Task-Specific Zero-shot Quantization-Aware Training for Object Detection

Changhao Li*, Xinrui Chen*, Ji Wang*, Kang Zhao, Jianfei Chen
Georgia Institute of Technology, Tsinghua University
* *Equal Contribution*

ICCV OCT 19-23, 2025 HONOLULU HAWAII

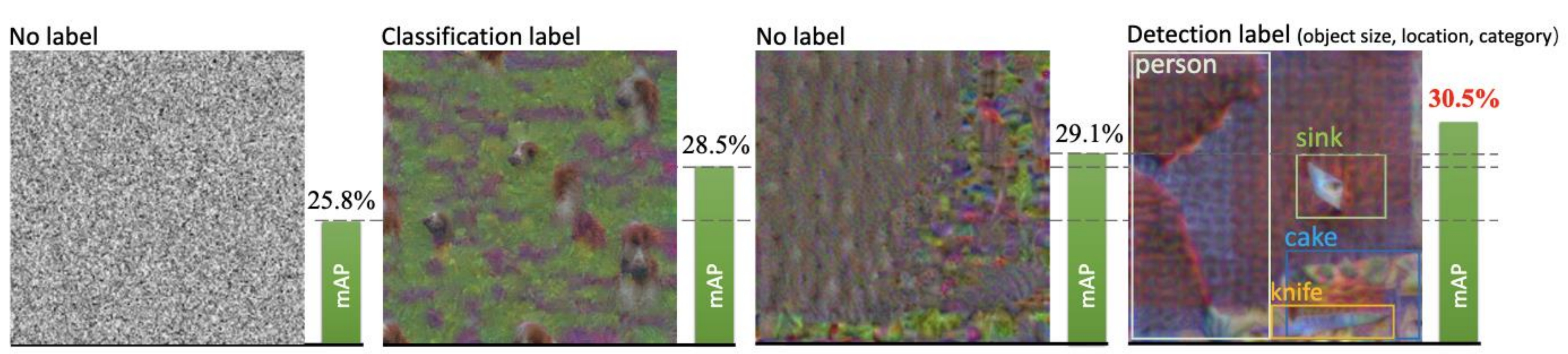
Paper Code Website

1. Background

- **Zero-shot quantization (ZSQ)**
  When training data is inaccessible due to size or privacy constraints, ZSQ enables quantization by inverting the network with randomly sampled labels to generate synthetic data.
- **ZSQ in Object Detection**
  Since object detection targets are inherently difficult, existing methods drop detection loss and use task-agnostic synthetic data, which omits task-specific signals and yields suboptimal results.

2. Motivation

- **Task-specific calibration set matters**



(a) Gaussian Noise (b) Task-specific (classification) (c) Task-agnostic (d) Task-specific (object detection, **Ours**)

Impact of different synthetic images on ZSQ with Mask-RCNN on MS-COCO.

