Adaptive Early-Learning Correction for Segmentation from Noisy Annotations

Sheng Liu*  Kangning Liu*  Weicheng Zhu  Yiqiu Shen  Carlos Fernandez-Granda1,2
1 NYU Center for Data Science  2 NYU Courant Institute of Mathematical Sciences

Abstract

Deep learning in the presence of noisy annotations has been studied extensively in classification, but much less in segmentation tasks. In this work, we study the learning dynamics of deep segmentation networks trained on inaccurately annotated data. We observe a phenomenon that has been previously reported in the context of classification: the networks tend to first fit the clean pixel-level labels during an “early-learning” phase, before eventually memorizing the false annotations. However, in contrast to classification, memorization in segmentation does not arise simultaneously for all semantic categories. Inspired by these findings, we propose a new method for segmentation from noisy annotations with two key elements. First, we detect the beginning of the memorization phase separately for each category during training. This allows us to adaptively correct the noisy annotations in order to exploit early learning. Second, we incorporate a regularization term that enforces consistency across scales to boost robustness against annotation noise. Our method outperforms standard approaches on a medical-imaging segmentation task where noises are synthesized to mimic human annotation errors. It also provides robustness to realistic noisy annotations present in weakly-supervised semantic segmentation, achieving state-of-the-art results on PASCAL VOC 2012.

1. Introduction

Semantic segmentation is a fundamental problem in computer vision. The goal is to assign a label to each pixel in an image, indicating its semantic category. Deep learning models based on convolutional neural networks (CNNs) achieve state-of-the-art performance [9, 39, 51, 65]. These models are typically trained in a supervised fashion, which requires pixel-level annotations. Unfortunately, gathering pixel-level annotations is very costly, and may require significant domain expertise in some applications [17, 32, 40, 48]. Furthermore, annotation noise is inevitable in some applications. For example, in medical imaging, segmentation annotation may suffer from inter-reader annotation variations [22, 63]. Learning to perform semantic segmentation from noisy annotations is thus an important topic in practice.

Prior works on learning from noisy labels focus on classification tasks [33, 46, 57]. There are comparatively fewer works on segmentation, where existing works focus on designing noise-robust network architecture [50] or incorporating domain specific prior knowledge [42]. We instead focus on improving the performance in a more general perspective by studying the learning dynamics. We observe that the networks tend to first fit the clean annotations during an “early-learning” phase, before eventually memorizing the false annotations, thus jeopardizing generalization performance. This phenomenon has been reported in the context of classification [33]. However, this phenomenon in semantic segmentation differs significantly from its counterpart in classification in the following ways:

• The noise in segmentation labels is often spatially dependent. Therefore, it is beneficial to leverage spatial information during training.

• In semantic segmentation, early learning and memorization do not occur simultaneously for all semantic categories due to pixel-wise imbalanced labels. Previous methods [28, 33] in noisy label classification often assume class...
balanced data and thus either detecting or handling wrong labels for different classes at the same time.

• The annotation noise in semantic segmentation can be ubiquitous (all examples have some errors) while the state-of-the-art methods in classification [28, 33, 67] assume that some samples are completely clean.

Inspired by these observations, we propose a new method, ADELE (ADAdaptive Early-Learning corrEction), that is designed for segmentation from noisy annotations. Our method detects the beginning of the memorization phase by monitoring the Intersection over Union (IoU) curve for each category during training. This allows it to adaptively correct the noisy annotations in order to exploit early-learning for individual classes. We also incorporate a regularization term to promote spatial consistency, which further improves the robustness of segmentation networks to annotation noise.

