline)	21.3	40.9	48.5	37.3	22.9	41.8	48.2	38.2	23.3	43.2	49.6	39.4
GradNorm [21]	22.1	41.5	47.7	37.4	23.7	42.0	47.7	38.4	24.2	43.1	49.4	39.3
Loss (hard mining) [33]	23.7	40.3	44.7	36.3	20.9	37.9	43.8	34.5	24.7	41.6	44.6	37.7
Cls. Entropy [35]	19.6	38.4	47.5	34.8	23.1	40.6	43.2	36.6	23.5	40.6	47.7	37.0
Entropy, Info., Diversity [35]	22.6	41.4	47.8	37.5	22.4	40.8	45.9	37.1	22.7	41.7	48.3	38.2
Image-AP [58]	23.3	42.3	49.8	38.3	24.2	43.4	50.7	39.4	24.7	43.6	51.5	40.0
Loss-learnability [13, 36]	23.6	42.5	50.1	38.9	22.0	41.4	49.6	38.1	25.9	44.3	51.5	40.4
DetGain (ours)	**24.4**	**44.1**	**50.8**	**40.0**	**26.0**	**45.0**	**51.1**	**40.9**	**26.1**	**45.5**	**53.1**	**41.6**

Table 5. **Complementarity between DetGain and Knowledge Distillation.** Results on FCOS–Res50 [45]. DetGain achieves performance comparable to modern KD methods [5, 48] and can be combined with them for further improvement. Unlike KD, DetGain requires no architectural or loss modifications and is less sensitive to teacher strength.

Teacher	**DetGain**	**KD**	**AP**
N.A	×	×	38.5
FCOS-Res50	✓	×	40.8 (+2.3)
FCOS-Res101	✓	×	41.1 (+2.6)
FCOS-Res50	×	PKD	38.2 (-0.3)
FCOS-Res101	×	PKD	40.9 (+2.4)
FCOS-Res50	×	CrossKD	39.2 (+0.7)
FCOS-Res101	×	CrossKD	41.1 (+2.6)
FCOS-Res101	✓	PKD	41.7 (+3.2)
FCOS-Res101	✓	**CrossKD**	**42.2** (+3.7)

in active learning for object detection [35, 50]. We also incorporate the image-wise AP metric proposed by [58]. All methods are run in the same online framework with batch-wise sampling and strong augmentation (not offline selection). For image-wise AP, we define a learnability signal as the teacher–student difference in per-image AP. Unlike this discrete, local metric, DetGain estimates each image's continuous marginal contribution to *global* mAP.

Empirically, DetGain delivers higher absolute accuracy and markedly *more stable* gains across diverse detectors and optimization objectives. Loss-/gradient-based strategies (e.g., hard mining, GradNorm) vary with the detector's internal loss scaling and dynamics, and can fluctuate when switching across anchor-based vs. anchor-free or one- vs. two-stage designs. By directly targeting dataset-level mAP, DetGain is architecture-agnostic and remains consistent across loss formulations and model families.

4.5. Discussion with Knowledge Distillation

Tab. 5 further explores the relationship between DetGain-based curation and existing KD methods that utilize the pretrained teacher [4, 5, 48, 57]. With FCOS–Res50 under a 2× schedule (matching the setting of CrossKD), DetGain alone is comparable with state-of-the-art KD methods (e.g., PKD [5], CrossKD [48]). Unlike KD methods, which modify the loss function or introduces additional training heads and feature-alignment modules, our method is *non-intrusive*: it operates purely on the data level without altering the model architecture, training objective, or inference procedure. This simplicity allows DetGain to be easily applied to any detector without code-level modification.

Compared with KD methods, DetGain is less sensitive to the strength of the teacher. Even when the teacher and student share the same architecture (e.g., ResNet-50), DetGain remains effective, whereas KD performance drops significantly with weaker teachers. Importantly, DetGain and KD are *complementary*: KD transfers feature-level knowledge, while DetGain enhances sample-level quality. Combining the two is equal to training the student on the most informative samples while aligning the internal representations.

5. Conclusion and Limitations

For object detection, we propose the first data-efficient online curation pipeline that pairs strong augmentation with DetGain-based subsampling to select informative samples aligned with the model's current state. DetGain is model-agnostic and consistently improves multiple detector families, while remaining robust under noisy annotations, pseudo-labels, and when combined with KD methods. Limitations include added training time from the pre-sampling stage. Moreover, the data augmentation used in this study is relatively naïve and could be further improved. We leave these directions for future investigation toward more efficient and adaptive online training for object detection task.

