ELDET: Early-Learning Distillation with Noisy Labels for Object Detection

Dongmin Choi[1*] **Sangbin Lee**[1,2*] **EungGu Yun**[1] **Jonghyuk Baek**[3†] **Frank C. Park**[1,2]

[1]SAIGE [2]Seoul National University [3]Flitto

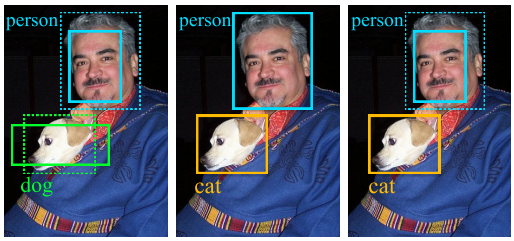
{dm.choi, sb.lee, eg.yun}@saige.ai, jonghyuk.baek@flitto.com, fcp@snu.ac.kr

Abstract

The performance of learning-based object detection algorithms, which attempt to both classify and locate objects within images, is determined largely by the quality of the annotated dataset used for training. Two types of labelling errors are prevalent: objects that are incorrectly classified (*categorization noise*) and inaccurate bounding boxes (*localization noise*); both errors typically occur together in large-scale datasets. In this paper, we propose a distillation-based method to train object detectors that takes into account both categorization and localization noise. The key insight underpinning our method is that the early-learning phenomenon – in which models trained on noisy data with mixed clean and noisy labels tend to first fit to the clean data, and memorize the noisy labels later – manifests earlier for localization noise than for categorization noise. We propose a method that uses models from the early-learning phase (before overfitting to noisy data occurs) as a teacher network. A plug-in module implementation compatible with general object detection architectures is developed, and its performance is validated against the state-of-the-art using PASCAL VOC, MS COCO and VinDr-CXR medical detection datasets.

1 Introduction

Object detection is a fundamental task in computer vision that requires both accurate classification and precise localization of objects within images. Object detection is essential for autonomous driving [12, 54, 53], medical imaging [20, 58, 37] and many other applications that rely on knowing both the type and location of objects in images.



(a) Localization noise (b) Categorization noise (c) Combined noise

Figure 1: Comparison of label noise types in object detection: (a) Localization only. (b) Categorization only. (c) Combined localization and categorization noise.

With few exceptions, methods for object detection are now almost entirely based on neural network models, *e.g.*, [43, 46, 35, 6, 44, 26] trained on large image datasets such as PASCAL VOC [11] and MS COCO [30]. While network architectures and algorithms for data processing and training may still affect object detection performance, by and large the performance of object detectors is determined mostly by the size and quality of the datasets [38].

[*]These authors contributed equally to this work.

[†]This work was primarily conducted while the author was at SAIGE.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

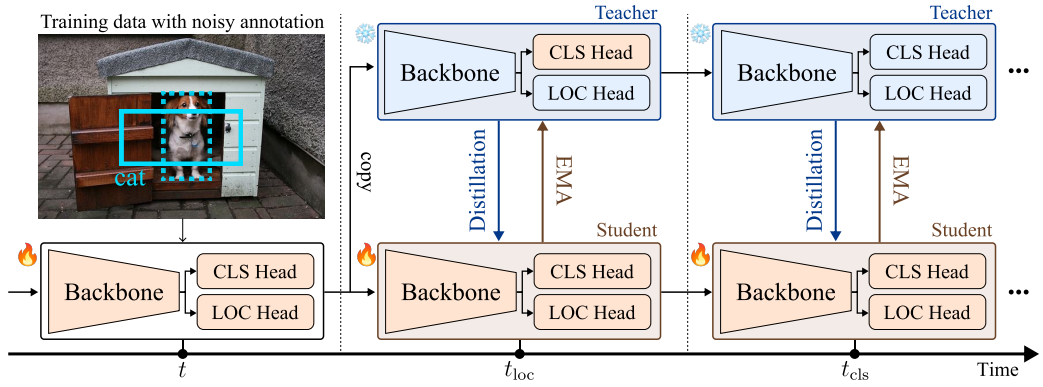


Figure 2: Overview of the ELDET workflow. The early-stage model is initially trained on a dataset with noisy labels until memorization occurs at the t_{loc}-th training iteration. Once the early-learning phase terminates for the localization task, a teacher model (blue), initialized from the current state of the student model (orange), is employed to guide subsequent training through knowledge distillation with its parameters frozen. The teacher model is progressively updated using an exponential moving average (EMA) of the student parameters after each iteration. Concurrently, the classification head is actively updated with a small momentum prior to the memorization phase for classification at iteration t_{cls}.

Acquiring high-quality annotations for large-scale datasets is difficult. Even with recently available methods for auto-labelling and label assist [64, 10], some manual effort is unavoidable, and human labeling errors are often the main source of noisy labels [55, 52]. The two main types of labeling errors are misclassification (*categorization noise*) and inaccurate bounding boxes (*localization noise*). In large-scale, crowd-sourced annotations [13, 63] both errors frequently occur together; for instance, labeling errors of the type shown in Figure 1 are quite common. As observed in [31, 3], the performance of object detection models that are overfit to such incorrect labels can be significantly degraded.

For the simpler task of image classification, numerous methods address noisy labels in [32, 34, 23, 25, 2, 27]. In comparison, only a few works address noisy labels in object detection, with the focus typically on a single type of noise (categorization or localization) rather than both simultaneously. For example, Liu *et al.* [36] propose a method to mitigate categorization noise by excluding unreliable samples, while Bär *et al.* [3] introduce a localization label refinement network to correct box errors. These approaches are limited in real-world settings where both types of noise often appear together.

In this paper, we propose a method to train object detectors that takes into account both categorization and localization noise in the annotations. The key insight underpinning our method is the *early-learning phenomenon* [32] – in which models trained on noisy data with mixed clean and false labels have been observed to first fit to the clean data, and that the false labels are memorized later – manifests differently for classification and localization errors. Specifically, we observe that *models tend to memorize localization error earlier than categorization noise*.

Leveraging this observation, we propose an object detection training algorithm based on knowledge distillation [19, 50], in which models from the early-learning phase (before overfitting occurs to noisy data) act as teacher networks. Specifically, once the early-learning phase terminates, we copy the model to create a frozen teacher network, while a student model continues training on the noisy data. To detect the transition from early-learning to memorization, we fit an exponential parametric curve to training metrics and identify the transition point when the curve's gradient significantly decreases [31, 33]. The teacher network is updated by an exponential moving average (EMA) of the student's parameters to train informative features from clean samples while maintaining robust knowledge. We further adjust EMA momentum for the classification and localization heads to handle the distinct early-learning termination points between these tasks effectively. This self-distillation workflow [21] helps the current model focus on learning from clean data while avoiding overfitting to noisy labels. The overview is illustrated in Figure 2.

We develop a practical implementation of our proposed method in the form of a plug-in module, compatible with various object detection architectures including both anchor-based [46, 44] and anchor-free models [49]. Our approach does not require architecture-specific modifications, making it widely applicable. We evaluate our approach on PASCAL VOC and MS COCO with noise

simulations. Additionally, we assess the domain robustness of our framework on the VinDr-CXR medical detection dataset [39], demonstrating its adaptability to specialized domains. Extensive experimental results demonstrate that our method significantly improves detection performance in noisy environments, effectively handling simultaneous classification and localization noise in real-world object detection tasks.

2 Related Works

Object Detection. Object detection has primarily evolved along two directions: anchor-based [35, 65, 43] and anchor-free detectors [22, 66, 6]. Anchor-based detectors such as RetinaNet [46] and Faster R-CNN [44] use pre-defined anchor boxes as reference points to predict objects. In contrast, anchor-free detectors like FCOS [49] remove the dependency on anchor boxes by directly predicting object locations, which results in simpler architectures and reduced computational demands.

These methods can also be categorized into one-stage and two-stage detectors. One-stage detectors (*e.g.,* RetinaNet [46] and Generalized Focal Loss (GFL) [26]) aim to directly predict class probabilities and bounding box coordinates from the entire image in a single shot. Two-stage detectors (*e.g.,* Faster R-CNN [44] and Cascaded R-CNN [5]) first generate region proposals and then refine them for accurate prediction. Given these significant architectural differences, methods that can effectively integrate with both structures are clearly needed [57]. To this aim, our proposed approach is designed as a plug-in module that can be readily incorporated into most existing architectures.

