AdaWAC: ADAPTIVELY WEIGHTED AUGMENTATION CONSISTENCY REGULARIZATION FOR VOLUMETRIC MEDICAL IMAGE SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Sample reweighting is an effective strategy for learning from training data coming from a mixture of different subpopulations. However, existing reweighting algorithms do not fully take advantage of the particular type of data distribution encountered in volumetric medical image segmentation, where the training data images are uniformly distributed but their associated data labels fall into two subpopulations—"label-sparse" and "label-dense"—depending on whether the data image occurs near the beginning/end of the volumetric scan or the middle. For this setting, we propose *AdaWAC* as an adaptive weighting algorithm that assigns label-dense samples to supervised cross-entropy loss and label-sparse samples to unsupervised consistency regularization. We provide a convergence guarantee for *AdaWAC* by appealing to the theory of online mirror descent on saddle point problems. Moreover, we empirically demonstrate that *AdaWAC* not only enhances segmentation performance and sample efficiency but also improves robustness to the subpopulation shift in labels.

1 INTRODUCTION

Modern machine learning has been revolutionizing the field of medical imaging, especially in computer-aided diagnosis with Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) scans. While the successes of most classical learning algorithms (*e.g.*, empirical risk minimization (ERM)) build upon the assumption that training samples are independently and identically distributed (*i.i.d.*), real-world volumetric medical images rarely fit into this picture. Specifically for medical image segmentation, as instantiated in Figure 1, the segmentation labels corresponding to different cross-sections of organs within a given volume tend to have distinct distributions. That is, the slices toward the beginning/end of the volume that contain no target organs have very few, if any, segmentation labels (which we refer to as "label-sparse"); whereas segmentation labels are prolific in the slices toward the middle of the volume ("label-dense"). Such discrepancy in labels results in distinct difficulty levels measured by the training cross-entropy (Wang et al., 2021b) and leads to various training schedulers (Tullis & Benjamin, 2011; Tang et al., 2018; Hacohen & Weinshall, 2019). Motivated by the separation between label-sparse and label-dense samples, we explore the following questions in this work:

What is the effect of separation between sparse and dense labels on segmentation?
Can we leverage such separation to improve the segmentation accuracy?

We first formulate the mixture of label-sparse and label-dense samples as a subpopulation shift in the conditional distribution of labels given images $P(\mathbf{y}|\mathbf{x})$. As illustrated in Figure 1, such subpopulation shift induces a separation in supervised cross-entropy between sparse and dense labels despite the uniformity of data images.

Prior works address the subpopulation shift issue by utilizing hard thresholding algorithms such as Trimmed Loss Estimator (Shen & Sanghavi, 2019), MKL-SGD (Shah et al., 2020), Ordered SGD (Kawaguchi & Lu, 2020), and quantile-based Kacmarz algorithm (Haddock et al., 2020). However, these trimmed-loss-based methods discard the samples from some subpopulations (*e.g.* samples with label corruption estimated by their losses) at each iteration, which results in loss of

information in the discarded data, leading to reduced sample efficiency. Relaxing the hard thresholding operator to soft thresholding is proposed to incorporate the information from both subpopulations (Wang et al., 2018; Sagawa et al., 2020). However, lowering the weights assigned to some subpopulations of data according to the properties of their labels reduces the importance of the data and labels simultaneously, suggesting that we may further improve the learning efficiency by exploiting the uniformity of data and the separation of labels separately.

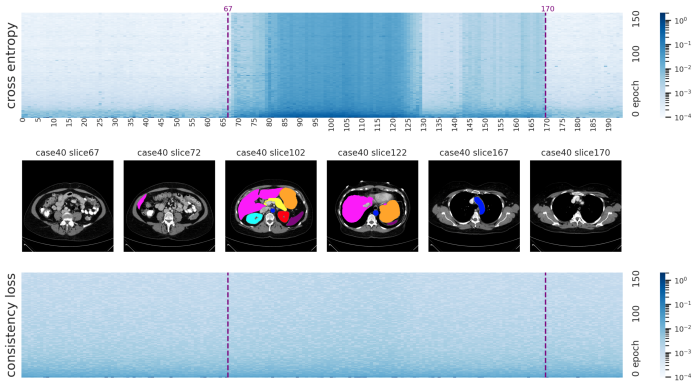

Figure 1: Evolution of cross-entropy losses versus consistency regularization terms for slices across one training volume (Case 40) in the Synapse dataset (Section 5) during training.

Instead of thresholding out or down-weighting the label-sparse samples, we exploit the image inputs of these samples via *augmentation consistency regularization*. Consistency regularization (Bachman et al., 2014; Laine & Aila, 2016; Sohn et al., 2020) aims to learn the proximity between data augmentations of the same samples; crucially, this set-up does not involve the data labels, and hence consistency regularization has become an essential strategy for utilizing unlabeled data. For medical imaging tasks, consistency regularization has been extensively studied in the semi-supervised learning setting (Bortsova et al., 2019; Zhao et al., 2019; Li et al., 2020; Wang et al., 2021a; Zhang et al., 2021; Zhou et al., 2021; Basak et al., 2022). By contrast, we will explore its potency in the fully supervised setting—leveraging the spare information in all image inputs, regardless of their label subpopulations.