- **Challenges in task-specific ZSQ for detection**

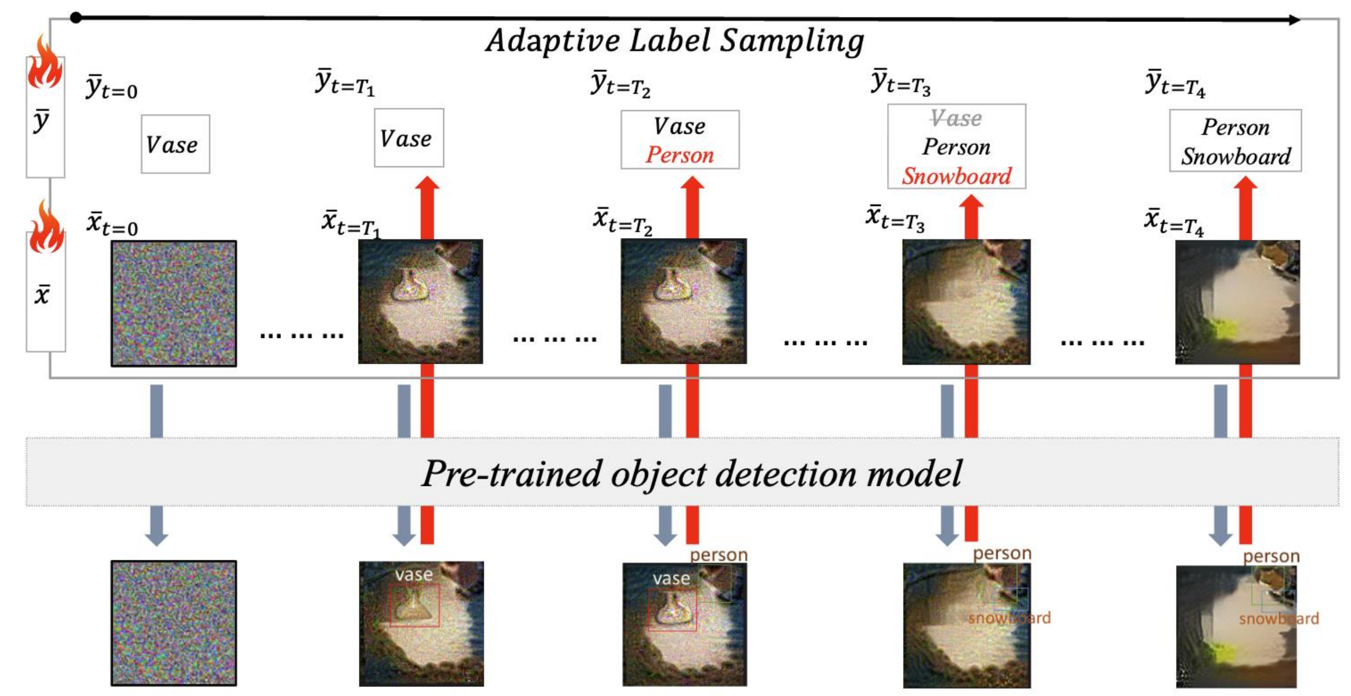
➢ **Label Reconstruction:**
Object locations and sizes remain unknown

➢ **Category Imbalance:**
Random sampling yields unrealistic data

➢ **Underexplored finetuning:**
Logits alignment alone may be insufficient

3. Task-Specific Zero-Shot Quantization-Aware Training

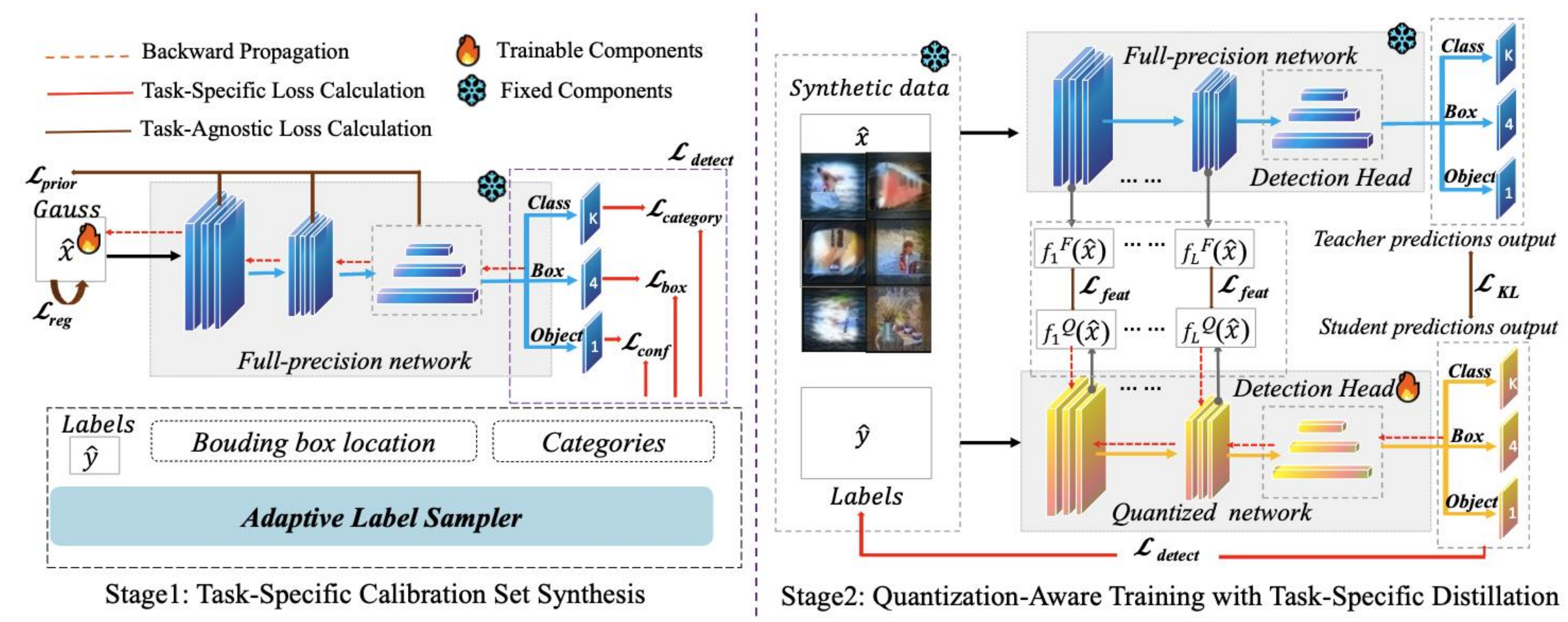
Stage1: Task-Specific Calibration Set Synthesis