Transition from Early Learning to Memorization Under Noisy Annotations. Liu *et al.* [32] observe the following *early-learning phenomenon* in deep learning models: during the initial stages of training, model gradients are dominated by clean labels, while memorization of noisy labels emerges in later stages. Based on this observation, they propose a regularization technique that leverages the predictions of earlier models to limit the impact of noisy labels.

Building on this foundation, several works have aimed to develop robust models under noisy annotations for image classification [34, 25, 2, 27]. Han *et al.* [15] introduce Co-teaching, in which two networks are trained simultaneously using the small-loss instances selected by its peer network. Li *et al.* [23] propose DivideMix, which models the distribution of losses to separate clean and noisy samples, then applies semi-supervised learning techniques to utilize both sets effectively.

Extending the concept to segmentation, Liu *et al.* [33] introduce an Adaptive Early Learning Correction (ADELE) framework for the weakly-supervised setting. ADELE monitors class-specific transitions from early-learning to memorization, refining noisy labels with pseudo-labels generated from early-phase model predictions for each class. While segmentation and object detection both require localization and classification, ADELE's focus on pixel-wise annotations and class-specific early learning is more suited to segmentation tasks where localization involves fine-grained pixel boundaries. In contrast, object detection involves both the accurate localization of bounding boxes and the classification of entire regions within the box, making it challenging to directly apply ADELE to this domain. Thus, when applying early learning to object detection with noisy labels, what is needed is a specialized approach that addresses the unique challenges of noisy labels.

Object Detection with Noisy Annotations. Some recent works that propose to train robust object detectors under noisy labels include [48, 24, 7]. Liu *et al.* [36] propose an adaptive framework which identifies reliable examples in noisy data by measuring instance-level domain properties and adjusting the training accordingly. While their approach effectively mitigates the impact of domain shifts with noise, it focuses primarily on domain adaptation scenarios and may not fully address the challenges posed by noisy labels in general object detection tasks.

ORSOD [31] introduces a dynamic loss decay mechanism to enhance the robustness of oriented object detection with noisy labels by adaptively reducing the influence of high-loss samples. However, ORSOD primarily addresses categorization noise and does not account for localization noise, which can significantly impact detection accuracy in real-world scenarios where both types of noise are present. In contrast, Bär *et al.* [3] propose a Localization Label Refinement Network (LLRN) to refine the noisy coordinates of bounding boxes. LLRN focuses on correcting localization errors by training a separate network to predict more accurate bounding boxes, which are then used to update the training data. While effective for localization noise, LLRN overlooks categorization noise,

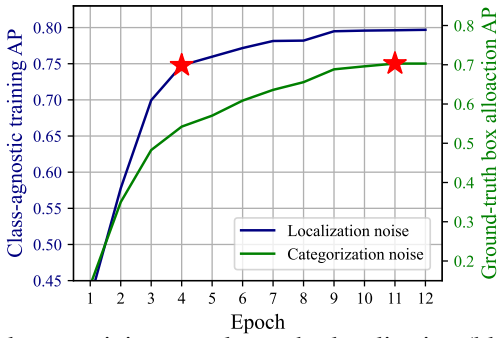


Figure 3: Evaluation results on training samples under localization (blue) or categorization noise (green) with class-agnostic and ground-truth box allocation metric, respectively. Red stars mark early-learning termination points for each metric of RetinaNet on PASCAL VOC with a noise ratio of 40%.

which presents clear limitations in real-world scenarios. In contrast to prior works, our approach simultaneously tackles both types of noise.

Knowledge Distillation. Knowledge distillation has been widely adopted to transfer informative features from a high-performance model (teacher) to an efficient network (student) [19]. Several works have proposed a distillation framework for object detection that improves the performance of small detectors [8, 51, 59]. Wang *et al.* [50] propose an efficient approach named CrossKD, which propagates the intermediate features of the student network to the heads of the teacher models. This cross-head approach enables task-oriented knowledge transfer between networks, improving the student's performance without significantly increasing computational costs.

In self-knowledge distillation [14, 56, 61, 21, 59, 28], a model serves as both teacher and student, gradually refining knowledge by jointly learning from the ground truth and past predictions of the model itself. Works adopting the co-teaching framework [15, 4, 29] demonstrate that self-knowledge distillation can address noisy labels by distilling knowledge from the model itself. While these works have shown promising results, they often suffer from high computational costs due to training multiple models simultaneously.

3 Priority in Memorization: Localization Over Classification

The dual-task nature of object detection—classification and localization—lead us to hypothesize that detectors may prioritize the coupled objectives differently when training under noisy labels. To investigate this phenomenon, we conduct empirical experiments to monitor how object detectors internalize noisy labels across the two tasks.

Specifically, we adopt task-specific metrics to independently monitor detector performance while isolating the impact of noise on each task. First, we evaluate detectors in a class-agnostic (CA) manner, treating all objects as a single category. This evaluation prevents potential interference from classification errors. For categorization noise, we replace the predicted bounding box coordinates with the corresponding ground truth if the overlap between the two exceeds a certain threshold. This ground-truth box allocation (GTBA) effectively alleviates the influence of localization errors, allowing for a more accurate assessment of classification performance. Details are discussed in Section 4.1.

As illustrated in Figure 3, our findings reveal a critical aspect of model behavior under noisy labels for object detection. We observe that *the memorization of localization noise occurs significantly earlier than the memorization of categorization noise*. This finding suggests that the model tends to focus on spatial aspects during the early learning phase, while requiring longer training times for classification. Based on this finding, we propose a technique that suppresses noise memorization across the two tasks.

4 Methods

In this section, we discuss our approach for training robust object detectors in the presence of noisy labels. We exploit the early learning phenomenon [32] to address the challenges posed by both localization and categorization noise.

4.1 Early-Learning and Memorization

As discussed in Section 3, the early learning phenomenon manifests differently in object detection for the two key tasks of localization and classification. Specifically, models tend to memorize localization noise earlier than categorization noise. To accurately capture any task-specific patterns, we employ specialized metrics that independently monitor the training progress of each task.

For localization, we utilize a class-agnostic (CA) metric that treats all detected objects as belonging to a single category (e.g., "object"). This evaluation isolates localization performance without the confounding influence of classification errors, allowing us to monitor how the model learns spatial information over time. For classification, we introduce a Ground Truth Box Allocation (GTBA) approach. Inspired by Moon *et al.* [60], we replace the box coordinates of the model prediction with the ground truth box if the Intersection over Union (IoU) between them exceeds a certain threshold τ. This allocation effectively eliminates the impact of localization errors, enabling a more accurate assessment of classification performance under noisy label conditions.

By employing these task-specific metrics, we can effectively discern the distinct dynamics of early-learning across localization and classification tasks, and determine the appropriate moments to intervene during training.

4.2 Early Learning Phase Detection

To detect the transition from early learning to memorization for each task, we follow Liu et al. [33] in fitting an exponential parametric function to model the rate of change with respect to the performance on noisy training sets over training:

$$f(t) = a\left(1 - e^{-b \cdot t^c}\right), \tag{1}$$

where t represents the training epoch, and $a > 0$, $b \geq 0$, and $c \geq 0$ are parameters that are fitted to the observed data. This parametric form captures the performance trend of the model on the training dataset.

To detect the transition point from early-learning to memorization, we monitor the relative change in the derivative of the metric. Specifically, the transition is identified when the rate of change deviates significantly from the initial learning gradient, as formalized by the following condition:

$$\frac{|f'(1) - f'(t)|}{|f'(1)|} > \gamma, \tag{2}$$

where γ is a threshold capturing the deviation from the initial learning rate. When this condition is met, it indicates that the model dynamics have shifted and it begins to memorize noisy annotations. Using this method, we determine the transition epochs t_{loc} and t_{cls} for localization and classification respectively, based on the metrics in Section 4.1.