Moreover, in light of the uniformity of unsupervised consistency on different slices throughout each volume, the augmentation consistency of the encoder layer outputs serves as a natural reference for separating samples from different subpopulations. Whereby, we introduce the *weighted augmentation consistency (WAC) regularization*—a minimax formulation that not only incorporates the consistency regularization but also leverages the consistency regularization as a reference for reweighting the cross-entropy and the augmentation consistency terms corresponding to different samples. At the saddle point, the WAC regularization automatically separates samples from different label subpopulations by assigning all weight to the consistency terms for label-sparse samples, and all weight to the cross-entropy terms for label-dense samples.

Furthermore, as an algorithm for solving the minimax problem posed by the WAC regularization, we propose *AdaWAC*—an *adaptive weighting* scheme inspired by a mirror-descent-based algorithm for distributionally-robust optimization (Sagawa et al., 2020). By adaptively adjusting the weights between the cross-entropy and consistency terms of different samples, *AdaWAC* comes with both a convergence guarantee and empirical success.

Overall, we summarize the main contributions of this work as follows:

- We cast the discrepancy between the sparse and dense labels within each volume as a subpopulation shift in the conditional distribution $P(\mathbf{y}|\mathbf{x})$ (Section 2).
- We propose WAC regularization which uses the consistency of encoder layer outputs (in a UNet architecture) as a natural reference to incentivize separation between samples with sparse and dense labels (Section 3), along with an adaptive weighting algorithm—*AdaWAC*—for solving the WAC regularization problem with a convergence guarantee (Section 4).
- We empirically demonstrate the potency of *AdaWAC* not only in enhancing segmentation performance and sample efficiency but also in improving distributional robustness (Section 5).

1.1 RELATED WORK

Sample reweighting. Sample reweighting is a popular strategy for coping with subpopulation shifts in training data where different weights are assigned to samples from different subpopulations. In particular, the distributionally-robust optimization (DRO) (Ben-Tal et al., 2013; Duchi et al., 2016; Duchi & Namkoong, 2018; Sagawa et al., 2020) considers a collection of training sample groups from different distributions where, with the explicit grouping of samples, the goal is to minimize the worst-case loss over the groups. Without prior knowledge on sample grouping, importance sampling (Needell et al., 2014; Zhao & Zhang, 2015; Alain et al., 2015; Loshchilov & Hutter, 2015; Gopal, 2016; Katharopoulos & Fleuret, 2018), iterative trimming (Kawaguchi & Lu, 2020; Shen & Sanghavi, 2019), and empirical-loss-based reweighting (Wu et al., 2022) are commonly incorporated in the stochastic optimization process for adaptive reweighting and separation of samples from different subpopulations.

Consistency regularization. Consistency regularization (Bachman et al., 2014; Laine & Aila, 2016; Sohn et al., 2020; Berthelot et al., 2019) is a popular way to exploit data augmentations that encourage the model to learn the vicinity among augmentations of the same sample, with the assumption that data augmentations generally preserve the semantic information in data.

For medical imaging, consistency regularization is generally leveraged as a semi-supervised learning tool (Bortsova et al., 2019; Zhao et al., 2019; Li et al., 2020; Wang et al., 2021a; Zhang et al., 2021; Zhou et al., 2021; Basak et al., 2022). In efforts to incorporate consistency regularization in medical image segmentation with augmentation-sensitive labels, Li et al. (2020) encourages transformation consistency between predictions with augmentations applied to the image inputs and the segmentation outputs. Basak et al. (2022) penalizes inconsistent segmentation outputs between teacher-student models, with MixUp (Zhang et al., 2017) applied on image inputs of the teacher model and segmentation outputs of the student model. Instead of enforcing consistency in the segmentation output space as above, our algorithm leverages the insensitivity of sparse labels to augmentations and encourages consistent encodings (in the latent space of encoder outputs) on label-sparse samples.

2 PROBLEM SETUP

Notations. We denote $[K] = \{1, \dots, K\}$ for any $K \in \mathbb{N}$. For an arbitrary tensor, we adapt the syntax for Python slicing on the subscript (except counting from 1) to represent its elements and subtensors. For example, $\mathbf{x}_{[i,j]}$ denotes the (i, j)-entry of the two-dimensional tensor $\mathbf{x}$, and $\mathbf{x}_{[i,:]}$ denotes its i-th row. Let $\mathbb{I}$ be a function onto $\{0, 1\}$ such that, for any event e, $\mathbb{I}\{e\} = 1$ if e is true and 0 otherwise. For any distribution P and $n \in \mathbb{N}$, let P^n denote the joint distribution of n samples drawn *i.i.d.* from P. We refer to an event as happening with high probability (*w.h.p.*) if it takes place with probability $1 - \Omega\left(\text{poly}\left(n\right)\right)^{-1}$.

2.1 PIXEL-WISE CLASSIFICATION WITH SPARSE AND DENSE LABELS

We consider the volumetric medical image segmentation as a pixel-wise multi-class classification problem where we aim to learn a pixel-wise classifier $h : \mathcal{X} \to [K]^d$ that serves as a good approximation for the ground truth $h^* : \mathcal{X} \to [K]^d$.