References

[1] Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S Morcos. Semdedup: Data-efficient learning at web-scale through semantic deduplication. *arXiv preprint arXiv:2303.09540*, 2023. 1

[2] Sharat Agarwal, Himanshu Arora, Saket Anand, and Chetan Arora. Contextual diversity for active learning. In *European Conference on Computer Vision*, pages 137–153. Springer, 2020. 3

[3] Zalán Borsos, Mojmir Mutny, and Andreas Krause. Coresets via bilevel optimization for continual learning and streaming. *Advances in neural information processing systems*, 33: 14879–14890, 2020. 3

[4] Shengcao Cao, Mengtian Li, James Hays, Deva Ramanan, Yu-Xiong Wang, and Liangyan Gui. Learning lightweight object detectors via multi-teacher progressive distillation. In *International Conference on Machine Learning*, pages 3577–3598. PMLR, 2023. 8

[5] Weihan Cao, Yifan Zhang, Jianfei Gao, Anda Cheng, Ke Cheng, and Jian Cheng. Pkd: General distillation framework for object detectors via pearson correlation coefficient. *Advances in Neural Information Processing Systems*, 35: 15394–15406, 2022. 2, 8

[6] Yuhang Cao, Kai Chen, Chen Change Loy, and Dahua Lin. Prime sample attention in object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11583–11591, 2020. 2

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2

[8] Thibault Castells, Philippe Weinzaepfel, and Jerome Revaud. Superloss: A generic loss for robust curriculum learning. *Advances in Neural Information Processing Systems*, 33:4308–4319, 2020. 2

[9] Kean Chen, Jianguo Li, Weiyao Lin, John See, Ji Wang, Lingyu Duan, Zhibo Chen, Changwei He, and Junni Zou. Towards accurate one-stage object detection with ap-loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5119–5127, 2019. 3

[10] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 6

[11] Jiwoong Choi, Ismail Elezi, Hyuk-Jae Lee, Clement Farabet, and Jose M Alvarez. Active learning for deep object detection via probabilistic modeling. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10264–10273, 2021. 3

[12] Cody Coleman, Christopher Yeh, Stephen Mussmann, Baharan Mirzasoleiman, Peter Bailis, Percy Liang, Jure Leskovec, and Matei Zaharia. Selection via proxy: Efficient data selection for deep learning. *arXiv preprint arXiv:1906.11829*, 2019. 2

[13] Talfan Evans, Nikhil Parthasarathy, Hamza Merzić, and Olivier J. Hénaff. Data curation via joint example selection further accelerates multimodal learning. In *Advances in Neural Information Processing Systems*, pages 141240–141260. Curran Associates, Inc., 2024. 1, 2, 3, 6, 7, 8

[14] Talfan Evans, Shreya Pathak, Hamza Merzic, Jonathan Schwarz, Ryutaro Tanno, and Olivier J Henaff. Bad students make great teachers: Active learning accelerates large-scale visual understanding. In *European Conference on Computer Vision*, pages 264–280. Springer, 2024. 1, 3, 7

[15] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. 6

[16] Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*, 2023. 1

[17] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31, 2018. 2

[18] Xinyue Hao, Shihao Hou, Yang Lu, Laura Sevilla-Lara, Anurag Arnab, Shreyank N Gowda, et al. Progressive data dropout: An embarrassingly simple approach to faster training. *arXiv preprint arXiv:2505.22342*, 2025. 2

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[20] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International conference on machine learning*, pages 2304–2313. PMLR, 2018. 2

[21] Tyler B Johnson and Carlos Guestrin. Training deep models faster with robust, approximate importance sampling. *Advances in Neural Information Processing Systems*, 31, 2018. 2, 7, 8

[22] Chieh-Chi Kao, Teng-Yok Lee, Pradeep Sen, and Ming-Yu Liu. Localization-aware active learning for object detection. In *Asian Conference on Computer Vision*, pages 506–522. Springer, 2018. 3

[23] Kenji Kawaguchi and Haihao Lu. Ordered sgd: A new stochastic optimization framework for empirical risk minimization. In *International Conference on Artificial Intelligence and Statistics*, pages 669–679. PMLR, 2020. 2

[24] Krishnateja Killamsetty, Sivasubramanian Durga, Ganesh Ramakrishnan, Abir De, and Rishabh Iyer. Grad-match: Gradient matching based data subset selection for efficient deep model training. In *International Conference on Machine Learning*, pages 5464–5474. PMLR, 2021. 3

[25] Hojun Lee, Suyoung Kim, Junhoo Lee, Jaeyoung Yoo, and Nojun Kwak. Coreset selection for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7682–7691, 2024. 2, 3