4.3 Early Learning Guidance via Distillation

Knowledge Distillation. To mitigate the effects of noisy labels and prevent memorization, we propose an early learning guidance mechanism. This mechanism involves using the model obtained at the end of the early learning phase as a teacher network in a knowledge distillation framework [19, 50, 21]. The teacher network, which has not yet overfitted to noisy labels, guides the current student model during subsequent training, helping it avoid overfitting to noise. The student model continues training on the noisy dataset but receives guidance from the teacher model to focus on learning from clean data. The distillation process is implemented by minimizing the divergence between the predictions of the teacher model and the student model. Specifically, we define the knowledge distillation loss:

$$\mathcal{L}_{\text{kd}} = \mathcal{L}^{\text{kd}}_{\text{cls}} + \mathcal{L}^{\text{kd}}_{\text{loc}}, \tag{3}$$

where $\mathcal{L}_{\text{cls}}^{\text{kd}}$ and $\mathcal{L}_{\text{loc}}^{\text{kd}}$ are the distillation losses for classification and localization, respectively. The total loss function for training the student model is then defined as

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{det}} + \lambda \mathcal{L}_{\text{kd}}, \tag{4}$$

where $\mathcal{L}_{\text{det}}$ is the original detection loss between the student predictions and the noisy ground-truth annotations, and λ is a hyperparameter that balances the detection loss and the distillation loss.

To enhance the efficiency of the distillation process, we adopt the CrossKD framework [50]. CrossKD allows for knowledge transfer between the teacher and student models by propagating the intermediate features of the student network to the teacher's detection heads. This cross-head approach enables task-oriented knowledge transfer without significant computational overhead.

Teacher Model Update via EMA. To ensure that the teacher model remains a reliable source of guidance throughout training, co-teaching frameworks [15, 29] that propagate gradients to both teacher and student networks can be used. However, this dual-training approach incurs high computational costs, as it requires training two models simultaneously. Instead, we update the teacher model using an exponential moving average (EMA) of the student model's weights without updating it directly. The EMA update rule is defined as follows:

$$\theta_{\text{teacher}} \leftarrow \alpha \theta_{\text{teacher}} + (1 - \alpha)\theta_{\text{student}}, \tag{5}$$

where θ_{teacher} and θ_{student} represent the weights of the teacher and student models, respectively, and α is the decay rate. Typically α is set close to 1 to ensure that the teacher model updates slowly, preserving its robustness to noisy labels. However, since the teacher model is initialized after the early-learning phase ends for localization but before it concludes for classification, using a high decay rate on the classification head can potentially hinder its performance. To preclude this possibility, we apply a small decay rate of $\alpha = 0.1$ specifically to the classification head during this period. This enables faster updates for the classification head, which allows classification performance to sufficiently converge before memorization occurs without compromising noise robustness.

Our framework combines EMA updates with knowledge distillation to enable effective teacher guidance and robust training, mitigating noisy label effects in object detection.

5 Experiments

5.1 Experimental Settings

Noise Simulations. To evaluate the robustness of object detectors under noisy labels, we simulate both localization and categorization noise. For localization noise, we perturb the ground truth bounding box coordinate while keeping its category [3]. Given a ground truth box $\mathbf{b} = (x_{\text{min}}, y_{\text{min}}, x_{\text{max}}, y_{\text{max}})$, the noisy box $\mathbf{b}^{\text{noisy}}$ is calculated as:

$$\mathbf{b}^{\text{noisy}} = \begin{bmatrix} x_{\text{min}} + \delta_x & y_{\text{min}} + \delta_y & x_{\text{max}} + \delta'_x & y_{\text{max}} + \delta'_y \end{bmatrix}^\top. \tag{6}$$

where $\delta_x, \delta_y, \delta'_x, \delta'_y$ are sampled from uniform distributions scaled by box size with magnitude controlled by $\epsilon = 0.5$, corresponding to up to 50% perturbation relative to width or height.

For categorization noise, we randomly replace the true class label c_i of each object with an incorrect label c'_i drawn uniformly from all other possible classes $\mathcal{C} \setminus \{c_i\}$:

$$c'_i \sim \mathcal{U}(\mathcal{C} \setminus \{c_i\}). \tag{7}$$

We apply these noise simulations to a random subset of the training data with various levels of 20%, 30%, and 40%.

Datasets. We conduct our experiments on PASCAL VOC [11] and MS COCO, which is a widely used benchmark for object detection. Following the standard protocol [49], we use the VOC 2007 and VOC 2012 `trainval` sets (16,551 images) for training, and perform evaluation on the VOC 2007 `test` set (4,952 images). MS COCO is a large-scale dataset with 80 object categories, featuring over 330K images and more than 2.5 million labeled instances. We use the COCO 2017 version, training on the `train` split (118K images) and evaluating on the `val` split (5K images).

To further validate the robustness of our approach, we run experiments on VinDr-CXR medical object detection dataset [40]. VinDr-CXR dataset consists of 15,000 chest X-ray images annotated with 14 classes, which represent various thoracic disease findings and abnormalities. For all datasets, we apply the aforementioned noise simulations only to the training data while keeping the validation and test sets with clean annotations.

Table 1: Comparison of performance under noise annotations on Pascal VOC reported as AP@50. The results are shown for localization, categorization and combined noise with noise levels at 20%, 30%, and 40% as well as the clean setting. For each detector, the best AP scores are highlighted in bold, with the second-best scores underlined.

Detector	Method	Clean	Localization Noise			Categorization Noise			Combined Noise		
			20%	30%	40%	20%	30%	40%	20%	30%	40%
RetinaNet	-	75.07	74.00	73.83	73.13	70.67	69.03	67.43	70.27	68.07	65.63
	ORSOD [31]	75.41	74.17	73.97	73.43	71.13	68.90	67.33	70.37	66.57	65.47
	ADELE [33]	74.69	73.67	74.10	71.03	71.03	69.23	67.07	70.13	68.20	65.87
	ELDET (ours)	**76.52**	**76.23**	**76.30**	**74.80**	**73.66**	**73.71**	**68.21**	**74.53**	**73.67**	**68.82**
FCOS	-	**72.23**	71.00	70.67	70.60	67.37	64.97	62.63	66.57	63.33	60.13
	ORSOD [31]	71.63	68.99	70.99	70.94	67.55	64.47	62.80	67.36	63.06	60.51
	ADELE [33]	71.59	71.20	70.70	70.57	67.53	65.40	62.67	66.73	63.87	60.70
	ELDET (ours)	72.00	**73.40**	**72.80**	**74.10**	**68.43**	**66.13**	**63.73**	**68.67**	**65.03**	**62.43**
Faster R-CNN	-	**73.89**	69.48	69.49	69.25	66.95	65.15	63.65	66.68	64.84	62.02
	ORSOD [31]	70.77	68.91	68.94	68.88	66.40	64.84	63.40	65.85	64.24	63.01
	ADELE [33]	71.48	69.15	68.54	68.91	67.76	65.65	63.59	68.83	65.28	62.61
	ELDET (ours)	73.34	**71.92**	**71.81**	**70.12**	**69.40**	**67.60**	**66.33**	**69.09**	**66.83**	**64.28**
GFL	-	73.00	73.23	71.68	71.50	69.54	66.18	62.77	67.70	64.69	56.30
	ORSOD [31]	74.11	73.87	73.12	72.51	**69.92**	**67.71**	63.64	68.49	64.73	61.28
	ADELE [33]	73.05	73.82	73.46	72.29	69.14	67.49	64.19	68.41	66.04	62.71
	ELDET (ours)	**75.28**	**75.23**	**74.57**	**74.44**	69.68	67.57	**65.35**	**69.82**	**66.77**	**63.57**

Table 2: Comparison of performance under noise annotations on MS COCO val2017, reported as AP@50. The results are shown for localization, categorization and combined noise with noise levels at 20%, 30%, and 40% as well as the clean setting. The best AP scores are highlighted in bold.

Detector	Method	Clean	Localization Noise			Categorization Noise			Combined Noise		
			20%	30%	40%	20%	30%	40%	20%	30%	40%
RetinaNet	-	44.41	43.97	42.61	43.12	41.86	40.66	39.90	42.61	42.33	41.95
	ELDET (ours)	**45.95**	**44.96**	**44.87**	**44.74**	**43.51**	**41.98**	**40.31**	**44.51**	**43.23**	**42.99**
FCOS	-	44.02	44.34	44.36	43.11	41.91	41.73	39.26	43.16	42.66	42.13
	ELDET (ours)	**45.89**	**45.17**	**44.94**	**44.83**	**43.36**	**42.66**	**41.02**	**44.39**	**43.70**	**43.59**
Faster R-CNN	-	43.55	42.80	42.86	42.48	40.36	39.26	37.21	40.91	40.83	39.57
	ELDET (ours)	**44.79**	**44.00**	**43.51**	**43.49**	**41.42**	**39.47**	**38.20**	**42.72**	**42.06**	**40.30**
GFL	-	47.24	47.19	46.08	45.17	45.32	44.53	43.31	46.30	45.58	44.81
	ELDET (ours)	**49.56**	**48.77**	**47.43**	**46.64**	**47.05**	**45.59**	**44.78**	**47.47**	**46.93**	**45.58**

Baselines. We compare ELDET with two baselines:

- **ORSOD** [31]: ORSOD addresses categorization noise by dynamically excluding samples with high classification loss during training, which prevents the model to memorize noisy samples with large loss.
- **ADELE** [33]: We extend the role of ADELE to object detection under noisy annotations. Specifically, detectors begin to be supervised by the pseudo-label from the early-models instead of raw noisy labels.