To verify the effectiveness of our method, we consider a setting where noisy annotations are synthesized and controllable. We also consider a practical setting – Weakly-Supervised Semantic Segmentation (WSSS), which aims to perform segmentation based on weak supervision signals, such as image-level labels [24, 54], bounding box [11, 44], or scribbles [30]. We focus on a popular pipeline in WSSS. This pipeline consists of two steps (See Figure 2). First, a classification model is used to generate pixel-level annotations. This is often achieved by applying variations of Class Activation Maps (CAM) [66] combined with post-processing techniques [3, 25]. Second, these pixel-level annotations are used to train a segmentation model (such as deeplabv1 [8]). Generated by a classification model, the pixel-wise annotations supplied to the segmentation model are inevitably noisy, thus the second step is indeed a noisy segmentation problem. We therefore apply ADELE to the second step. In summary, our main contributions are:

• We propose a novel approach (ADELE) to perform semantic segmentation with noisy pixel-level annotations, which exploits early learning by adaptively correcting the annotations using the model output.

• We evaluate ADELE on the thoracic organ segmentation task where annotations are corrupted to resemble human errors. ADELE is able to avoid memorization, outperforming standard baselines. We also perform extensive experiments to study ADELE on various types and levels of noises.

• ADELE achieves the state of the art on PASCAL VOC 2012 for WSSS. We show that ADELE can be combined with several different existing methods for extracting pixel-level annotations [3, 14, 52] in WSSS, consistently improving the segmentation performance by a substantial margin.

2. Methodology

2.1. Early learning and memorization in segmentation from noisy annotations

In a typical classification setting with label noise, a subset of the images are incorrectly labeled. It has been observed in prior works that deep neural networks tend to first fit the training data with clean labels during an early-learning phase, before eventually memorizing the examples with incorrect labels [4, 33]. Here, we show that this phenomenon also occurs in segmentation when the available pixel-wise annotations are noisy (i.e. some of the pixels are incorrect). We consider two different problems. First, segmentation in medical imaging, where annotation noise is mainly due to human error. Second, the annotation noise in weakly-supervised semantic segmentation due to the bias of classification models, as they mostly focus on discriminative regions, and the post-processing errors may result in systematic over or under segmentation.

Given noisy annotations for which we know the ground truth, we can quantify the early-learning and memorization phenomena by analyzing the model output on the pixels that are incorrectly labeled:

• **early learning IoU\(_{el}\):** We quantify early learning using the overlap (measured in terms of the Intersection over Union (IoU) metric) between the outputs and the corresponding ground truth label on the pixels that are incorrectly labeled, denoted by IoU\(_{el}\).

• **memorization IoU\(_{m}\):** We quantify memorization using the overlap (measured in IoU) between the CNN outputs and the incorrect labels, denoted by IoU\(_{m}\).

Figure 3 demonstrates the phenomena of early-learning and memorization on a randomly corrupted CT-scan segmentation dataset (SegTHOR [27]). We analyze the learning
Figure 3. We visualize the effect of early learning ($\text{IoU}_e$, green curves) and memorization ($\text{IoU}_m$, red curves) on incorrectly annotated pixels with (solid lines) and without (dashed lines) ADELE for each foreground category of a medical dataset SegThor [27]. The model is a UNet trained with noisy annotations that mimic human errors. $\text{IoU}_e$ is the IOU between the model output and the ground truth computed over the incorrectly-labeled pixels. $\text{IoU}_m$ is the IOU between the model output and the incorrect annotations. For all classes, $\text{IoU}_m$ increases substantially as training proceeds because the model gradually memorizes the incorrect annotations. This occurs at different speeds for different categories. In contrast, $\text{IoU}_e$ first increases during an early-learning stage where the model learns to correctly segment the incorrectly-labeled pixels, but eventually decreases as memorization occurs. Like memorization, early-learning also happens at varying speeds for the different semantic categories. See Figure 10 in Appendix for the plot on PASCAL VOC.