Recall the separation of cross-entropy losses between samples with different fractions of non-background labels during training from Figure 1. We refer to a sample $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times [K]^d$ as *label-sparse* if most pixels in $\mathbf{y}$ are labeled as the background such that the cross-entropy loss on $(\mathbf{x}, \mathbf{y})$ converges rapidly in the early stage of training[1]. Otherwise, we say that $(\mathbf{x}, \mathbf{y})$ is *label-dense*. Formally, we describe such variation as a subpopulation shift in the label distribution.

Definition 1 (Mixture of label-sparse and label-dense distributions)**.** Let P_0 and P_1 be the distributions of *label-sparse and label-dense samples* with distinct conditional distributions $P_0\left(\mathbf{y}|\mathbf{x}\right)$

[1]Although the sparsity of non-background pixels in the segmentation label is a key feature of label-sparse samples (as the name suggests), the unknown cut-off on such sparsity degenerates it as a sufficient condition for the rapid convergence of cross-entropy loss (Figure 1). Instead of making distinction with the sparsity of non-background pixels, we formalize a natural separation between label-sparse and label-dense samples in Assumption 1, based on which our algorithm can distinguish different samples spontaneously.

and $P_1(\mathbf{y}|\mathbf{x})$, respectively, but the same marginal distribution $P(\mathbf{x})$ such that $P_i(\mathbf{x},\mathbf{y}) = P_i(\mathbf{y}|\mathbf{x})P(\mathbf{x})$ $(i = 0, 1)$. For $\xi \in [0,1]$, we define a data distribution P_ξ where each sample $(\mathbf{x},\mathbf{y}) \sim P_\xi$ is drawn either from P_1 with probability ξ or from P_0 with probability $1-\xi$.

We aim to learn a pixel-wise classifier from a function class $\mathcal{H} \ni h_\theta = \operatorname{argmax}_{k\in[K]} f_\theta(\mathbf{x})_{[j,:]}$ for all $j \in [d]$ where the underlying function $f_\theta \in \mathcal{F}$, parameterized by some $\theta \in \mathcal{F}_\theta$, admits an encoder-decoder structure:

$$\mathcal{F} = \left\{ f_\theta = \phi_\theta \circ \psi_\theta \;\middle|\; \phi_\theta : \mathcal{X} \to \mathcal{Z},\ \psi_\theta : \mathcal{Z} \to [0,1]^{d\times K} \right\}. \tag{1}$$

Here ϕ_θ, ψ_θ correspond to the encoder and decoder functions, respectively; $(\mathcal{F}_\theta, \langle\cdot,\cdot\rangle_{\mathcal{F}})$ denotes the inner product space of parameters with the induced norm $\|\cdot\|_{\mathcal{F}}$ and dual norm $\|\cdot\|_{\mathcal{F},*}$; $(\mathcal{Z}, \varrho)$ is a latent metric space.

To learn from segmentation labels, we consider the *averaged cross-entropy loss*:

$$\ell_{CE}(\theta;(\mathbf{x},\mathbf{y})) = -\frac{1}{d}\sum_{j=1}^{d}\sum_{k=1}^{K}\mathbb{I}\left\{\mathbf{y}_{[j]}=k\right\}\cdot\log\left(f_\theta(\mathbf{x})_{[j,k]}\right) = -\frac{1}{d}\sum_{j=1}^{d}\log\left(f_\theta(\mathbf{x})_{[j,\mathbf{y}_{[j]}]}\right). \tag{2}$$

We assume proper learning where there exists $\theta^* \in \bigcap_{\xi\in[0,1]} \operatorname{argmin}_{\theta\in\mathcal{F}_\theta} \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim P_\xi}\left[\ell_{CE}(\theta;(\mathbf{x},\mathbf{y}))\right]$ that is invariant with respect to ξ.

2.2 Augmentation Consistency Regularization

Despite the invariance of f_{θ^*} to P_ξ on the population loss, in practice we have a finite number of training samples and the predominance of label-sparse samples in the training set introduces difficulties due to the class imbalances. As a specific extreme scenario for the pixel-wise classifier with encoder-decoder structure (Equation (1)), when the label-sparse samples are predominant ($\xi \ll 1$), a decoder function ψ_θ that predicts every pixel as background can achieve near-optimal cross-entropy loss, regardless of what the encoder function ϕ_θ is, considerably compromising the test performance (*cf.* Table 1). To encourage legit encoding even in absence of sufficient dense labels, we leverage the unsupervised consistency regularization on the *encoder function* ϕ_θ based on data augmentations.

Let $\mathcal{A}$ be a distribution over transformations on $\mathcal{X}$ where for any $\mathbf{x} \in \mathcal{X}$, each $A \sim \mathcal{A}$ $(A : \mathcal{X} \to \mathcal{X})$ induces an augmentation $A(\mathbf{x})$ of $\mathbf{x}$ that perturbs low-level information in $\mathbf{x}$. We aim to learn an encoder function $\phi_\theta : \mathcal{X} \to \mathcal{Z}$ that is capable of filtering out low-level information from $\mathbf{x}$ and therefore provides similar encodings for augmentations of the same sample. Recalling the metric ϱ on $\mathcal{Z}$, for a given scaling hyperparameter $\lambda_{AC} > 0$, we measure the similarity between augmentations with a consistency regularization term on $\phi_\theta(\cdot)$: for any $A_1, A_2 \sim \mathcal{A}^2$,

$$\ell_{AC}(\theta;\mathbf{x},A_1,A_2) \triangleq \lambda_{AC}\cdot\varrho\Big(\phi_\theta(A_1(\mathbf{x})),\phi_\theta(A_2(\mathbf{x}))\Big). \tag{3}$$