5.2 Experimental Results

We use the mean Average Precision at an Intersection over Union (IoU) threshold of 0.5 (AP@50) for PASCAL VOC and COCO, and 0.4 (AP@40) for VinDr-CXR. To validate the compatibility of our method with various architectures, we conducted experiments with different detectors including RetinaNet [46], FCOS [49], Faster R-CNN [44] and GFL [26].

PASCAL VOC. Table 1 presents a quantitative comparison of ELDET against baseline methods on PASCAL VOC. ELDET consistently outperforms the comparison models across all scenarios. Notably, ELDET using the RetinaNet significantly surpasses all existing baselines with a large gap under the combined noise condition with a 40% noise level. This substantial improvement

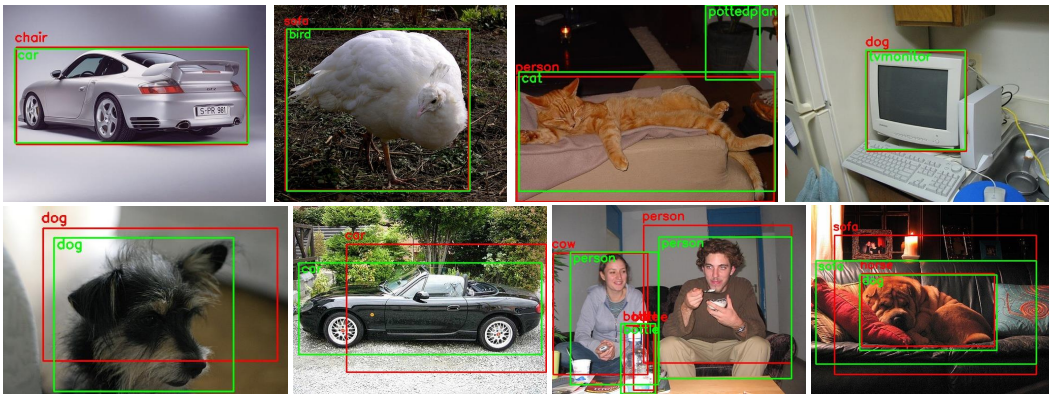


Figure 4: Qualitative results of ELDET on the training set of PASCAL VOC, where green boxes represent the predictions of our method and red boxes indicate the noisy annotations. Our model effectively avoids memorizing both localization and categorization noise.

Table 3: Comparison of detection performance under various noise levels on VinDr-CXR. Best AP scores are highlighted in bold.

Detector	ELDET	Combined Noise		
		20%	30%	40%
RetinaNet	–	29.54	27.44	26.75
	✓	**31.09**	**29.61**	**31.06**
FCOS	–	28.51	27.09	25.32
	✓	**34.40**	**32.55**	**32.74**
Faster R-CNN	–	30.97	29.16	27.87
	✓	**32.91**	**32.52**	**31.61**
GFL	–	29.06	26.76	26.65
	✓	**36.29**	**33.90**	**27.60**

Table 4: Ablation study of task-specific metrics on PASCAL VOC with noise ratio of 40% using RetinaNet.

	Class-agnostic	Ground-truth Box Allocation	AP
(1)	–	–	66.13
(2)	✓	–	67.93
(3)	–	✓	65.70
(4)	✓	✓	**68.82**

underscores superior capability of ELDET to simultaneously handle both types of noise. Qualitative results are illustrated in Figure 4

In contrast, ORSOD and ADELE exhibit limited effectiveness in managing both types of noise. According to the original experimental results of ORSOD, it demonstrate negligible performance gain when integrated with certain architectures (*e.g.*, ReDet [16]), which indicates the low robustness of dynamic loss decay across different detectors. On the other hand, ADELE, which is initially designed for weakly-supervised settings where the ground truth mask quality is notably low, corrects the ground truth with early-learning phase predictions when memorization begins. However, this substitution can occur performance drop in our setting because direct supervision from early-phase predictions may prevent the model from effectively learning informative features.

COCO. Table 2 shows that incorporating ELDET consistently improves AP on the MS COCO dataset compared to the baseline without it. These gains on a large-scale benchmark underscore the robustness of our method and demonstrate its ability to generalize to complex, real-world data.

VinDr-CXR. As shown in Table 3, ELDET maintains its superior performance over the baselines on VinDr-CXR dataset. It is notable that ELDET boosts the performance of FCOS from 25.32 to 32.74 under a 40% combined noise condition. It shows the robustness and adaptability of our proposed method in challenging settings with complex noise patterns in medical imaging. By effectively handling noisy annotations in multi-domain, ELDET showcases its potential for widespread scenarios across various fields where annotation noise is prevalent to happen.

Ablation Study on Task-Specific Early Learning. We conduct an ablation study on PASCAL VOC with a 40% noise ratio using RetinaNet to evaluate the contribution of each metric discussed in Section 4.1. Table 4 summarizes the results under different configurations: (1) without any task-specific metric, (2) using only the class-agnostic (CA) metric for localization, (3) utilizing only the ground-truth box allocation (GTBA) metric for classification, and (4) leveraging both CA and GTBA

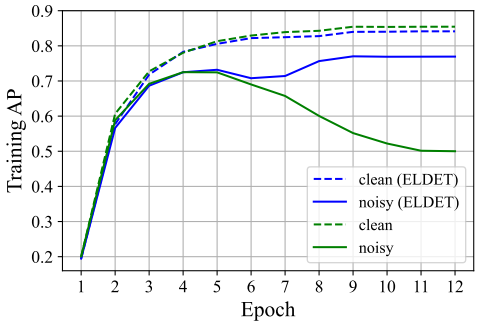


Figure 5: Performance of RetinaNet (green) compared to the setting with ELDET (blue) on the PASCAL VOC training samples. Both models are trained on datasets with 20% combined noisy labels yet evaluated on the original clean annotations. Solid lines represent evaluation on the samples with noisy labels during training, while dashed lines represent evaluation on clean data.

Table 5: Accuracy (%) of noisy label filtering by comparing predictions on noisy samples with the clean labels, evaluated on PASCAL VOC training set using RetinaNet with 20% noise.

ELDET	Noise Type	
	Localization	Categorization
-	73.74	22.66
✓	**80.98**	**74.71**

Table 6: Ablation study of exponential moving average (EMA) of student parameters to update the teacher network. The results are evaluated on PASCAL VOC using RetinaNet with 40% noise.

EMA	Noise Type		
	Localization	Categorization	Combination
-	73.73	66.90	66.03
✓	**74.53**	**73.67**	**68.82**

metrics. Our final setting with both techniques achieves the best performance, which indicates that task-specific monitoring effectively contributes to robust learning under noisy conditions. The CA manner mitigates the impact of categorization noise, while the GTBA evaluation safeguards detectors against localization noise. The proposed combination of metrics enable ELDET to successfully detect the transition from the early-learning phase to memorization for each task respectively.

Robustness to Noisy Annotations. Figure 5 illustrates the performance comparison between RetinaNet and our setting with ELDET when trained on noisy PASCAL VOC but evaluated on the clean labels after each epoch. For example, the solid blue line denotes the *evaluation-with-clean* performance of ELDET on *training-with-noise* samples with 20% noise ratio. This experiment assesses the ability to resist overfitting to noisy annotations and generalize to clean labels. RetinaNet without ELDET shows a significant drop on *evaluation-with-clean* for *training-with-noise* samples (solid green line), which implies the occurrence of memorization to the noisy labels. In contrast, ELDET with high accuracy throughout training demonstrates the robustness in avoiding memorization of noise and its effectiveness in generalizing to clean label even unseen while training. A slight performance drop at the beginning of distillation is observed, which reflects the typical transient instability in early distillation stages caused by conflicting supervision and the delayed stabilization of the EMA teacher.

