# Adaptive Early-Learning Correction for Segmentation from Noisy Annotations

Sheng Liu*[1] Kangning Liu*[1] Weicheng Zhu[1] Yiqiu Shen[1] Carlos Fernandez-Granda[1,2]
[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]

## 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.

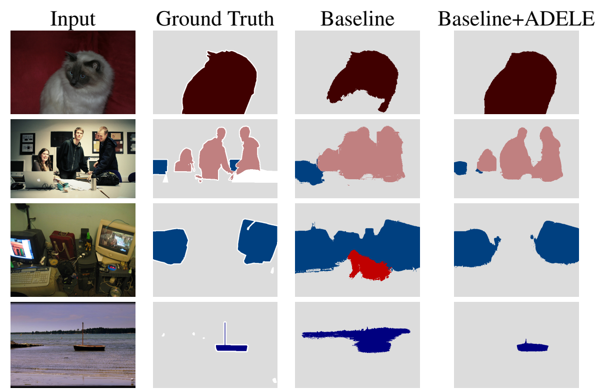


Figure 1. Visualization of the segmentation results of the baseline method SEAM [52] and the baseline combined with the proposed ADaptive Early-Learning corrEction (ADELE). Our proposed ADELE improves segmentation quality. More examples can be found in Appendix A.1.

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

*The first two authors contribute equally, order decided by coin flipping.
[1]Code is available at https://github.com/Kangningthu/ADELE

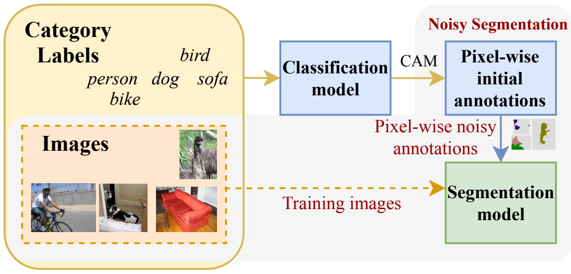


Figure 2. A prevailing pipeline for training WSSS. We aim to improve the segmentation model from noisy annotations.

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 (ADaptive 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 analyze the behavior of segmentation networks when trained with noisy pixel-level annotations. We show that the training dynamics can be separated into an early-learning and a memorization stage in segmentation with annotation noise. Crucially, we discover that these dynamics differ across each semantic category.

- 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** $\textbf{IoU}_{el}$: 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 $\text{IoU}_{el}$.

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

Figure 3 demonstrates the phenomena of early-learning and memorization on a randomly corrupted CT-scan segmentation dataset (SegTHOR [27]). We analyze the learning

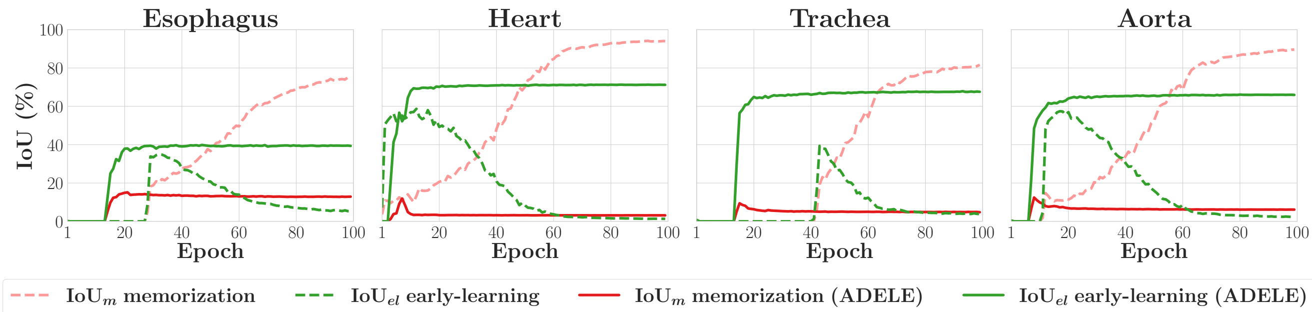


Figure 3. We visualize the effect of early learning (IoU$_{el}$, green curves) and memorization (IoU$_m$, red curves) on *incorrectly annotated pixels* with (solid lines) and without (dashed lines) ADELE for each foreground category of a medical dataset SegThor [27]. The model is a UNet trained with noisy annotations that mimic human errors. IoU$_{el}$ is the IOU between the model output and the ground truth computed over the incorrectly-labeled pixels. IoU$_m$ is the IOU between the model output and the incorrect annotations. For all classes, 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, IoU$_{el}$ 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.