For the n training samples $\{(\mathbf{x}_i,\mathbf{y}_i)\}_{i\in[n]} \sim P_\xi^n$, we consider n pairs of data augmentation transformations $\{(A_{i,1},A_{i,2})\}_{i\in[n]} \sim \mathcal{A}^{2n}$. In the basic version, we encourage the similar encoding $\phi_\theta(\cdot)$ of the augmentation pairs $(A_{i,1}(\mathbf{x}_i), A_{i,2}(\mathbf{x}_i))$ for all $i \in [n]$ via consistency regularization:

$$\min_{\theta\in\mathcal{F}_{\theta^*}(\gamma)} \frac{1}{n}\sum_{i=1}^{n}\ell_{CE}(\theta;(\mathbf{x}_i,\mathbf{y}_i)) + \ell_{AC}(\theta;\mathbf{x}_i,A_{i,1},A_{i,2}). \tag{4}$$

We enforce consistency on $\phi_\theta(\cdot)$ in light of the encoder-decoder architecture: the encoder is generally designed to abstract essential information and filters out low-level non-semantic perturbations (*e.g.*, those introduced by augmentations), while the decoder recovers the low-level information for the pixel-wise classification. Therefore, with different $A_1, A_2 \sim \mathcal{A}$, the encoder output $\phi_\theta(\cdot)$ tends to be more consistent than the other intermediate layers, especially for label-dense samples.

3 Weighted Augmentation Consistency (WAC) Regularization

As the motivation, we begin with a key observation about the averaged cross-entropy:

Remark 1 (Separation of averaged cross-entropy loss on P_0 and P_1). As demonstrated in Figure 1, the sparse labels from P_0 tend to be much easier to learn than the dense ones from P_1, leading to considerable separation of averaged cross-entropy losses on the sparse and dense labels after a sufficient number of training epochs – $\ell_{CE}(\theta;(\mathbf{x},\mathbf{y})) \ll \ell_{CE}(\theta;(\mathbf{x}',\mathbf{y}'))$ for most label-sparse samples $(\mathbf{x},\mathbf{y}) \sim P_0$ and label-dense samples $(\mathbf{x}',\mathbf{y}') \sim P_1$.

Although Equation (4) with consistency regularization alone can boost the segmentation accuracy during testing (*cf.* Table 4), it does not take the separation between label-sparse and label-dense samples into account. In Section 5, we will empirically demonstrate that proper exploitation of such separation, like the formulation introduced below, can bring compelling further improvements.

Concretely, we formalized the notion of separation between P_0 and P_1 based on the consistency regularization term (Equation (3)) with the following assumption [2].

Assumption 1 (N-separation between P_0 and P_1)**.** Given a sufficiently small $\gamma > 0$, let $\mathcal{F}_{\theta^*}(\gamma) = \{\theta \in \mathcal{F}_\theta \mid \|\theta - \theta^*\|_{\mathcal{F}} \leq \gamma\}$ be a compact and convex neighborhood of well-trained pixel-wise classifiers[3]. We say that P_0 *and* P_1 *are* N*-separated over* $\mathcal{F}_{\theta^*}(\gamma)$ if there exists $\omega > 0$ such that both:

(i) $\ell_{CE}(\theta;(\mathbf{x},\mathbf{y})) < \ell_{AC}(\theta;\mathbf{x},A_1,A_2)$ for all $\theta \in \mathcal{F}_{\theta^*}(\gamma)$ given $(\mathbf{x},\mathbf{y}) \sim P_0$
(ii) $\ell_{CE}(\theta;(\mathbf{x},\mathbf{y})) > \ell_{AC}(\theta;\mathbf{x},A_1,A_2)$ for all $\theta \in \mathcal{F}_{\theta^*}(\gamma)$ given $(\mathbf{x},\mathbf{y}) \sim P_1$

hold with probability $1 - \Omega\left(N^{1+\omega}\right)^{-1}$ over $((\mathbf{x},\mathbf{y}),(A_1,A_2)) \sim P_\xi \times \mathcal{A}^2$.

This assumption is motivated by the empirical observation that the perturbation in $\phi_\theta(\cdot)$ induced by $\mathcal{A}$ is more uniform across P_0 and P_1 than the averaged cross-entropy, as instantiated in Figure 3.

Under Assumption 1, up to a proper scaling hyperparameter λ_{AC}, the consistency regularization (Equation (3)) can separate the averaged cross-entropy loss (Equation (2)) on N label-sparse and label-dense samples with probability $1 - \Omega(N^\omega)^{-1}$ by the union bound (as explained formally in Appendix A). In particular, the larger N correspond to the stronger separation between P_0 and P_1.