Figure 4 illustrates the effect of early learning and memorization on the model output. In the medical-imaging application, the noisy annotations (third column) are synthesized to resemble human annotation errors which either miss or encompass the ground truth regions (compare to second column). Right after early learning, these regions are identified by the segmentation model (fourth column), but after memorization the model overfits to the incorrect annotations and forgets how to segment these regions correctly (fifth column). Similar effects are observed in WSSS, in which the noisy annotations generated by the classification model are missing some object regions, perhaps because they are not particularly discriminative (e.g. the body of the dog, cat and people in the first, second, and fourth row respectively, or the upper half of the bus in the third row). The segmentation model first identify these regions but eventually overfits to the incorrect annotations. Our goal in this work is to modify the training of segmentation models on noisy annotations in order to prevent memorization. This is achieved by combining two strategies described in the next two sections. Figure 3 and Figure 4 shows that the resulting method substantially mitigates memorization (solid red lines) and promotes continued learning beyond the early-learning stage (solid green lines).

2.2. Adaptive label correction based on early-learning

The early-learning phenomenon described in the previous section suggests a strategy to enhance segmentation models: correcting the annotations using the model output. Similar ideas have inspired works in classification with noisy labels [33, 37, 46, 60]. However, different from the classification task where the noise is mainly sample-wise, the annotation noise is ubiquitous across examples and distributed in a pixel-wise manner. There is a key consideration for this approach to succeed: the annotations cannot be corrected too soon, because this degrades their quality. Determining when to correct the pixel-level annotations using the model output is challenging for two reasons:

- Correcting all classes at the same time can be sub-optimal.
- During training, we do not have access to the performance of the model on ground-truth annotations (otherwise we would just use them to train the model in the first place!).

To overcome these challenges we propose to update the annotations corresponding to different categories at different times by detecting when early learning has occurred and memorization is about to begin using the training performance of the model.

In our experiments, we observe that the segmentation performance on the training set (measured by the IoU be-
between the model output and the noisy annotations) improves rapidly during early learning, and then much more slowly during memorization (see the rightmost graph in Figure 5). We propose to use this deceleration to decide when to update the noisy annotations. To estimate the deceleration we first fit the following exponential parametric model to the training IoU using least squares:

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

(1)

where $t$ represents training time and $0 < a \leq 1, b \geq 0$, and $c \geq 0$ are fitting parameters. Then we compute the derivative $f'(t)$ of the parametric model with respect to $t$ at $t = 1$ and at the current iteration.$^2$ For each semantic category, the annotations are corrected when the relative change in derivative is above a certain threshold $r$, i.e. when

$$\left|\frac{f'(1) - f'(t)}{|f'(1)|}\right| > r,$$

(2)

$^2$The derivative is given by $f'(t) = abce^{-bt^c}tc^{-1}$.
which we set to 0.9, and at every subsequent epoch. We only correct annotations for which the model output has confidence above a certain threshold \( \tau \), which we set to 0.8. A detailed description about the label correction is attached in the Appendix B. As shown in Table 2, adaptive label correction based on early learning improves segmentation models in the medical-imaging applications and WSSS, both on its own and in combination with multiscale-consistency regularization. Figure 4 shows some examples of annotation corrections (rightmost column).

### 2.3. Multiscale consistency

As we previously mentioned, model outputs after early-learning are used to correct noisy annotations. Therefore, the quality of model outputs is crucial for the effectiveness of the proposed method. Following a common procedure that has shown to result in more accurate segmentation from the outputs [31, 58], we average model outputs corresponding to multiple rescaled copies of inputs to form the final segmentation, and use them to correct labels. Furthermore, we incorporate a regularization that imposes consistency of the outputs across multi-scales and is able to make averaged outputs more accurate (See the right graph of Figure 6). This idea is inspired by consistency regularizations, a popular concept in the semi-supervised learning literature [6, 15, 23, 26, 36, 43, 47] that encourages the model to produce predictions that are robust to arbitrary semantically-preserving spatial perturbations. In segmentation with noisy annotation, we introduce the consistency loss to provide an extra supervision signal to the network, preventing the network from only training on the noisy segmentation annotations, and overfitting to them. This regularization effect is also observed in the literature of classification with label noise [10, 28]. Since our method uses the network predictions to correct labels, it is crucial to avoid overfitting to the noisy segmentation.