curve on the incorrectly-annotated pixels during the training process. The plots show the IoU$_m$ (dashed red line) and IoU$_{el}$ (dashed green line) at different training epochs. For all classes, the IoU between the output and the incorrect labels (IoU$_m$) increases substantially as training proceeds because the model gradually memorizes the incorrect annotations. This memorization process occurs at **varying speeds** for different semantic categories (compare *heart* and *Aorts* with *Traches* or *Esophagus* in the SegThor dataset). The IoU between the output and the correct labels (IoU$_{el}$) follows a completely different trajectory: it first increases during an early-learning stage where the model learns to correctly segment the incorrectly-labeled pixels, but eventually decreases as memorization occurs (for the WSSS dataset, we observe a very similar phenomenon shown in Figure 11 in the Appendix). Like memorization, early-learning also happens at varying speeds for the different semantic categories.

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-

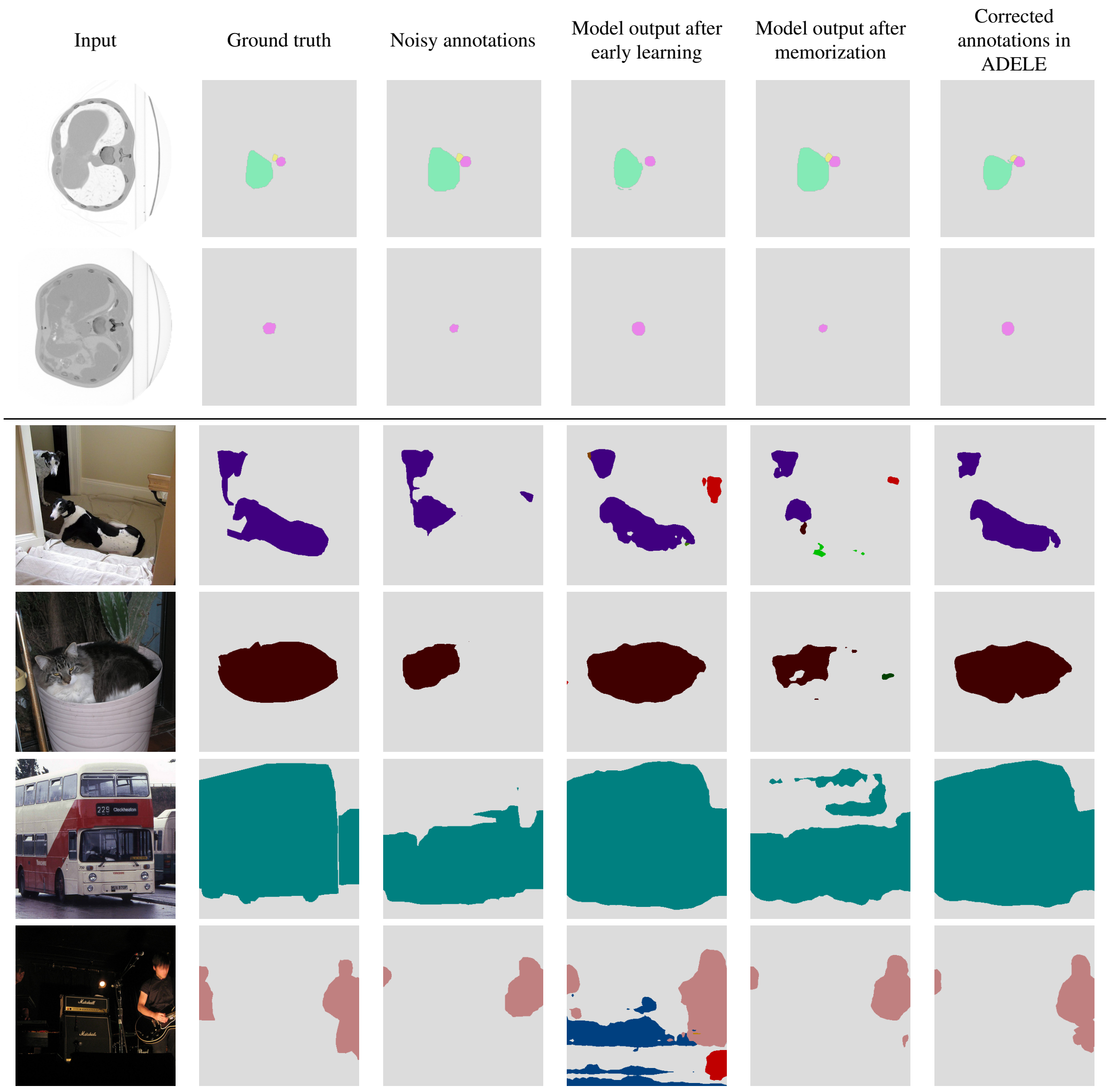


Figure 4. Visual examples illustrating the early-learning and memorization phenomena. For several images in a medical dataset Segthor [27] (top tow rows) and the WSSS dataset VOC 2012 [13] (bottom four rows), we show the ground-truth annotations (second column), noisy annotations (third column) obtained by a synthetic corruption process for the medical data and by the classification-based SEAM [52] model for WSSS, the output of a model segmentation model trained on the noisy annotations after early learning (fourth column), and the output of the same model after memorization (fifth column). The model for the medical dataset is a UNet. The WSSS model is