With Assumption 1, we introduce a minimax formulation that incentivizes the separation of label-sparse and label-dense samples automatically by introducing a flexible weight $\boldsymbol{\beta}_{[i]} \in [0,1]$ that balances $\ell_{CE}(\theta;(\mathbf{x}_i,\mathbf{y}_i))$ and $\ell_{AC}(\theta;\mathbf{x}_i,A_{i,1},A_{i,2})$ for each of the n samples.

$$\widehat{\theta}^{WAC}, \widehat{\boldsymbol{\beta}} \in \underset{\theta\in\mathcal{F}_{\theta^*}(\gamma)}{\operatorname{argmin}} \underset{\boldsymbol{\beta}\in[0,1]^n}{\operatorname{argmax}} \left\{ \widehat{L}^{WAC}(\theta,\boldsymbol{\beta}) \triangleq \frac{1}{n}\sum_{i=1}^{n} \widehat{L}_i^{WAC}(\theta,\boldsymbol{\beta}) \right\} \tag{5}$$
$$\widehat{L}_i^{WAC}(\theta,\boldsymbol{\beta}) \triangleq \boldsymbol{\beta}_{[i]} \cdot \ell_{CE}(\theta;(\mathbf{x}_i,\mathbf{y}_i)) + (1-\boldsymbol{\beta}_{[i]}) \cdot \ell_{AC}(\theta;\mathbf{x}_i,A_{i,1},A_{i,2}).$$

With convex and continuous loss and regularization terms (formally in Proposition 1), Equation (5) has a saddle point where $\widehat{\boldsymbol{\beta}}$ separates the label-sparse and label-dense samples under Assumption 1.

Proposition 1 (Formal proof in Appendix A)**.** *Assume that* $\ell_{CE}(\theta;(\mathbf{x},\mathbf{y}))$ *and* $\ell_{AC}(\theta;\mathbf{x},A_1,A_2)$ *are convex and continuous in* θ *for all* $(\mathbf{x},\mathbf{y}) \in \mathcal{X}\times[K]^d$ *and* $A_1,A_2 \sim \mathcal{A}^2$*;* $\mathcal{F}_{\theta^*}(\gamma) \subset \mathcal{F}_\theta$ *is compact and convex. If* P_0 *and* P_1 *are* n*-separated (Assumption 1), then there exists* $\widehat{\boldsymbol{\beta}} \in \{0,1\}^n$ *and* $\widehat{\theta}^{WAC} \in \operatorname{argmin}_{\theta\in\mathcal{F}_{\theta^*}(\gamma)} \widehat{L}^{WAC}\left(\theta,\widehat{\boldsymbol{\beta}}\right)$ *such that*

$$\min_{\theta\in\mathcal{F}_{\theta^*}(\gamma)} \widehat{L}^{WAC}\left(\theta,\widehat{\boldsymbol{\beta}}\right) = \widehat{L}^{WAC}\left(\widehat{\theta}^{WAC},\widehat{\boldsymbol{\beta}}\right) = \max_{\boldsymbol{\beta}\in[0,1]^n} \widehat{L}^{WAC}\left(\widehat{\theta}^{WAC},\boldsymbol{\beta}\right). \tag{6}$$

Further, $\widehat{\boldsymbol{\beta}}$ *separates the label-sparse and label-dense samples—*$\boldsymbol{\beta}_{[i]} = \mathbb{I}\{(\mathbf{x}_i,\mathbf{y}_i) \sim P_1\}$—w.h.p..

In other words, for n samples drawn from a mixture of the n-separated P_0 and P_1, at the saddle point, Equation (5) automatically identifies the label-sparse samples with $\boldsymbol{\beta}_{[i]} = 0$, learning more from the unsupervised consistency regularization, and the label-dense ones with $\boldsymbol{\beta}_{[i]} = 1$, emphasizing more on the supervised averaged cross-entropy loss.

[2] We note that although Assumption 1 can be rather strong, it is only required for the separation guarantee of label-sparse and label-dense samples with high probability in Proposition 1, but not for the adaptive weighting algorithm introduced in Section 4 or in practice for the experiments.

[3] With pretrained initialization, we assume that the optimization algorithm is always probing in $\mathcal{F}_{\theta^*}(\gamma)$.

4 ADAPTIVELY WEIGHTED AUGMENTATION CONSISTENCY (*AdaWAC*)

Remark 2 (Connection to hard thresholding algorithms). The saddle point of Equation (5) is closely related to hard thresholding algorithms like Ordered SGD (Kawaguchi & Lu, 2020) and iterative trimmed loss (Shen & Sanghavi, 2019). In each iteration, these algorithms update the model only on a proper subset of training samples based on the (ranking of) current empirical risks. Compared to hard thresholding algorithms, (i) Equation (5) additionally leverages the unused samples (*e.g.*, label-sparse samples) for unsupervised consistency regularization on data augmentations, which is known for improving generalization and feature learning even in supervised settings (Yang et al., 2022; Shen et al., 2022); (ii) meanwhile, it does not require prior knowledge of the sample subpopulations (*e.g.*, ξ for P_ξ) which is essential for hard thresholding algorithms.

Equation (5) further facilitates the more flexible optimization process. As we will empirically show in Table 2, despite the close relation between Equation (5) and the hard thresholding algorithms (Remark 2), such updating strategies may be suboptimal for solving Equation (5).