To be more specific, let \( s \) be the number of scaling operations. In our experiments we set \( s = 3 \) (downscaling \( \times 0.7 \), no scaling, and upscaling \( \times 1.5 \)). We denote by \( p_k(x) \), \( 1 \leq k \leq s \), the model predictions for an input \( x \) rescaled according to these operations (see Figure 6). We propose to use a regularization term \( \mathcal{L}_{\text{Multiscale}} \) to promote consistency between \( p_k(x) \), \( 1 \leq k \leq s \), and the average \( q(x) = \frac{1}{s} \sum_{k=1}^{s} p_k(x) \):

\[
\mathcal{L}_{\text{Multiscale}}(x) = -\frac{1}{s} \sum_{k=1}^{s} \text{KL}(p_k(x) \parallel q(x)),
\]

where KL denotes the Kullback-Leibler divergence. The term is only applied to the input \( x \) where the maximum entry of \( q(x) \) is above a threshold \( \rho \) (equal to 0.8 for all experiments). The regularization is weighted by a parameter \( \lambda \) (set to one in all experiments) and then combined with a cross-entropy loss based on the available annotations. As shown in Tables 2, with multiscale consistency regularization, adaptive label correction further improves segmentation performance in both medical-imaging applications and the
Weakly supervised semantic segmentation (WSSS).

In contrast, we propose to work learns high-level spatial structures for fluorescence initial pixel-level annotations, by modifying the classification model. These techniques mostly focus on improving the annotations [6]. The classification model is first used to produce pixel-level annotations. These methods mainly focus on improving the classification model itself [29, 52, 53, 55], or by post-processing these annotations [2, 3, 49]. However, the resulting annotations are still noisy [62] (see Figure 4). Our goal is to improve the segmentation model by adaptively accounting for this noise. Similar approach to our method has been observed in object detection where network outputs are dynamically used for training [21]. In semantic segmentation, the work that is most similar to our label-correction strategy is [18], which is inspired by traditional seeded region-growing techniques [1]. This method estimates the foreground using an additional model [19], and initializes the foreground segmentation estimate with classification-based annotations. This estimate is used to train a segmentation model, which is then used to iteratively update the estimate. ADELE seeks to correct the initial annotations, as opposed to growing them, and does not need to identify the foreground estimate or an initial subset of highly-accurate annotations.

3. Related work

Classification from noisy labels. Early learning and memorization were first discovered in image classification from noisy labels [33]. Several methods exploit early learning to improve classification models by correcting the labels or adding regularization [33, 37, 46, 57, 60]. Here we show that segmentation from noisy labels also exhibits early learning and memorization. However, these dynamics are different for different semantic categories. ADELE exploits this to perform correction in a class-adaptive fashion.

Segmentation from noisy annotations. Segmentation from noisy annotations is an important problem, especially in the medical domain [5]. Some recent works address this problem by explicitly taking into account systematic human labeling errors [63], and by modifying the segmentation loss to increase robustness [42, 50]. [35] propose to discover noisy gradient by collecting information from two networks connected with mutual attention. [34] shows that the network learns high-level spatial structures for fluorescence microscopy images. These structures are then leveraged as supervision signals to alleviate influence from wrong annotations. These methods mainly focus on improving the robustness by exploiting some setting-specific information (e.g. network architecture, dataset, requiring some samples with completely clean annotation). In contrast, we propose to study the learning dynamics of noisy segmentation and propose ADELE, which performs label correction by exploiting early learning.