a standard DeepLab-v1 network trained with the SEAM annotations. As suggested by the graphs in Figure 3 after early learning the model corrects some of the annotation errors, but these appear again after memorization. ADELE is able to correct the labels leveraging the early learning output, thereby avoiding memorization (sixth column). We set the background color to light gray for ease of visualization.

tween 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:

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

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

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

[2]The derivative is given by $f'(t) = abce^{-bt^{c}}t^{c-1}$.

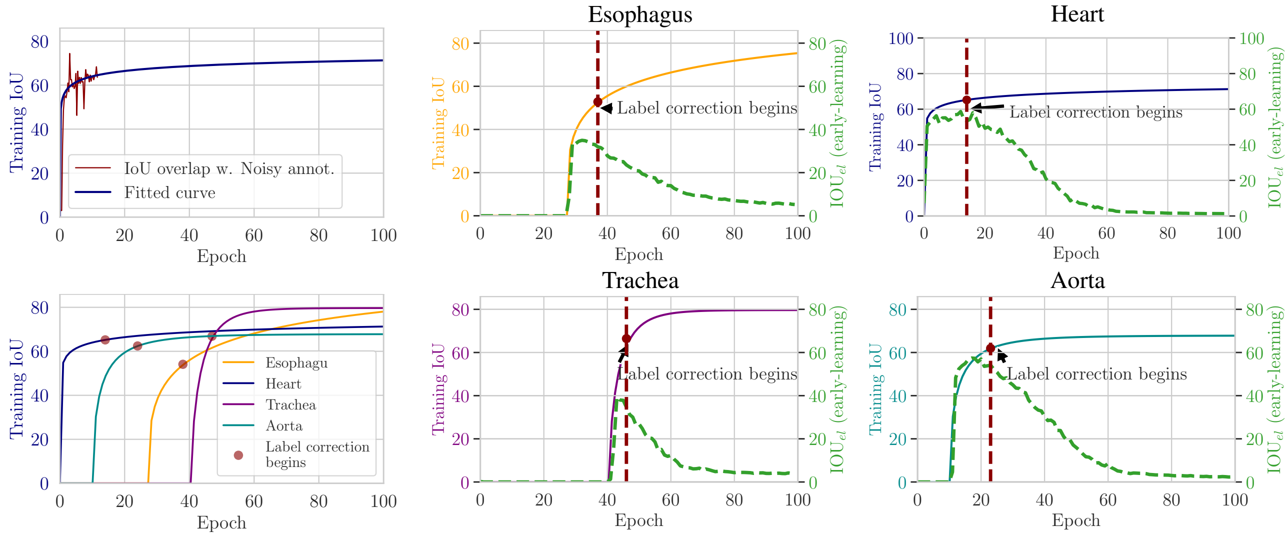


Figure 5. Illustration of the proposed curve fitting method to decide when to begin label correction in ADELE (Results on SegThor). **First column:** On the top, we plot the IoU between the model predictions and the initial noisy annotations for the same model used in Figures 3 and 4 and the corresponding fit with the parametric model in Equation 1. The label correction beginning iteration is based on the relative slope change of the fitted curve. The bottom image shows the label correction times for different semantic categories, showing that they are quite different. **Second and third columns:** the green lines show the IoU$_{el}$ for different categories *Esophagus*, *Heart*, *Trachea* and *Aorta*. The IoU$_{el}$ equals the IoU between the model output and the ground truth computed over the incorrectly-labeled pixels, and therefore quantifies early-learning. The label correction begins close to the end of the early-learning phase, as desired. More result in section A.1 in Appendix shows that this also occurs for VOC 2012.