Algorithm 1 Adaptively Weighted Augmentation Consistency (*AdaWAC*)

Input: Training samples $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i\in[n]} \sim P_\xi^n$, augmentations $\{(A_{i,1}, A_{i,2})\}_{i\in[n]} \sim \mathcal{A}^{2n}$, maximum number of iterations $T \in \mathbb{N}$, learning rates $\eta_\theta, \eta_{\boldsymbol{\beta}} > 0$, pretrained initialization for the pixel-wise classifier $\theta_0 \in \mathcal{F}_{\theta^*}(\gamma)$.
Initialize the sample weights $\boldsymbol{\beta}_0 = \mathbf{1}/2 \in [0,1]^n$.
for $t = 1, \ldots, T$ **do**
  Sample $i_t \sim [n]$ uniformly
  $\mathbf{b} \leftarrow \left[(\boldsymbol{\beta}_{t-1})_{[i_t]}, 1 - (\boldsymbol{\beta}_{t-1})_{[i_t]}\right]$
  $\mathbf{b}_{[1]} \leftarrow \mathbf{b}_{[1]} \cdot \exp\left(\eta_{\boldsymbol{\beta}} \cdot \ell_{CE}\left(\theta_{t-1}; (\mathbf{x}_{i_t}, \mathbf{y}_{i_t})\right)\right)$
  $\mathbf{b}_{[2]} \leftarrow \mathbf{b}_{[2]} \cdot \exp\left(\eta_{\boldsymbol{\beta}} \cdot \ell_{AC}\left(\theta_{t-1}; \mathbf{x}_{i_t}, A_{i_t,1}, A_{i_t,2}\right)\right)$
  $\boldsymbol{\beta}_t \leftarrow \boldsymbol{\beta}_{t-1}, (\boldsymbol{\beta}_t)_{[i_t]} \leftarrow \mathbf{b}_{[1]} / \|\mathbf{b}\|_1$
  $\theta_t \leftarrow \theta_{t-1} - \eta_\theta \cdot \Big((\boldsymbol{\beta}_t)_{[i_t]} \cdot \nabla_\theta \ell_{CE}\left(\theta_{t-1}; (\mathbf{x}_{i_t}, \mathbf{y}_{i_t})\right)$
    $+ \left(1 - (\boldsymbol{\beta}_t)_{[i_t]}\right) \cdot \nabla_\theta \ell_{AC}\left(\theta_{t-1}; \mathbf{x}_{i_t}, A_{i_t,1}, A_{i_t,2}\right) \Big)$
end for

Inspired by the breakthrough made by Sagawa et al. (2020) in the distributionally-robust optimization (DRO) setting where gradient updating on weights is shown to enjoy better convergence guarantees than hard thresholding, in Algorithm 1, we introduce an adaptive weighting algorithm for solving Equation (5) based on online mirror descent. In contrast to the commonly used stochastic gradient descent (SGD), the flexibility of online mirror descent in choosing the associated norm space not only allows gradient updates on sample weights, but also grants distinct learning dynamics to sample weights $\boldsymbol{\beta}_t$ and model parameters θ_t, which leads to the following convergence guarantee.

Proposition 2 (Formally in Proposition 3, proof in Appendix B, assumptions instantiated in Example 1)**.** *Assume that $\ell_{CE}(\theta; (\mathbf{x}, \mathbf{y}))$ and $\ell_{AC}(\theta; \mathbf{x}, A_1, A_2)$ are convex and continuous in θ for all $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times [K]^d$ and $A_1, A_2 \sim \mathcal{A}^2$; $\mathcal{F}_{\theta^*}(\gamma) \subset \mathcal{F}_\theta$ is convex and compact. If there exist* [4] *(i) $C_{\theta,*} > 0$ such that $\frac{1}{n}\sum_{i=1}^n \left\|\nabla_\theta \widehat{L}_i^{WAC}(\theta, \boldsymbol{\beta})\right\|_{\mathcal{F},*}^2 \le C_{\theta,*}^2$ and (ii) $C_{\boldsymbol{\beta},*} > 0$ such that $\frac{1}{n}\sum_{i=1}^n \max\left\{\ell_{CE}(\theta; (\mathbf{x}_i, \mathbf{y}_i)), \ell_{AC}(\theta; \mathbf{x}_i, A_{i,1}, A_{i,2})\right\}^2 \le C_{\boldsymbol{\beta},*}^2$ for all $\theta \in \mathcal{F}_{\theta^*}(\gamma)$, $\boldsymbol{\beta} \in [0,1]^n$, then with $\eta_\theta = \eta_{\boldsymbol{\beta}} = \frac{2}{\sqrt{5T\left(\gamma^2 C_{\theta,*}^2 + 2nC_{\boldsymbol{\beta},*}^2\right)}}$, Algorithm 1 provides*

$$\mathbb{E}\left[\max_{\boldsymbol{\beta}\in[0,1]^n} \widehat{L}^{WAC}\left(\overline{\theta}_T, \boldsymbol{\beta}\right) - \min_{\theta\in\mathcal{F}_{\theta^*}(\gamma)} \widehat{L}^{WAC}\left(\theta, \overline{\boldsymbol{\beta}}_T\right)\right] \le 2\sqrt{5\left(\gamma^2 C_{\theta,*}^2 + 2nC_{\boldsymbol{\beta},*}^2\right)/T}$$

where $\overline{\theta}_T = \frac{1}{T}\sum_{t=1}^T \theta_t$ and $\overline{\boldsymbol{\beta}}_T = \frac{1}{T}\sum_{t=1}^T \boldsymbol{\beta}_t$.