- **How to obtain task-specific labels?**
  Adaptive Label Sampling: start with single object labels and Gaussian noise, then progressively aligning image with labels.
- **How to synthesize task-specific labeled data?**

$$\min_{x} \ \alpha_{prior}\mathcal{L}_{prior}(x) + \alpha_{detect}\mathcal{L}_{detect}(\phi(x), \mathrm{y}) + \mathcal{L}_{reg}(x).$$

where $\mathcal{L}_{detect}$ denotes the object detection loss, incorporating task-specific information.



Overall architecture of our framework.

Stage2: QAT with Task-Specific Distillation

- **Prediction-matching Distillation**

$$\min_{\theta'} \mathcal{L}_{KD} = \frac{\tau^2}{N}\sum_{i=1}^{N} KL(z^F(\hat{x}_i;\theta), z^Q(\hat{x}_i;\theta')),$$

- **Feature-level Distillation**

$$\min_{\theta'} \mathcal{L}_{feat} = \frac{1}{NL}\sum_{i=1}^{N}\sum_{l=1}^{L} \| f_l^F(\hat{x}_i;\theta) - f_l^Q(\hat{x}_i;\theta') \|_2^2.$$

- **Task-Specific Quantization-Aware Training**

$$\min_{\theta'} \mathcal{L}^Q = \beta_{KL}\mathcal{L}_{KD} + \beta_{feat}\mathcal{L}_{feat} + \beta_{detect}\mathcal{L}_{detect}$$

4. Experiment

Comparison with real data QATs on MS-COCO validation using different detection architectures. WBAB = weights and activations quantized to B-bit.

Method	Real Data	Num Data	Prec.	mAP / mAP50 YOLOv5-s	YOLOv5-m	YOLOv5-l	YOLO11-s	YOLO11-m	YOLO11-l
Pre-trained	✓	120k(full)	FP	37.4/56.8	45.4/64.1	49.0/67.3	47.0/65.0	51.5/70.0	53.4/72.5
LSQ	✓	120k(full)	W8A8	35.7/54.9	43.2/62.2	46.0/64.9	44.9/61.8	49.1/66.2	50.4/67.4
LSQ+	✓	120k(full)		35.4/54.6	43.3/62.4	46.3/64.9	45.1/61.8	49.6/66.7	50.9/67.7
LSQ	✓	2k		31.6/50.6	36.5/55.6	40.3/59.1	44.0/60.8	47.6/64.5	48.8/65.8
LSQ+	✓	2k		31.5/50.3	36.6/55.8	40.1/58.6	43.8/60.7	47.8/64.7	48.5/65.3
Ours	×	2k		**35.8/55.0**	**43.6/62.3**	**47.3/65.6**	**45.6/62.3**	**50.0/66.5**	**51.8/68.4**
LSQ	✓	120k(full)	W6A6	31.5/49.9	41.3/60.0	43.3/62.1	43.0/59.7	47.4/64.2	48.6/65.3
LSQ+	✓	120k(full)		32.3/50.9	**41.3/60.3**	43.4/62.3	**43.2/59.8**	**47.6/64.3**	**48.9/65.8**
LSQ	✓	2k		28.9/47.2	35.0/53.9	37.7/55.7	41.5/58.3	45.0/61.9	45.8/62.5
LSQ+	✓	2k		28.6/46.7	34.2/52.6	37.5/55.8	41.6/58.2	44.8/61.7	45.9/62.8
Ours	×	2k		**32.7/51.4**	41.0/59.7	**45.1/63.3**	43.0/59.3	47.1/63.2	48.4/64.6
LSQ	✓	120k(full)	W4A8	32.2/51.0	41.0/59.9	44.6/63.5	42.4/59.1	47.6/64.4	48.7/65.6
LSQ+	✓	120k(full)		32.3/51.1	41.2/60.1	44.4/63.2	**42.7/59.3**	**47.8/64.8**	49.4/66.3
LSQ	✓	2k		28.1/46.5	35.8/54.6	39.0/57.5	40.9/57.5	45.2/62.4	46.1/63.0
LSQ+	✓	2k		29.3/47.8	37.8/56.9	40.6/59.7	40.7/57.3	45.2/62.3	46.4/63.4
Ours	×	2k		**33.0/52.5**	**42.6/61.7**	**46.2/64.7**	42.6/58.9	47.7/64.1	**49.4/65.7**