Noisy Annotation Filtering for Data Curation. In addition, we probe the effectiveness of ELDET in validating the quality of annotations on PASCAL VOC using RetinaNet with a noise ratio of 20%. To measure how successfully models identify noisy labels, we compare the model predictions with the original clean ground truth on noisy samples.

In the case of categorization noise, a prediction is considered to be correct if the class prediction is same with the original category. For localization noise, a prediction is valid if IoU compared to the clean box surpasses a threshold of 0.65. The results in Table 5 reveal that detectors with ELDET not only produce accurate predictions on noisy data but also demonstrate a strong capacity to distinguish between clean and noisy annotations. This capability highlights the potential utility of ELDET for identifying and filtering out noisy labels, which suggests practical applications in data curation or data cleaning.

Effect of EMA on Knowledge Distillation. We evaluate the impact of the exponential moving average (EMA) on our framework on PASCAL VOC using RetinaNet with a noise ratio of 40%. The teacher network without EMA is not updated after early-learning phase with frozen parameters. Table 6 shows that incorporating EMA significantly improves performance across all noise settings. The improvement in the case of categorization noise shows that the teacher network suffers from limited classification performance in the early-learning phase but that EMA effectively overcomes this limitation. These results underscore the effectiveness of EMA in enhancing the teacher guidance.

Table 7: Performance of ELDET on Deformable DETR under combined noise conditions on PASCAL VOC (AP@50).

Method	Clean	20% Noise	30% Noise	40% Noise
Baseline	74.27	68.60	65.39	62.52
ELDET	**74.52**	**68.82**	**65.84**	**62.91**

Table 8: Performance comparison of DINO with query denoising vs. ELDET on PASCAL VOC (40% noise ratio, AP).

Method	Clean	Loc.	Cat.	Combined
Query Denoising	75.37	74.26	67.16	66.26
ELDET	**76.72**	**75.63**	**67.83**	**68.29**

Table 9: Hyperparameter sensitivity analysis for our proposed method under different noise conditions. Results are evaluated on the PASCAL VOC dataset using RetinaNet with 40% noise. The table reports performance across various values of τ and γ. The best AP scores are highlighted in bold, with the second-best scores underlined.

τ	γ	Noise Type		
		Localization	Categorization	Combination
0.1	0.9	74.53	**73.67**	68.82
0.3	0.9	**76.67**	70.55	**73.46**
0.5	0.9	75.41	67.71	66.68
0.1	0.7	<u>76.11</u>	73.07	<u>73.39</u>
0.1	0.8	75.69	<u>73.39</u>	71.81

Transformer-based Detectors. To further demonstrate the generality of ELDET as a plug-and-play module across different architectures, we evaluate it on transformer-based detectors. The results verify ELDET's effectiveness in transformer-specific denoising frameworks and its robustness under challenging noisy conditions. For Deformable DETR [67] on PASCAL VOC with combined noise (as shown in Table 7), ELDET consistently improves performance over the baseline across all noise levels, highlighting its capability to strengthen DETR-based models. Furthermore, to benchmark against native transformer denoising, we replace DINO's query denoising [62] with ELDET and evaluate on PASCAL VOC at 40% noise across all scenarios (Table 8). ELDET surpasses the native method under every noisy condition, with a notable improvement under combined noise (68.29 vs. 66.26 AP). These results validate ELDET's strong adaptability to label noise and confirm its seamless integration into transformer-based detectors such as DINO, extending its applicability beyond convolutional architectures.

Impact of Hyperparameters We conduct ablation studies on two key hyperparameters in our ELDET framework: the IoU threshold τ used in the Ground Truth Box Allocation (GTBA) process, and the deviation threshold γ used for early-learning phase detection. Table 9 presents the Average Precision (AP) scores under different settings of τ and γ across localization, categorization, and combined noise types. While certain configurations like $\tau = 0.3$ and $\gamma = 0.9$ achieve the highest AP under localization and combined noise, the first line with $\tau = 0.1$ and $\gamma = 0.9$ provides strong and balanced performance across all noise conditions.

6 Conclusion

This paper introduced ELDET, a self-knowledge distillation framework that leverages the early-learning phenomenon to address both localization and categorization noise in object detection. By using early-stage models as teacher networks, ELDET effectively mitigates the memorization of noisy labels, resulting in improved robustness and performance across diverse datasets such as PASCAL VOC, MS COCO and VinDr-CXR. The proposed framework is compatible with various detection architectures, making it practical for real-world applications. Future research could explore extending ELDET to video object detection or integrating it into active learning pipelines for automated data curation.

Acknowledgments This work was supported by funding from SAIGE. FC Park was further supported in part by IITP-MSIT grants RS-2021-II212068, 2022-220480, RS-2022-II220480, SNU-AIIS, SNU-IPAI, SNU-IAMD, and the SNU Institute for Engineering Research.

References

[1] Sotiris Anagnostidis, Gregor Bachmann, Lorenzo Noci, and Thomas Hofmann. The curious case of benign memorization. *arXiv preprint arXiv:2210.14019*, 2022.

[2] Yingbin Bai, Erkun Yang, Bo Han, Yanhua Yang, Jiatong Li, Yinian Mao, Gang Niu, and Tongliang Liu. Understanding and improving early stopping for learning with noisy labels. *Advances in Neural Information Processing Systems*, 34:24392–24403, 2021.

[3] Andreas Bär, Jonas Uhrig, Jeethesh Pai Umesh, Marius Cordts, and Tim Fingscheidt. A novel benchmark for refinement of noisy localization labels in autolabeled datasets for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3851–3860, 2023.

[4] Shivendra Bhardwaj, Abbas Ghaddar, Ahmad Rashid, Khalil Bibi, Chengyang Li, Ali Ghodsi, Philippe Langlais, and Mehdi Rezagholizadeh. Knowledge distillation with noisy labels for natural language understanding. *arXiv preprint arXiv:2109.10147*, 2021.

[5] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018.

[6] 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.

[7] Krystian Chachuła, Jakub Łyskawa, Bartłomiej Olber, Piotr Frątczak, Adam Popowicz, and Krystian Radlak. Combating noisy labels in object detection datasets. *arXiv preprint arXiv:2211.13993*, 2022.

[8] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. *Advances in neural information processing systems*, 30, 2017.

[9] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[10] Michael Desmond, Michael Muller, Zahra Ashktorab, Casey Dugan, Evelyn Duesterwald, Kristina Brimijoin, Catherine Finegan-Dollak, Michelle Brachman, Aabhas Sharma, Narendra Nath Joshi, et al. Increasing the speed and accuracy of data labeling through an ai assisted interface. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*, pages 392–401, 2021.

[11] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010.

[12] Di Feng, Ali Harakeh, Steven L Waslander, and Klaus Dietmayer. A review and comparative study on probabilistic object detection in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 23(8):9961–9980, 2021.

[13] Hui Guo, Boyu Wang, and Grace Yi. Label correction of crowdsourced noisy annotations with an instance-dependent noise transition model. *Advances in Neural Information Processing Systems*, 36:347–386, 2023.

[14] Sangchul Hahn and Heeyoul Choi. Self-knowledge distillation in natural language processing. *arXiv preprint arXiv:1908.01851*, 2019.

[15] 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.

[16] Jiaming Han, Jian Ding, Nan Xue, and Gui-Song Xia. Redet: A rotation-equivariant detector for aerial object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2786–2795, 2021.

[17] 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.

[18] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1921–1930, 2019.

[19] Geoffrey Hinton. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[20] Paul F Jaeger, Simon AA Kohl, Sebastian Bickelhaupt, Fabian Isensee, Tristan Anselm Kuder, Heinz-Peter Schlemmer, and Klaus H Maier-Hein. Retina u-net: Embarrassingly simple exploitation of segmentation supervision for medical object detection. In *Machine Learning for Health Workshop*, pages 171–183. PMLR, 2020.

[21] Kyungyul Kim, ByeongMoon Ji, Doyoung Yoon, and Sangheum Hwang. Self-knowledge distillation with progressive refinement of targets. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6567–6576, 2021.

[22] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European conference on computer vision (ECCV)*, pages 734–750, 2018.

[23] Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394*, 2020.

[24] Junnan Li, Caiming Xiong, Richard Socher, and Steven Hoi. Towards noise-resistant object detection with noisy annotations. *arXiv preprint arXiv:2003.01285*, 2020.