Weakly supervised semantic segmentation (WSSS). Recent methods for WSSS [3, 14, 61] are mostly based on the approach introduced by Ref. [24, 54], where a classification model is first used to produce pixel-level annotations [66], which are then used to train a segmentation model. These techniques mostly focus on improving the initial pixel-level annotations, by modifying the classification model itself [29, 52, 53, 55], or by post-processing these annotations [2, 3, 49]. However, the resulting annotations are still noisy [62] (see Figure 4). Our goal is to improve the segmentation model by adaptively accounting for this noise. Similar approach to our method has been observed in object detection where network outputs are dynamically used for training [21]. In semantic segmentation, the work that is most similar to our label-correction strategy is [18], which is inspired by traditional seeded region-growing techniques [1]. This method estimates the foreground using an additional model [19], and initializes the foreground segmentation estimate with classification-based annotations. This estimate is used to train a segmentation model, which is then used to iteratively update the estimate. ADELE seeks to correct the initial annotations, as opposed to growing them, and does not need to identify the foreground estimate or an initial subset of highly-accurate annotations.

4. Segmentation on Medical Images with Annotation Noise

Segmentation from noisy annotations is a fundamental challenge in the medical domain, where available annotations are often hampered by human error [63]. Here, we evaluate ADELE on a segmentation task where the goal is to identify organs from computed tomography images.

Settings. The dataset consists of 3D CT scans from the SegTHOR dataset [27]. Each pixel is assigned to the esophagus, heart, trachea, aorta, or background. We treat each 2D slice of the 3D scan as an example, resizing to 256 × 256 pixels. We randomly split the slices into a training set of 3638 slices, a validation set of 570 slices, and a test set of 580 slices. Each patient only appears in one of these subsets. We generate annotation noise by applying random degrees of dilation and erosion to the ground-truth segmentation labels, mimicking common human errors [63] (see Figure 4). In the main experiment, the noisy annotation is with a mIoU of 0.6 w.r.t the ground truth annotation. We further control the
Ablation study for each part of ADELE.

with multi-scale inputs as our with or without class-adaptively correcting labels, on the test set of 

improves the performance. Most importantly, combining any of these methods with label correction would substantially improve the performance. ADELE, which combines label correction with the proposed regularization, achieves the best performance. We also include ablation studies for the hyperparameters $r$, $\tau$ and $\rho$ in Appendix C. Additional segmentation results are provided in Appendix A.1.

5. Noisy Annotations in Weakly-supervised Semantic Segmentation

We adopt a prevailing pipeline for training WSSS (described in detail in Section 1), in which some pixel-wise annotations are generated using image-level labels to supervise a segmentation network. These pixel-wise annotations are noisy. Therefore, we apply ADELE to this WSSS pipeline.

We evaluate ADELE on a standard WSSS dataset – PASCAL VOC 2012 [13], which has 21 annotation classes (including background), and contains 1464, 1449 and 1456 images in the training, validation (val) and test sets respectively. Following [41,45,52,59,61,62], we use an augmented training set with 10582 images with annotations from [16].

Baseline Models. To demonstrate the broad applicability of our approach, we apply ADELE using pixel-level annotations generated by three popular WSSS models: AffinityNet [3], SEAM [52] and ICD [14], which do not rely on external datasets or external saliency maps. The annotations are produced by a classification model combined with the post-processing specified in [3,14,52]. We provide details on the training procedure in Section B in the Appendix. We use the same inference pipeline as SEAM [52], which includes multi-scale inference [3,14,52,64] and CRF [25].

Comparison with the state-of-the-art. Table 3 compares the performance of the proposed method ADELE to state-of-the-art WSSS methods on PASCAL VOC 2012. ADELE improves the performance of AffinityNet [3], SEAM [52] and ICD [14] substantially on the validation and test sets. Moreover, ADELE combined with SEAM [52] and ICD [14] achieves state-of-the-art performance on both sets. Although it uses only image-level labels, ADELE outperforms state-of-the-art methods [20,45,59,64] that rely on external saliency models [19]. To show that our method is complementary with other more advanced WSSS methods, we have conducted an experiment with a recent WSSS method NSROM [59], which uses external saliency models. ADELE+NSROM achieves mIoU of 71.6 and 72.0 on the validation and test set respectively, which is the SoTA for WSSS with ResNet segmentation backbone (see Appendix A.2).