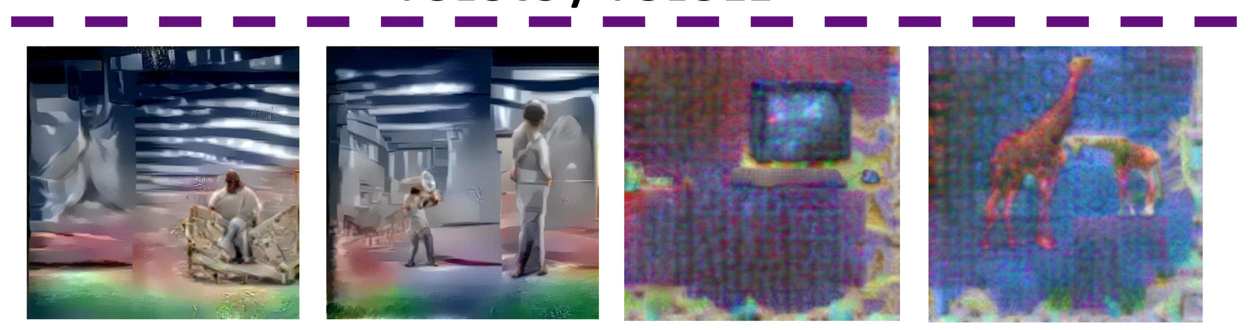
YOLOv5 / YOLO11

Dataset	Method	Real Data	Num Data	Precision	mAP
VOC	Pre-trained	✓	5k(full)	FP32	75.6
	LSQ	✓	5k(full)	W8A8	72.4
	LSQ	✓	50		70.9
	Ours	×	50		**72.9**
MS-COCO	Pre-trained	✓	120k(full)	FP32	38.1
	LSQ	✓	120k(full)	W8A8	35.0
	LSQ	✓	2k		32.9
	Ours	×	2k		**35.2**
	LSQ	✓	120k(full)	W4A8	34.6
	LSQ	✓	2k		32.3
	Ours	×	2k		**34.6**

CNN-based Mask R-CNN

Method	Real Data	Num Data	Prec.	mAP / mAP50 Swin-T	Swin-S
Pre-trained	✓	120k(full)	FP	46.0/68.1	48.5/70.2
LSQ	✓	120k(full)	W8A8	45.9/68.0	48.1/69.7
LSQ	✓	2k		44.4/65.9	47.0/68.6
Ours	×	2k		45.1/66.7	47.1/68.8
LSQ	✓	120k(full)	W6A6	44.7/66.8	47.1/68.8
LSQ	✓	2k		41.2/62.9	44.4/65.9
Ours	×	2k		42.0/63.0	45.1/65.8
LSQ	✓	120k(full)	W4A8	45.5/64.7	47.8/69.4
LSQ	✓	2k		43.3/65.2	45.9/67.3
Ours	×	2k		43.0/64.2	46.2/67.1

Transformer-based Mask R-CNN



Visualization of task-specific calibration data

➢ Left: Adaptive Label Sampling accurately reconstructs object locations.

➢ Right: Its category distribution matches MS-COCO.