[25] Shikun Li, Xiaobo Xia, Shiming Ge, and Tongliang Liu. Selective-supervised contrastive learning with noisy labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 316–325, 2022.

[26] 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.

[27] Yifan Li, Hu Han, Shiguang Shan, and Xilin Chen. Disc: Learning from noisy labels via dynamic instance-specific selection and correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24070–24079, 2023.

[28] Zhihui Li, Pengfei Xu, Xiaojun Chang, Luyao Yang, Yuanyuan Zhang, Lina Yao, and Xiaojiang Chen. When object detection meets knowledge distillation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10555–10579, 2023.

[29] Jiehua Lin, Yan Zhao, Shigang Wang, and Yu Tang. A robust training method for object detectors in remote sensing image. *Displays*, 81:102618, 2024.

[30] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[31] Guozhang Liu, Ting Liu, Mengke Yuan, Tao Pang, Guangxing Yang, Hao Fu, Tao Wang, and Tongkui Liao. Dynamic loss decay based robust oriented object detection on remote sensing images with noisy labels. *arXiv preprint arXiv:2405.09024*, 2024.

[32] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. *Advances in neural information processing systems*, 33:20331–20342, 2020.

[33] Sheng Liu, Kangning Liu, Weicheng Zhu, Yiqiu Shen, and Carlos Fernandez-Granda. Adaptive early-learning correction for segmentation from noisy annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2606–2616, 2022.

[34] Sheng Liu, Zhihui Zhu, Qing Qu, and Chong You. Robust training under label noise by over-parameterization. In *International Conference on Machine Learning*, pages 14153–14172. PMLR, 2022.

[35] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016.

[36] Xinyu Liu, Wuyang Li, Qiushi Yang, Baopu Li, and Yixuan Yuan. Towards robust adaptive object detection under noisy annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14207–14216, 2022.

[37] Yang Liu, Zhuo Ma, Ximeng Liu, Siqi Ma, and Kui Ren. Privacy-preserving object detection for medical images with faster r-cnn. *IEEE Transactions on Information Forensics and Security*, 17:69–84, 2019.

[38] Jiaxin Ma, Yoshitaka Ushiku, and Miori Sagara. The effect of improving annotation quality on object detection datasets: A preliminary study. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4850–4859, 2022.

[39] Duc Nguyen, DungNB, Ha Q. Nguyen, Julia Elliott, NguyenThanhNhan, and Phil Culliton. Vinbigdata chest x-ray abnormalities detection. `https://kaggle.com/competitions/vinbigdata-chest-xray-abnormalities-detection`, 2020. Kaggle.

[40] Ha Q Nguyen, Khanh Lam, Linh T Le, Hieu H Pham, Dat Q Tran, Dung B Nguyen, Dung D Le, Chi M Pham, Hang TT Tong, Diep H Dinh, et al. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. *Scientific Data*, 9(1):429, 2022.

[41] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[42] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[43] J Redmon. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[44] 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.

[45] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019.

[46] T-YLPG Ross and GKHP Dollár. Focal loss for dense object detection. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2980–2988, 2017.

[47] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[48] Kwangrok Ryoo, Yeonsik Jo, Seungjun Lee, Mira Kim, Ahra Jo, Seung Hwan Kim, Seungryong Kim, and Soonyoung Lee. Universal noise annotation: Unveiling the impact of noisy annotation on object detection. *arXiv preprint arXiv:2312.13822*, 2023.

[49] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: A simple and strong anchor-free object detector. *IEEE transactions on pattern analysis and machine intelligence*, 44(4):1922–1933, 2020.

[50] 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.

[51] Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. Distilling object detectors with fine-grained feature imitation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4933–4942, 2019.

[52] Xudong Wang, Rohit Girdhar, Stella X Yu, and Ishan Misra. Cut and learn for unsupervised object detection and instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3124–3134, 2023.

[53] Yue Wang, Alireza Fathi, Abhijit Kundu, David A Ross, Caroline Pantofaru, Tom Funkhouser, and Justin Solomon. Pillar-based object detection for autonomous driving. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 18–34. Springer, 2020.

[54] Bichen Wu, Forrest Iandola, Peter H Jin, and Kurt Keutzer. Squeezedet: Unified, small, low power fully convolutional neural networks for real-time object detection for autonomous driving. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 129–137, 2017.

[55] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3060–3069, 2021.

[56] Ting-Bing Xu and Cheng-Lin Liu. Data-distortion guided self-distillation for deep neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 5565–5572, 2019.

[57] 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.

[58] Ruixin Yang and Yingyan Yu. Artificial convolutional neural network in object detection and semantic segmentation for medical imaging analysis. *Frontiers in oncology*, 11:638182, 2021.

[59] Zhendong Yang, Zhe Li, Xiaohu Jiang, Yuan Gong, Zehuan Yuan, Danpei Zhao, and Chun Yuan. Focal and global knowledge distillation for detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4643–4652, 2022.

[60] Moon Ye-Bin, Dongmin Choi, Yongjin Kwon, Junsik Kim, and Tae-Hyun Oh. Eninst: Enhancing weakly-supervised low-shot instance segmentation. *Pattern Recognition*, 145:109888, 2024.

[61] Sukmin Yun, Jongjin Park, Kimin Lee, and Jinwoo Shin. Regularizing class-wise predictions via self-knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13876–13885, 2020.

[62] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.

[63] Jing Zhang, Victor S Sheng, Tao Li, and Xindong Wu. Improving crowdsourced label quality using noise correction. *IEEE transactions on neural networks and learning systems*, 29(5): 1675–1688, 2017.

[64] Shikun Zhang, Omid Jafari, and Parth Nagarkar. A survey on machine learning techniques for auto labeling of video, audio, and text data. *arXiv preprint arXiv:2109.03784*, 2021.

[65] Qijie Zhao, Tao Sheng, Yongtao Wang, Zhi Tang, Ying Chen, Ling Cai, and Haibin Ling. M2det: A single-shot object detector based on multi-level feature pyramid network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9259–9266, 2019.

[66] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019.

[67] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction are fully supported by the technical contributions and experimental results presented in the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our approach in the supplementary material.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide all necessary information to reproduce the main experimental results that support our claims, including model architecture, training schedule, hyperparameters, and dataset.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We intend to release anonymized code with the camera-ready version. At submission time, the code is not included in the supplementary material.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We include all experimental details in the supplementary material, such as hyperparameter values, implementation-specific choices (e.g., optimizer, batch size, number of training epochs).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We report single-run experimental results without statistical error bars or variation across random seeds. While the results are stable across informal reruns, we acknowledge the need for reporting variance in future revisions.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report the compute environment used for all experiments, including GPU model, training time, and memory usage in the supplementary material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research complies with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss potential future directions stemming from our work.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our work does not involve the release of models or datasets

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We properly credit all external assets used in our work.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We plan to release the new assets introduced in this work, along with proper documentation, at the time of the camera-ready submission.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our work does not involve any human subjects or crowdsourcing experiments.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our research does not involve any experiments with human subjects, and therefore does not require IRB or equivalent ethical approval.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: We did not use any large language models (LLMs) as part of the core methodology.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

Appendix

A Limitations

While our framework demonstrates robust empirical gains under controlled noisy settings, several limitations remain. First, the foundational assumption of our method that localization and categorization exhibit temporally distinct early learning behaviors is based on empirical observations. Although supported by quantitative trends, a theoretical explanation of this phenomenon is beyond the scope of this work and remains an open question. Second, the categorization noise is synthetically generated by uniformly sampling incorrect labels from the set of classes excluding the ground truth. While this is a common strategy in prior work, it does not reflect the structured or semantically biased errors that typically arise in real-world annotation processes. Third, the detection of early-learning phase transitions is performed using curve fitting and gradient slope change, following the heuristic methodology proposed in prior work [33]. However, this procedure is sensitive to the choice of a hyperparameter such as the slope threshold γ, which can alter the estimated transition point and downstream performance. These limitations underscore the need for more realistic noise modeling, theoretically grounded dynamics analysis, and robust, data-adaptive mechanisms for to identify learning phase transitions.

Table 10: Training resource comparison (training time and GPU memory usage) across various detectors and methods.