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 semantic-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 $q(x) = \frac{1}{s}\sum_{k=1}^{s} p_k(x)$:

$$\mathcal{L}_{\text{Multiscale}}(x) = -\frac{1}{s}\sum_{k=1}^{s} \text{KL}\left(p_k(x) \parallel q(x)\right), \tag{3}$$

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

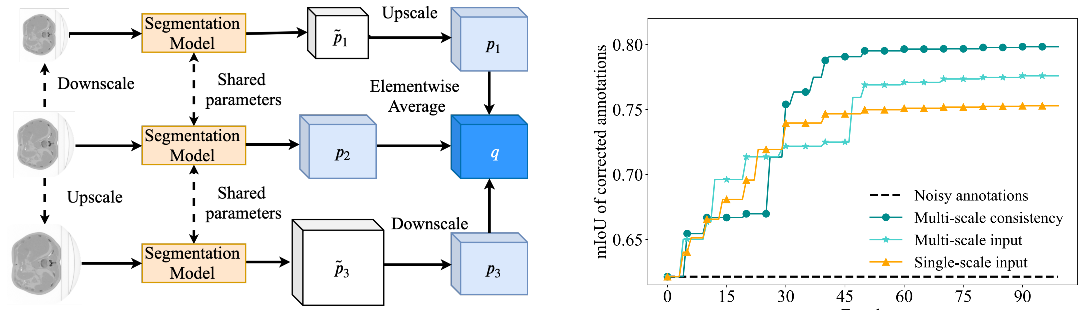


Figure 6. **Left:** In the proposed multiscale-consistency regularization, rescaled copies of the same input (here upscaled $\times 1.5$ and downscaled $\times 0.7$) are fed into the segmentation model. The outputs ($\tilde{p}_1$, $p_2$ and $\tilde{p}_3$) are rescaled to have the same dimensionality ($p_1$, $p_2$ and $p_3$). Regularization promotes consistency between these rescaled outputs and their elementwise average $q$. **Right:** Multi-scale consistency regularization leads to more accurate corrected annotations (results on SegThor, results for VOC 2012 can be found in Figure 12).

WSSS.

## 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 $256 \times 256$ 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

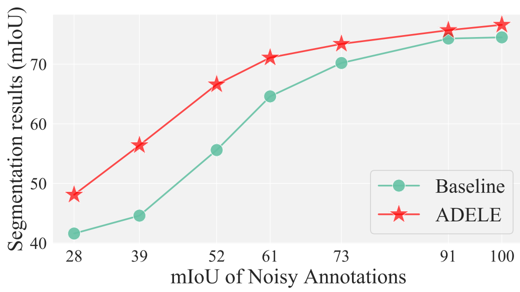


Figure 7. The performance comparison of the baseline and ADELE on the test set of SegTHOR [27]. The model is trained on noisy annotations with various levels of corruption (measured in mIoU with the clean ground truth annotations). ADELE is able to improve the model performance across a wide range of corruption levels.

degree of dilation and erosion to simulate noisy annotation sets with different noise levels for testing the model robustness. We corrupt all annotations in the training set, but not in the validation and test sets. Our **evaluation metric** is Mean Intersection over Union (mIoU).

| | Baseline | ADELE w/o class adaptive | ADELE |
|---|---|---|---|
| Best val | 62.6±2.3 | 40.7±2.5 | **71.1±0.7** |
| Max test | 63.3±2.0 | 40.7±2.4 | **71.2±0.6** |
| Last Epoch | 59.1±1.3 | 40.5±2.3 | **70.8±0.7** |

Table 1. The mIoU (%) comparison of the baseline and ADELE with or without class-adaptively correcting labels, on the test set of SegTHOR [27]. We report the test mIoU of the model that performs best on the validation set (**Best Val**), the test mIoU at the last epoch (**Last Epoch**), and the highest test performance during training (**Max Test**). We report mean and standard deviation after training the model with five realizations of noisy annotations.