[4] Following the convention, we use $*$ in subscription to denote the dual spaces. For instance, recalling the parameter space $\mathcal{F}_\theta$ characterized by the norm $\|\cdot\|_{\mathcal{F}}$ from Section 2.1, we use $\|\cdot\|_{\mathcal{F},*}$ to denote its dual norm; while $C_{\theta,*}, C_{\boldsymbol{\beta},*}$ upper bound the dual norms of the gradients with respect to θ and $\boldsymbol{\beta}$.

In addition to the convergence guarantee, Algorithm 1 also demonstrates superior performance over hard thresholding algorithms for segmentation problems in practice (Table 2). An intuitive explanation is that instead of filtering out all the label-sparse samples via hard thresholding, the adaptive weighting allows the model to learn from some sparse labels at the early epochs, while smoothly down-weighting ℓ_{CE} of these samples since learning sparse labels tends to be easier (Remark 1). With the learned model tested on a mixture of label-sparse and label-dense samples, learning sparse labels at the early stage is crucial for accurate segmentation.

5 EXPERIMENTS

In this section, we investigate the proposed *AdaWAC* algorithm (Algorithm 1) on different medical image segmentation tasks with different UNet-like architectures. We first demonstrate the performance improvements brought by *AdaWAC* in terms of sample efficiency and robustness to subpopulation shift (Table 1). Then, we verify the empirical advantage of *AdaWAC* compared to the closely related hard thresholding algorithms as discussed in Remark 2 (Table 2). Our ablation study (Table 4) further illustrates the indispensability of both sample reweighting and consistency regularization, the deliberate combination of which leads to the superior performance of *AdaWAC*[5].

Experiment setup. We conduct experiments on two volumetric medical image segmentation tasks: abdominal CT segmentation for Synapse multi-organ dataset (Synapse)[6] and cine-MRI segmentation for Automated cardiac diagnosis challenge dataset (ACDC)[7], with two UNet-like architectures: TransUNet (Chen et al., 2021) and UNet Ronneberger et al. (2015) (deferred to Appendix E.1). For the main experiments with TransUNet in Section 5, we follow the official implementation in (Chen et al., 2021) and use ERM+SGD as the baseline. We evaluate segmentations with two standard metrics—the average Dice-similarity coefficient (DSC) and average 95-percentile of Hausdorff distance (HD95). Dataset and implementation details are deferred to Appendix D. Given the sensitivity of medical image semantics to perturbations, our experiments only involve simple augmentations (*i.e.*, rotation and mirroring) adapted from (Chen et al., 2021).

5.1 SEGMENTATION PERFORMANCE OF *AdaWAC* WITH TRANSUNET

Segmentation on Synapse. Figure 2 visualizes the segmentation predictions on 6 Synapse test slices given by models trained via *AdaWAC* (ours) and via the baseline (ERM+SGD) with TransUNet (Chen et al., 2021). We observe that *AdaWAC* provides more accurate predictions on the segmentation boundaries and captures small organs better than the baseline.

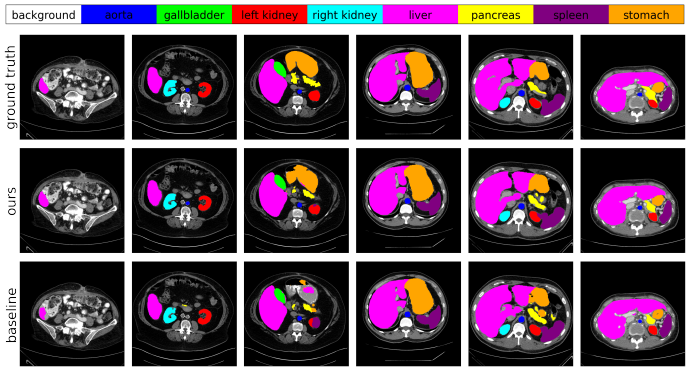

Figure 2: Visualization of segmentation predictions against the ground truth (in grayscale) on Synapse. Top to bottom: ground truth, ours (*AdaWAC*), baseline.

[5]We release our code anonymously at https://anonymous.4open.science/r/adawac-F5F8.
[6]https://www.synapse.org/#!Synapse:syn3193805/wiki/217789
[7]https://www.creatis.insa-lyon.fr/Challenge/acdc/

Visualization of *AdaWAC*. As shown in Figure 3, with $\ell_{CE}(\theta_t; (\mathbf{x}_i, \mathbf{y}_i))$ (Equation (2)) of label-sparse versus label-dense slices weakly separated in the early epochs, the model further learns to distinguish $\ell_{CE}(\theta_t; (\mathbf{x}_i, \mathbf{y}_i))$ of label-sparse/label-dense slices during training. By contrast, $\ell_{AC}(\theta_t; \mathbf{x}_i, A_{i,1}, A_{i,2})$ (Equation (3)) remains mixed for all the slices in the entire training process. As a result, the CE weights of label-sparse slices are much smaller than those of label-dense ones, pushing *AdaWAC* to learn more image representations but less pixel classification for slices with sparse labels and learn more pixel classification for slices with dense labels.

