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

1 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 [9]. 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.

2 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. 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 [7].

3 Detailed Experimental Settings

3.1 Implementation Details

Our proposed method is implemented using the MMDetection framework [2] built on PyTorch [11]. 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 [7] where only a linear scheduler is used. For PASCAL VOC [3] 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 [10], 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 [15] pre-trained ResNet-50 [4], and trained on NVIDIA GPUs.

3.2 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 MMDetection¹).

3.3 Baselines

ORSOD [8] 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

$$\mathcal{L}_{\text{DLD}} = \begin{cases} \mathcal{L}_{\text{cls}}(X), & \text{if } t_i < t_{\text{el}}, \\ \alpha \cdot \mathcal{L}_{\text{cls}}(X_k) + \mathcal{L}_{\text{cls}}(X_r), & \text{if } t_i \geq t_{\text{el}}, \end{cases} \quad (1)$$

where \mathcal{L}_{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:

$$\alpha = \exp\left(-\frac{c}{t_i - t_{\text{el}}}\right), \quad (2)$$

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 [9] 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 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.

3.4 Knowledge Distillation Loss Functions

We adopt the knowledge distillation loss functions used in CrossKD [17] to guide the student models in mimicking the un-memorized knowledge of teacher models. For RetinaNet [14], we use the Quality Focal Loss [6] for classification and the Generalized IoU Loss [13] for localization. In the case of FCOS [16], the classification loss is implemented with Focal Loss [14], while the localization loss employs IoU Loss. For Faster R-CNN [12], the classification loss is based on KL Divergence, and the localization loss uses L1 Loss. Lastly, for GFL [6], the classification loss is also Quality Focal Loss, but the localization loss relies on KD Divergence Loss. These loss functions ensure effective

¹<https://github.com/open-mmlab/mmdetection>

Table 1: Evaluation on the compatibility with various knowledge 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 [17] | 74.53 | 73.67 | 68.82 |
| FGD [18] | 73.11 | 67.85 | <u>66.61</u> |
| OFD [5] | <u>73.36</u> | 67.50 | 66.03 |

Table 2: 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 |

knowledge transfer by aligning the outputs of the student models with those of early-phase teacher models.

4 Additional Experimental Results

4.1 Compatible with Different Distillation Techniques

To demonstrate the flexibility of our ELDET framework, we investigate its compatibility with various knowledge distillation techniques beyond CrossKD [17]. Specifically, we integrate FGD [18] and OFD [5] into our framework and evaluate their performance under different types of noise.

As shown in Table 1, 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.

4.2 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 2 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

Table 3: 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 4: Termination epoch of the early-learning phase across various noise ratios on PASCAL VOC with FCOS.

| Noise Ratio | Noise Type | |
|-------------|--------------|----------------|
| | Localization | Categorization |
| 20% | 3.00 | 9.00 |
| 30% | 3.33 | 10.00 |
| 40% | 3.00 | 11.00 |

noise, the first line with $\tau = 0.1$ and $\gamma = 0.9$ provides strong and balanced performance across all noise conditions.

4.3 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 3 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 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.

4.4 Qualitative Examples on VinDr-CXR

fig. 1 presents a qualitative comparison of the detection results of the baseline FCOS [16] and our proposed ELDET method on the VinDr-CXR [10] 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.

4.5 Distinctive Early Learning Termination.

We further investigate when the model begins to memorize noisy annotations for localization and classification tasks separately. As reported in table 4, models tends to memorize localization noise significantly earlier compared to categorization noise on PASCAL VOC with FCOS detector. Moreover, it is noticeable that early-learning termination for categorization noise is prolonged as the noise level increases. This observation outlines the necessity of our task-specific guidance mechanism which indicates the appropriate moment to initiate teacher-student distillation for each task.

References

- [1] Sotiris Anagnostidis, Gregor Bachmann, Lorenzo Noci, and Thomas Hofmann. The curious case of benign memorization. *arXiv preprint arXiv:2210.14019*, 2022.
- [2] 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.
- [3] 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.
- [4] 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.
- [5] Byeongho Heo, Jeessoo 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.
- [6] 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.
- [7] 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.
- [8] 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.
- [9] 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.
- [10] 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.
- [11] 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.
- [12] 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.
- [13] 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.
- [14] 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.
- [15] 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.

- [16] 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.
- [17] 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.
- [18] 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.