**Results.** For a fair comparison, we choose a UNet trained with multi-scale inputs as our **baseline**. We report the mIoU of the baseline and ADELE on the test set of SegTHOR dataset in Table 1. ADELE outperforms the baseline method at all three evaluation epochs. Moreover, correcting labels at the same time for all classes will have a detrimental effect on the performance.

**Impacts of noise levels.** Figure 7 provides empirical evidence that ADELE is robust to a wide range of noises. The mIoU of noisy annotations (x-axis) indicates the correctness of the noisy annotations. Thus the smaller the mIoU shows the higher noise levels. The improvements achieved by ADELE are substantial when the noise levels are moderate.

**Ablation study for each part of ADELE.** We perform an ablation study to understand how different parts of ADELE contribute to the final performance. From Table 2, we observe that the model trained with multiple rescaled versions of the input (illustrated in left graph of Figure 6) performs better than the model trained only with the original scale of the input. The proposed spatial consistency regularization further 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

| | SegTHOR | | | PASCAL VOC 2012 | | |
|---|---|---|---|---|---|---|
| Label correction | Single scale | Multiscale input augmentation | Multiscale consistency regularization | Single scale | Multiscale input augmentation | Multiscale consistency regularization |
| ✗ | 58.8 | 60.7 | 62.5 | 64.5 | 65.5 | 66.7 |
| ✓ | 65.2 | 69.8 | **72.2** | 65.6 | 67.3 | **69.3** |

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.

| | Previous methods | | | | | | | **ADELE +** | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | DSRG [18] | ICD [14] | SCE [7] | AffinityNet [3] | SSDD [41] | SEAM [52] | CONTA [62] | AffinityNet [3] | SEAM [52] | ICD [14] |
| | ResNet-101 | | | ResNet-38 | | | | | | |
| Val | 61.4 | 64.1 | 66.1 | 61.7 | 64.9 | 64.5 | 66.1 | 64.8 | **69.3** | 68.6 |
| Test | 63.2 | 64.3 | 65.9 | 63.7 | 65.5 | 65.7 | 66.7 | 65.5 | 68.8 | **68.9** |

Table 3. Comparison with state-of-the-art methods on the Pascal VOC 2012 dataset using mIoU (%). The **best** and the best previous method 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.

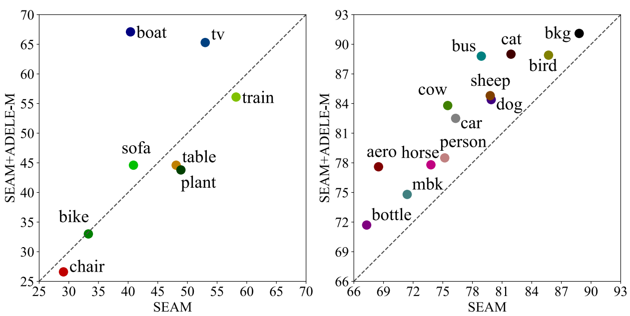


Figure 8. Category-wise comparison of the IoU (%) of SEAM [52] and SEAM combined with the proposed method ADELE on the validation set of PASCAL VOC 2012. We separate the categories based on IoUs for better visualization.

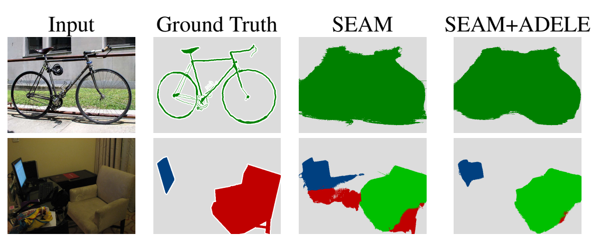


Figure 9. Visualization of the segmentation results of both methods for several examples. ADELE fails to improves segmentation for the bicycle and chair due to highly structured segmentation errors. We set the background color to gray for ease of visualization.

set. Figure 1 shows examples where ADELE successfully improves the SEAM segmentation. Figure 9 shows examples where it does not. In both the output of SEAM has highly structured segmentation errors: the prediction encompasses the bike but completely fails to capture its inner structure, and the chair is missclassified as a sofa. This supports the conclusion that ADELE provides less improvement when the baseline method performs poorly.