Figure 8 compares the performance of SEAM and the performance of ADELE combined with SEAM on the validation set separately for each semantic category. ADELE improves performance for most categories, with the exception of a few categories where the baseline model does not perform well (e.g. chair, bike). On Figure 1 and 9, we show some qualitative segmentation results from the validation.
Table 2. Ablation study for ADELE on SegTHOR [27] and PASCAL VOC 2012 [13]. We report the mIoU achieved at the last epoch on the validation set for both dataset. Class-adaptive label correction mechanism achieves the best performance when combined with multi-scale consistency regularization.

<table>
<thead>
<tr>
<th>Previous methods</th>
<th>ADELE +</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet-101</td>
<td>ResNet-38</td>
</tr>
<tr>
<td>Val</td>
<td></td>
</tr>
<tr>
<td>61.4</td>
<td>64.1</td>
</tr>
<tr>
<td>66.1</td>
<td>64.9</td>
</tr>
<tr>
<td>64.5</td>
<td>66.1</td>
</tr>
<tr>
<td>65.8</td>
<td>68.9</td>
</tr>
<tr>
<td>Test</td>
<td></td>
</tr>
<tr>
<td>63.2</td>
<td>64.3</td>
</tr>
<tr>
<td>65.9</td>
<td>63.7</td>
</tr>
<tr>
<td>65.5</td>
<td>66.7</td>
</tr>
<tr>
<td>65.5</td>
<td>68.8</td>
</tr>
<tr>
<td>69.3</td>
<td>68.6</td>
</tr>
</tbody>
</table>

Table 3. Comparison with state-of-the-art methods on the Pascal VOC 2012 dataset using mIoU (%). The best performance under each set are highlighted in red and blue respectively. The version of CONTA [62] reported here is deployed combined with SEAM [52]. The results clearly show that ADELE outperforms other approaches.

6. Limitations

The success of ADELE seems to rely to some extent on the quality of the initial annotations. When these annotations are of poor quality, ADELE may only produce a marginal improvement or even have negative impact (see Figure 8 and 9). An related limitation is that when the annotation noise is highly structured, early-learning may not occur, because there may not be sufficient information in the noisy annotations to correct the errors. In that case label correction based on early-learning will be unsuccessful. Illustrative examples are provided in the fifth and sixth rows of Figure 1, where the initial annotations completely encompass the bicycle, and completely misclassify the chair as a sofa.

7. Conclusion

In this work, we introduce a novel method to improve the robustness of segmentation models trained on noisy annotations. Inspired from the early-learning phenomenon, we proposed ADELE to boost the performance on the segmentation of thoracic organ, where noise is incorporated to resemble human annotation errors. Moreover, standard segmentation networks, equipped with ADELE, achieve the state-of-the-art results for WSSS on PASCAL VOC 2012. We hope that this work will trigger interest in the design of new forms of segmentation methods that provide robustness to annotation noise, as this is a crucial challenge in applications such as medicine. We also hope that the work will motivate further study of the early-learning and memorization phenomena in settings beyond classification.

Acknowledgments SL, KL, WZ, and YS were partially supported by NSF NRT grant HDR-1922658. SL was partially supported by NSF grant DMS 2009752 and Alzheimer’s Association grant AARG-NTF-21-848627. KL was partially supported by NIH grant (R01LM013316). YS was partially supported by NIH grant (P41EB017183, R21CA225175). CFG acknowledges support from NSF OAC 2103936.
References


[57] Xiaobo Xia, Tongliang Liu, Bo Han, Chen Gong, Nannan Wang, Zongyuan Ge, and Yi Chang. Robust early-learning: Hindering the memorization of noisy labels. In *ICLR*, 2021.