[26] Mingkun Li and Ishwar K Sethi. Confidence-based active learning. *IEEE transactions on pattern analysis and machine intelligence*, 28(8):1251–1261, 2006. 2

[27] Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Advances in neural information processing systems*, 33:21002–21012, 2020. 2, 6

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2, 6

[29] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 2

[30] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. 1

[31] Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. Deep learning for generic object detection: A survey. *International journal of computer vision*, 128(2):261–318, 2020. 2

[32] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 2

[33] Ilya Loshchilov and Frank Hutter. Online batch selection for faster training of neural networks. *arXiv preprint arXiv:1511.06343*, 2015. 1, 2, 7, 8

[34] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[35] Peng Mi, Jianghang Lin, Yiyi Zhou, Yunhang Shen, Gen Luo, Xiaoshuai Sun, Liujuan Cao, Rongrong Fu, Qiang Xu, and Rongrong Ji. Active teacher for semi-supervised object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14482–14491, 2022. 3, 8

[36] Sören Mindermann, Jan M Brauner, Muhammed T Razzak, Mrinank Sharma, Andreas Kirsch, Winnie Xu, Benedikt Höltgen, Aidan N Gomez, Adrien Morisot, Sebastian Farquhar, et al. Prioritized training on points that are learnable, worth learning, and not yet learnt. In *International Conference on Machine Learning*, pages 15630–15649. PMLR, 2022. 1, 2, 3, 6, 7, 8

[37] Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning*, pages 6950–6960. PMLR, 2020. 3

[38] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1

[39] Elias Ramzi, Nicolas Thome, Clément Rambour, Nicolas Audebert, and Xavier Bitot. Robust and decomposable average precision for image retrieval. *Advances in Neural Information Processing Systems*, 34:23569–23581, 2021. 3

[40] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016. 2, 6

[41] Soumya Roy, Asim Unmesh, and Vinay P Namboodiri. Deep active learning for object detection. In *BMVC*, page 375, 2018. 3

[42] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017. 3

[43] Vatsal Shah, Xiaoxia Wu, and Sujay Sanghavi. Choosing the sample with lowest loss makes sgd robust. In *International Conference on Artificial Intelligence and Statistics*, pages 2120–2130. PMLR, 2020. 2

[44] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 761–769, 2016. 2

[45] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019. 2, 6, 8

[46] Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. An empirical study of example forgetting during deep neural network learning. *arXiv preprint arXiv:1812.05159*, 2018. 2

[47] Vishaal Udandarao, Nikhil Parthasarathy, Muhammad Ferjad Naeem, Talfan Evans, Samuel Albanie, Federico Tombari, Yongqin Xian, Alessio Tonioni, and Olivier J Hénaff. Active data curation effectively distills large-scale multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14422–14437, 2025. 1, 2, 3, 6, 7

[48] Jiabao Wang, Yuming Chen, Zhaohui Zheng, Xiang Li, Ming-Ming Cheng, and Qibin Hou. Crosskd: Cross-head knowledge distillation for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16520–16530, 2024. 2, 8

[49] Jiachen T Wang, Prateek Mittal, Dawn Song, and Ruoxi Jia. Data shapley in one training run. *arXiv preprint arXiv:2406.11011*, 2024. 2

[50] Jiaxi Wu, Jiaxin Chen, and Di Huang. Entropy-based active learning for object detection with progressive diversity constraint. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9397–9406, 2022. 3, 8

[51] Chenhongyi Yang, Lichao Huang, and Elliot J Crowley. Plug and play active learning for object detection. In *Proceedings*

of the IEEE/CVF conference on computer vision and pattern recognition, pages 17784–17793, 2024. 3

[52] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 6

[53] Tianning Yuan, Fang Wan, Mengying Fu, Jianzhuang Liu, Songcen Xu, Xiangyang Ji, and Qixiang Ye. Multiple instance active learning for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5330–5339, 2021. 2, 3

[54] Haoyang Zhang, Ying Wang, Feras Dayoub, and Niko Sunderhauf. Varifocalnet: An iou-aware dense object detector. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8514–8523, 2021. 2, 6

[55] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[56] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9759–9768, 2020. 2, 6

[57] Zhaohui Zheng, Rongguang Ye, Ping Wang, Dongwei Ren, Wangmeng Zuo, Qibin Hou, and Ming-Ming Cheng. Localization distillation for dense object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9407–9416, 2022. 8

[58] Changyuan Zhou, Yumin Guo, Qinxue Lv, and Ji Yuan. Optimizing object detection via metric-driven training data selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7348–7355, 2024. 3, 8

[59] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021. 2, 6