## 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 missclassify 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


[1] Rolf Adams and Leanne Bischof. Seeded region growing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(6):641–647, 1994.

[2] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[3] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[4] Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. In *International Conference on Machine Learning*, 2017.

[5] Andrew J Asman and Bennett A Landman. Formulating spatially varying performance in the statistical fusion framework. *IEEE Transactions on Medical Imaging*, 31(6):1326–1336, 2012.

[6] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, 2019.

[7] Yu-Ting Chang, Qiaosong Wang, Wei-Chih Hung, Robinson Piramuthu, Yi-Hsuan Tsai, and Ming-Hsuan Yang. Weakly-supervised semantic segmentation via sub-category exploration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[8] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014.

[9] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2017.

[10] Hao Cheng, Zhaowei Zhu, Xingyu Li, Yifei Gong, Xing Sun, and Yang Liu. Learning with instance-dependent label noise: A sample sieve approach. *arXiv preprint arXiv:2010.02347*, 2020.

[11] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Ieee, 2009.

[13] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015.

[14] Junsong Fan, Zhaoxiang Zhang, Chunfeng Song, and Tieniu Tan. Learning integral objects with intra-class discriminator for weakly-supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[15] Geoff French, Timo Aila, Samuli Laine, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, high-dimensional perturbations. *arXiv preprint arXiv:1906.01916*, 2019.

[16] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *Proceedings of the IEEE International Conference on Computer Vision*, 2011.

[17] Mohammad Havaei, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle. Brain tumor segmentation with deep neural networks. *Medical Image Analysis*, 35:18–31, 2017.

[18] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[19] Huaizu Jiang, Jingdong Wang, Zejian Yuan, Yang Wu, Nanning Zheng, and Shipeng Li. Salient object detection: A discriminative regional feature integration approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[20] Peng-Tao Jiang, Qibin Hou, Yang Cao, Ming-Ming Cheng, Yunchao Wei, and Hong-Kai Xiong. Integral object mining via online attention accumulation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[21] Zequn Jie, Yunchao Wei, Xiaojie Jin, Jiashi Feng, and Wei Liu. Deep self-taught learning for weakly supervised object localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[22] Eytan Kats, Jacob Goldberger, and Hayit Greenspan. A soft staple algorithm combined with anatomical knowledge. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 510–517. Springer, 2019.

[23] Jongmok Kim, Jooyoung Jang, and Hyunwoo Park. Structured consistency loss for semi-supervised semantic segmentation. *arXiv preprint arXiv:2001.04647*, 2020.

[24] Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *Proceedings of the European Conference on Computer Vision*. Springer, 2016.

[25] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *Advances in Neural Information Processing Systems*, 24, 2011.

[26] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *International Conference on Learning Representations*, 2018.

[27] Zoé Lambert, Caroline Petitjean, Bernard Dubray, and Su Kuan. Segthor: Segmentation of thoracic organs at risk in ct images. In *2020 Tenth International Conference on Image*

*Processing Theory, Tools and Applications (IPTA)*. IEEE, 2020.

[28] Junnan Li, Richard Socher, and Steven C.H. Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations*, 2020.

[29] Qizhu Li, Anurag Arnab, and Philip HS Torr. Weakly-and semi-supervised panoptic segmentation. In *Proceedings of the European Conference on Computer Vision*, 2018.

[30] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[31] Di Lin, Yuanfeng Ji, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Multi-scale context intertwining for semantic segmentation. In *Proceedings of the European Conference on Computer Vision*, 2018.

[32] Kangning Liu, Yiqiu Shen, Nan Wu, Jakub Chledowski, Carlos Fernandez-Granda, and Krzysztof J Geras. Weakly-supervised high-resolution segmentation of mammography images for breast cancer diagnosis. In *Medical Imaging with Deep Learning*, 2021.

[33] 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*, 2020.

[34] Yaoru Luo, Guole Liu, Wenjing Li, Yuanhao Guo, and Ge Yang. Deep neural networks learn meta-structures to segment fluorescence microscopy images. *arXiv preprint arXiv:2103.11594*, 2021.

[35] Shaobo Min, Xuejin Chen, Zheng-Jun Zha, Feng Wu, and Yongdong Zhang. A two-stream mutual attention network for semi-supervised biomedical segmentation with noisy labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4578–4585, 2019.