Dataset	Detector	Method	Training time (hours)	Memory Usage (GB)
PASCAL VOC (20 classes)	RetinaNet	-	2.51	11,418
		ORSOD [31]	3.96	12,135
		ADELE [33]	4.11	11,496
		ELDET	4.23	17,967
	FCOS	-	2.13	10,507
		ORSOD [31]	3.67	10,666
		ADELE [33]	4.05	10,631
		ELDET	3.59	14,892
MS COCO (80 classes)	RetinaNet	-	8.03	11,418
		ELDET	15.43	17,967

B Compute Resources

All experiments were conducted using GPUs with 24GB VRAM (NVIDIA RTX 3090 and 4090). Our framework maintains a teacher model that is initialized at the end of the localization early learning phase and subsequently updated via exponential moving average (EMA) of the student model's parameters. Unlike co-teaching methods that require simultaneous gradient updates to two networks, our approach avoids full dual-model training. Instead, it only requires forward passes through the frozen teacher resulting in moderate memory and compute overhead roughly equivalent to running a single training model alongside a lightweight inference model.

To quantitatively assess computational efficiency, we measured both total training time and GPU memory usage across different configurations. One computational consideration in our setup is the monitoring of early-learning dynamics. To detect phase transitions in learning, we compute validation metrics on the entire training set at every epoch. This process, while crucial to identify transition points accurately, incurs additional time cost especially for large-scale datasets such as MS COCO [30]. This behavior is primarily due to the need for per-epoch validation over the entire training set to detect the early-learning transition point—a step required by all these methods rather than being specific to ELDET.

As shown in Table 10, both ADELE and ORSOD exhibit training time increases comparable to ELDET, primarily due to this per-epoch validation process shared across methods. Although ELDET shows a slightly larger memory footprint, this is mainly due to the additional teacher network used for knowledge distillation. Since the teacher model remains frozen and participates only in forward

passes without backpropagation, the actual computational overhead remains modest. Overall, the additional cost of ELDET is comparable to that of other robust-learning frameworks and is acceptable given the consistent performance improvements observed across datasets.

C Detailed Experimental Settings

C.1 Implementation Details

Our proposed method is implemented using the MMDetection framework [9] built on PyTorch [42]. All input images are resized to 512×512 for consistency. Training is conducted using Stochastic Gradient Descent (SGD) with a momentum of 0.9 and a weight decay of 10^{-4}. The learning rate follows a step schedule, decreasing by a factor of 10 at predefined epochs, except for MS COCO [30] where only a linear scheduler is used. For PASCAL VOC [11] and MS COCO, the learning rate is set to 0.01, and the training spans 12 epochs with a batch size of 32. In contrast, for VinDr-CXR [40], the learning rate is set to 0.005, and the training spans 20 epochs with a batch size of 16. We exclude the "No finding" class in VinDr-CXR data for fair comparison of noisy training scenario. All detectors are initialized with a ImageNet [47] pre-trained ResNet-50 [17], and trained on NVIDIA GPUs.

C.2 ELDET Hyperparameter Details

For the ground-truth box allocation (GTBA), we set the IoU threshold $\tau = 0.1$, replacing predicted box coordinates with ground-truth locations when the IoU exceeds τ. We consider the model to begin memorizing noisy labels when the relative change in the derivative of the performance metric exceeds the criterion with $\gamma = 0.9$. The exponential moving average (EMA) momentum α is set to 0.999 for the overall model and adjusted to $\alpha_{\text{cls}} = 0.1$ for the classification head during the period after localization memorization and before memorizing categorization noise. Other hyperparameters are same as the original setting (*e.g.,* the loss weight λ of MMDetection[†].

C.3 Baselines

ORSOD [31] tackles categorization noise by adopting a dynamic decay mechanism to progressively down-weight the top-k samples with the highest classification loss. The dynamic loss decay function is defined as

$$\mathcal{L}_{\text{DLD}} = \begin{cases} \mathcal{L}_{\text{cls}}(X), & \text{if } t_i < t_{\text{el}}, \\ \alpha \cdot \mathcal{L}_{\text{cls}}(X_k) + \mathcal{L}_{\text{cls}}(X_r), & \text{if } t_i \geq t_{\text{el}}, \end{cases} \tag{8}$$

where $\mathcal{L}_{\text{cls}}$ is the classification loss, X_k and X_r represent the top-k and remaining samples, respectively, t_i denotes the current training epoch, and t_{el} is the early-learning termination epoch. The decay factor α is defined as:

$$\alpha = \exp\left(-\frac{c}{t_i - t_{\text{el}}}\right), \tag{9}$$

where c is a constant controlling the rate of decay (set to 10 in our experiments). This adaptive mechanism ensures that high-loss samples have reduced impact in later training epochs, which promotes more stable and noise-resilient learning. However, a limitation of ORSOD is that it only suppresses the classification loss without explicitly addressing localization noise in the annotations, which may limit its effectiveness in handling noisy box-level annotations.

ADELE [33] was originally developed for semantic segmentation tasks with noisy annotations, leveraging the observation that early-learning concludes at different times for each class. By updating the labels of pixels where the model's prediction score exceeds a certain threshold (*e.g.,* 0.8) at the class-specific early-learning endpoints, ADELE effectively refines noisy annotations, enabling robust segmentation performance even in the presence of noise. To adapt ADELE for object detection, we modified the approach to account for the inherent differences between segmentation and detection tasks. Instead of utilizing class-specific early-learning endpoints, we defined a unified early-learning endpoint across all classes. At this point, model predictions are used to refine annotations by replacing

[†] `https://github.com/open-mmlab/mmdetection`

Table 11: Evaluation on the compatibility with various knowldege distillation techniques. The results are evaluated on PASCAL VOC using RetinaNet with 40% noise. The best AP scores are highlighted in bold, with the second-best scores underlined.

KD	Noise Type		
	Localization	Categorization	Combination
-	70.27	<u>68.07</u>	65.63
CrossKD [50]	**74.53**	**73.67**	**68.82**
FGD [59]	73.11	67.85	<u>66.61</u>
OFD [18]	<u>73.36</u>	67.50	66.03

noisy labels with more reliable predictions that meet strict criteria: (1) a prediction score of at least 0.5, and (2) an Intersection-over-Union (IoU) exceeding 0.5 with the corresponding ground-truth bounding box. For such cases, both the coordinates and the class label of the original ground truth are updated to match the model's prediction.

C.4 Knowledge Distillation Loss Functions

We adopt the knowledge distillation loss functions used in CrossKD [50] to guide the student models in mimicking the un-memorized knowledge of teacher models. For RetinaNet [46], we use the Quality Focal Loss [26] for classification and the Generalized IoU Loss [45] for localization. In the case of FCOS [49], the classification loss is implemented with Focal Loss [46], while the localization loss employs IoU Loss. For Faster R-CNN [44], the classification loss is based on KL Divergence, and the localization loss uses L1 Loss. Lastly, for GFL [26], the classification loss is also Quality Focal Loss, but the localization loss relies on KD Divergence Loss. These loss functions ensure effective knowledge transfer by aligning the outputs of the student models with those of early-phase teacher models.

D Additional Experimental Results

D.1 Compatible with Different Distillation Techniques

To demonstrate the flexibility of our ELDET framework, we investigate its compatibility with various knowledge distillation techniques beyond CrossKD [50]. Specifically, we integrate FGD [59] and OFD [18] into our framework and evaluate their performance under different types of noise.

As shown in Table 11, integrating FGD and OFD into our framework yields improvements over the baseline without distillation under localization and the combined noise. These improvements indicate that our ELDET framework is compatible with different KD techniques and can benefit from them. However, CrossKD consistently outperforms the other distillation methods across all noise types. These results suggest that while our framework can effectively incorporate various KD methods, CrossKD provides the most substantial improvements in our experiments. This superiority may be attributed to CrossKD's ability to facilitate task-oriented knowledge transfer without only focusing on transferring fine-grained feature embeddings from the teacher. Anagnostidis *et al.* [1] found that neural networks are tolerant to label noise except in the last layer, which indicates the vulnerability of the later layers of detectors to noisy annotations. In other words, direct distillation from the classification head of the teacher to that of the student using CrossKD can mitigate the memorization of noisy labels unlike other approaches.