[36] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1979–1993, 2018.

[37] Scott E. Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training deep neural networks on noisy labels with bootstrapping. *CoRR*, abs/1412.6596, 2015.

[38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015.

[39] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[40] Jo Schlemper, Ozan Oktay, Liang Chen, Jacqueline Matthew, Caroline Knight, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention-gated networks for improving ultrasound scan plane detection. *arXiv preprint arXiv:1804.05338*, 2018.

[41] Wataru Shimoda and Keiji Yanai. Self-supervised difference detection for weakly-supervised semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[42] Yucheng Shu, Xiao Wu, and Weisheng Li. Lvc-net: Medical image segmentation with noisy label based on local visual cues. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019.

[43] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020.

[44] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[45] Guolei Sun, Wenguan Wang, Jifeng Dai, and Luc Van Gool. Mining cross-image semantics for weakly supervised semantic segmentation. In *Proceedings of the European Conference on Computer Vision*. Springer, 2020.

[46] Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa. Joint optimization framework for learning with noisy labels. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[47] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, 2017.

[48] Michael Treml, José Arjona-Medina, Thomas Unterthiner, Rupesh Durgesh, Felix Friedmann, Peter Schuberth, Andreas Mayr, Martin Heusel, Markus Hofmarcher, Michael Widrich, et al. Speeding up semantic segmentation for autonomous driving. In *MLITS, NIPS Workshop*, 2016.

[49] Paul Vernaza and Manmohan Chandraker. Learning random-walk label propagation for weakly-supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[50] Guotai Wang, Xinglong Liu, Chaoping Li, Zhiyong Xu, Jiugen Ruan, Haifeng Zhu, Tao Meng, Kang Li, Ning Huang, and Shaoting Zhang. A noise-robust framework for automatic segmentation of covid-19 pneumonia lesions from ct images. *IEEE Transactions on Medical Imaging*, 39(8):2653–2663, 2020.

[51] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[52] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[53] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, July 2017.

[54] Yunchao Wei, Xiaodan Liang, Yunpeng Chen, Xiaohui Shen, Ming-Ming Cheng, Jiashi Feng, Yao Zhao, and Shuicheng Yan. Stc: A simple to complex framework for weakly-supervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11):2314–2320, 2016.

[55] Yunchao Wei, Huaxin Xiao, Honghui Shi, Zequn Jie, Jiashi Feng, and Thomas S Huang. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[56] Zifeng Wu, Chunhua Shen, and Anton Van Den Hengel. Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognition*, 90:119–133, 2019.

[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.

[58] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *CVPR*, 2018.

[59] Yazhou Yao, Tao Chen, Guosen Xie, Chuanyi Zhang, Fumin Shen, Qi Wu, Zhenmin Tang, and Jian Zhang. Non-salient region object mining for weakly supervised semantic segmentation. *arXiv preprint arXiv:2103.14581*, 2021.

[60] Kun Yi and Jianxin Wu. Probabilistic end-to-end noise correction for learning with noisy labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[61] Bingfeng Zhang, Jimin Xiao, Yunchao Wei, Mingjie Sun, and Kaizhu Huang. Reliability does matter: An end-to-end weakly supervised semantic segmentation approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[62] Dong Zhang, Hanwang Zhang, Jinhui Tang, Xiansheng Hua, and Qianru Sun. Causal intervention for weakly-supervised semantic segmentation. *arXiv preprint arXiv:2009.12547*, 2020.

[63] Le Zhang, Ryutaro Tanno, Mou-Cheng Xu, Chen Jin, Joseph Jacob, Olga Ciccarelli, Frederik Barkhof, and Daniel C Alexander. Disentangling human error from the ground truth in segmentation of medical images. *arXiv preprint arXiv:2007.15963*, 2020.

[64] Tianyi Zhang, Guosheng Lin, Weide Liu, Jianfei Cai, and Alex Kot. Splitting vs. merging: Mining object regions with discrepancy and intersection loss for weakly supervised semantic segmentation. In *Proceedings of the European Conference on Computer Vision*, 2020.

[65] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[66] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[67] Tianyi Zhou, Shengjie Wang, and Jeff Bilmes. Robust curriculum learning: From clean label detection to noisy label self-correction. In *International Conference on Learning Representations*, 2020.