D.2 EMA Decay Rates

We analyze the impact of the momentum α, α_{cls} and the exponential moving average (EMA) update cycle of the student parameters for the update of the teacher network. Table 12 shows that a small momentum of $\alpha = 0.9$ reduces performance, suggesting that a strong momentum is crucial to maintaining the stability of the teacher network. Similarly, setting $\alpha_{\text{cls}} = 0.999$ or $\alpha_{\text{cls}} = 1.0$ (*i.e.*, updating the classification head slowly before early-learning terminates for classification task) results

Table 12: Evaluation of the EMA (Exponential Moving Average) strategy with varying parameters α, α_{cls}, and interval settings. Results are reported as AP@50 on the PASCAL VOC dataset using RetinaNet with 40% combined noise. The table highlights the performance impact of different EMA configurations. The best AP scores are highlighted in bold, with the second-best scores underlined.

α	α_{cls}	Interval	AP@50
0.999	0.1	1	**68.82**
0.999	0.1	3	<u>68.60</u>
0.999	0.1	5	68.52
0.9	0.1	1	65.54
0.999	0.999	1	66.88
1.0	1.0	1	66.03

Table 13: Detection performance comparison under various noise levels on Oxford Pets using RetinaNet. Best AP scores are highlighted in bold.

ELDET	Noise Level	Localization Noise	Categorization Noise	Combined Noise
-	30%	89.30	79.40	84.30
✓	30%	**89.90**	**83.60**	**88.50**
-	50%	81.00	76.80	66.50
✓	50%	**85.90**	**78.60**	**79.50**
-	70%	79.10	71.40	64.00
✓	70%	**83.10**	**80.10**	**85.90**

in lower AP. This confirms that using a smaller decay rate $\alpha_{\text{cls}} = 0.1$ for the classification head is important to allow it to adapt more quickly, preventing the teacher from lagging behind the student's learning on classification tasks.

D.3 Qualitative Examples on VinDr-CXR

Figure 6 presents a qualitative comparison of the detection results of the baseline FCOS [49] and our proposed ELDET method on the VinDr-CXR [40] training set. This comparison underscores the inherent challenges associated with localization and categorization anomalies in medical images. Despite the presence of noisy labels, the detector utilizing ELDET demonstrates significantly better alignment with the ground-truth annotations compared to the baseline FCOS. It highlights the effectiveness of our method in mitigating the adverse effects of both localization and categorization noise. Furthermore, this result emphasizes the robustness of ELDET in diverse domains, demonstrating its applicability not only to real-world images but also to the challenging domain of medical imaging.

D.4 Results on Smaller Datasets

Although we have already evaluated our method on the relatively small dataset, VinDr-CXR [40], which contains more than ten thousand samples, we further investigated whether the proposed approach remains effective on smaller datasets. We conducted additional experiments on the Oxford Pets [41], which consists of 27 classes with approximately 200 images for each class. As shown in table 13, applying ELDET led to consistently higher performance compared to the baseline without ELDET.

D.5 Distinctive Early Learning Termination.

We further investigate when the model begins to memorize noisy annotations for localization and classification tasks respectively. As reported in Table 14, models tends to memorize localization noise significantly earlier compared to categorization noise on PASCAL VOC and COCO with various detectors. This observation outlines the necessity of our task-specific guidance mechanism which indicates the appropriate moment to initiate teacher-student distillation for each task.

Table 14: Termination epochs of the localization (t_{loc}) and classification (t_{cls}) early-learning phases, and their difference.

Dataset	Detector	Noise Level	t_{loc}	t_{cls}	Difference
PASCAL VOC (20 classes)	RetinaNet	20%	3	7	+4
		30%	4	4	+0
		40%	4	11	+7
	FCOS	20%	3	9	+6
		30%	4	10	+6
		40%	3	11	+8
	Faster R-CNN	20%	7	8	+1
		30%	4	11	+7
		40%	3	4	+1
	GFL	20%	3	8	+5
		30%	3	9	+6
		40%	6	12	+6
	Average	-	3.92	8.67	+4.75
MS COCO (80 classes)	RetinaNet	20%	8	8	+0
		30%	4	6	+2
		40%	5	9	+4
	FCOS	20%	4	8	+4
		30%	4	8	+4
		40%	3	8	+4
	Faster R-CNN	20%	6	11	+5
		30%	5	11	+6
		40%	6	12	+6
	GFL	20%	7	12	+5
		30%	8	8	+0
		40%	4	12	+8
	Average	-	5.42	9.42	+4

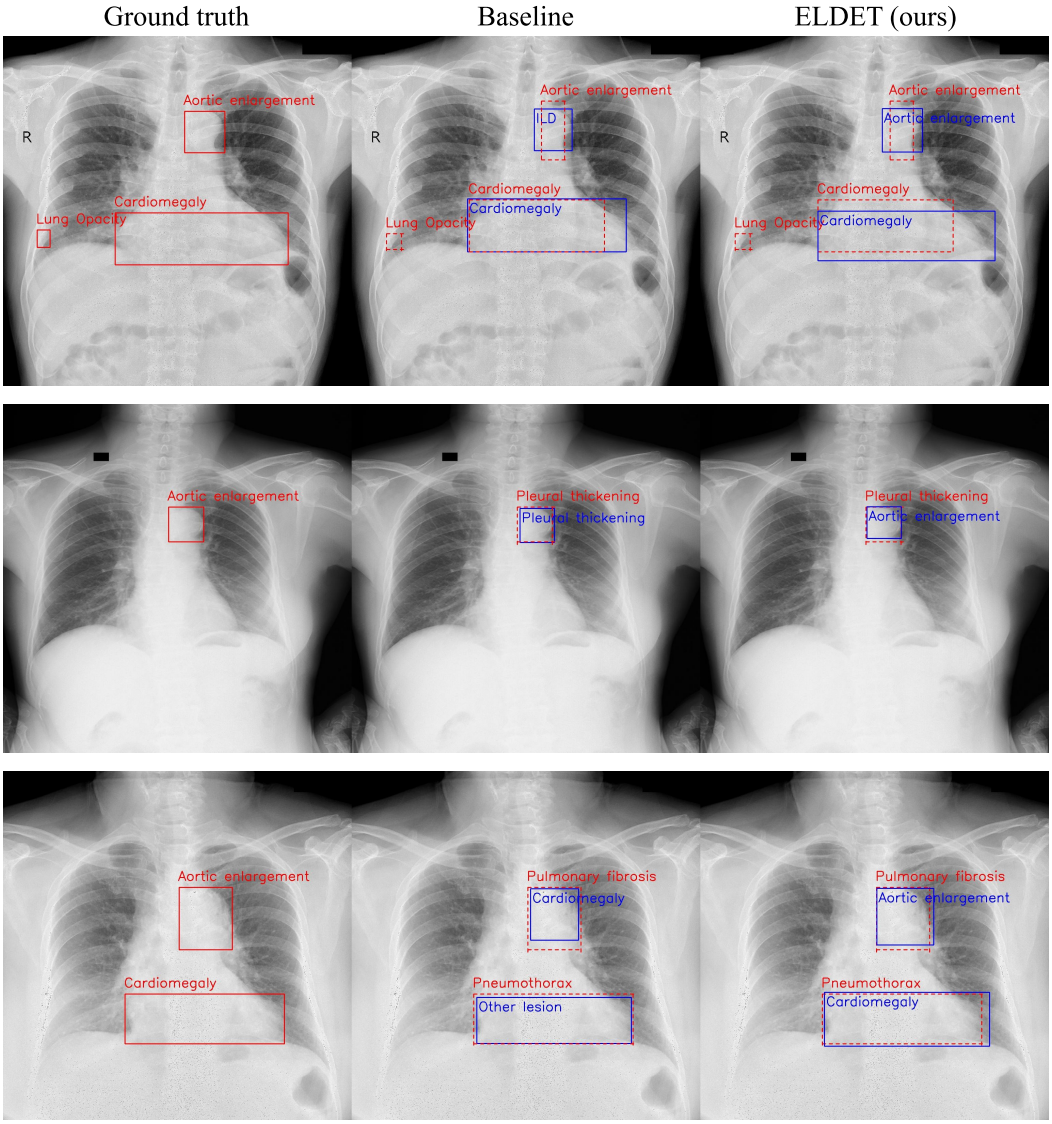


Figure 6: Qualitative Analysis of FCOS on the VinDr-CXR training set. Blue boxes denote model predictions, while red boxes with bold outline represent ground truth annotations. The left panel illustrates the original clean annotations, the middle panel displays predictions from the baseline FCOS, and the right panel shows model outputs with our proposed ELDET. Dotted red boxes in the middle and right panels highlight noisy labels encountered during training. Our proposed ELDET method demonstrates superior capability in mitigating the effects of both localization and categorization noise.