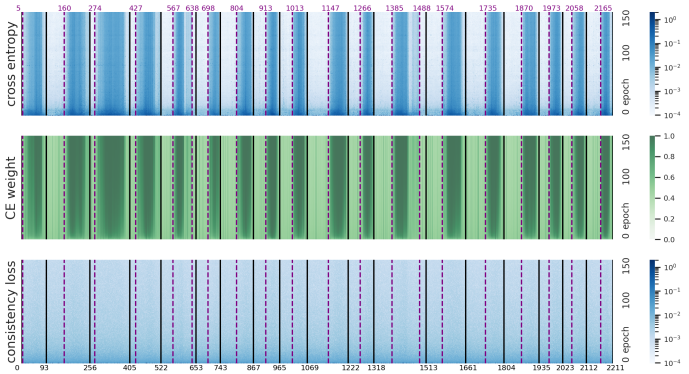

Figure 3: $\ell_{CE}(\theta_t; (\mathbf{x}_i, \mathbf{y}_i))$ (top), CE weights $\boldsymbol{\beta}_t$ (middle), and $\ell_{AC}(\theta_t; \mathbf{x}_i, A_{i,1}, A_{i,2})$ (bottom) of the entire Synapse training process. The x-axis indices slices 0–2211. The y-axis enumerates epochs 0–150. Individual volumes (cases) are partitioned by black lines; while the purple lines separate slices with/without non-background pixels.

Sample efficiency and robustness. We first demonstrate the *sample efficiency* of *AdaWAC* in comparison to the baseline (ERM+SGD) when training only on different subsets of the full Synapse training set ("**full**" in Table 1). Specifically, (i) **half-slice** contains slices with even indices only in each volume[8]; (ii) **half-vol** consists of 9 volumes uniformly sampled from the total 18 volumes in **full** where different volumes tend to have distinct ξs (*i.e.*, ratios of label-dense samples); (iii) **half-sparse** takes the first half slices in each volume, most of which tend to be label-sparse (*i.e.*, ξs are made to be small). As shown in Table 1, the model trained with *AdaWAC* on **half-slice** generalizes as well as a baseline model trained on **full**, if not better. Moreover, the **half-vol** and **half-sparse** experiments illustrate the *robustness* of *AdaWAC* to subpopulation shift. Furthermore, such sample efficiency and distributional robustness of *AdaWAC* extend to the more widely used UNet architecture. We defer the detailed results and discussions on UNet to Appendix E.1.

Table 1: *AdaWAC* with TransUNet trained on the full Synapse and its subsets.

Training	Method	DSC ↑	HD95 ↓	Aorta	Gallbladder	Kidney (L)	Kidney (R)	Liver	Pancreas	Spleen	Stomach
full	baseline	75.94 ± 0.68	32.91 ± 8.80	87.16	54.70	81.04	74.37	93.99	57.34	84.25	74.66
	AdaWAC	$\mathbf{78.83 \pm 0.38}$	$\mathbf{27.50 \pm 1.88}$	87.65	55.96	82.89	80.21	93.97	61.40	89.57	79.01
half-slice	baseline	74.62 ± 0.78	31.62 ± 8.37	86.14	44.23	79.09	78.46	93.50	55.78	84.54	75.24
	AdaWAC	$\mathbf{77.37 \pm 0.40}$	$\mathbf{29.56 \pm 1.09}$	86.89	55.96	82.15	78.63	94.34	57.36	86.60	77.05
half-vol	baseline	71.08 ± 0.90	46.83 ± 2.91	84.38	46.71	78.19	74.55	92.02	48.03	76.28	68.47
	AdaWAC	$\mathbf{73.81 \pm 0.94}$	$\mathbf{35.33 \pm 0.92}$	84.37	48.14	80.32	77.39	93.23	52.78	83.50	70.79
half-sparse	baseline	31.74 ± 2.78	69.72 ± 1.37	65.71	8.33	59.46	51.59	51.18	10.72	6.92	0.00
	AdaWAC	$\mathbf{41.03 \pm 2.12}$	$\mathbf{59.04 \pm 12.32}$	71.27	8.33	69.14	63.09	64.29	17.74	30.77	3.57

Comparison with hard thresholding algorithms. Table 2 illustrates the empirical advantage of *AdaWAC* over the hard thresholding algorithms, as suggested in Remark 2. In particular, we consider the following hard thresholding algorithms: (i) **trim-train** learns only from slices with at least one non-background pixel and trims the rest in each iteration on the fly; (ii) **trim-ratio** ranks the cross-entropy loss $\ell_{CE}(\theta_t; (\mathbf{x}_i, \mathbf{y}_i))$ in each iteration (mini-batch) and trims samples with the lowest cross-entropy losses at a fixed ratio – the ratio of all-background slices in the full training set

[8]Such sampling is equivalent to doubling the time interval between two consecutive scans or halving the scanning frequency in practice, resulting in the halving of sample size.

($1 - \frac{1280}{2211} \approx 0.42$); (iii) **ACR** further incorporates the augmentation consistency regularization directly via addition of $\ell_{AC}(\theta_t; \mathbf{x}_i, A_{i,1}, A_{i,2})$ without reweighting; (iv) **pseudo-*AdaWAC*** simulates the sample weights $\boldsymbol{\beta}$ at the saddle point and learns via $\ell_{CE}(\theta_t; (\mathbf{x}_i, \mathbf{y}_i))$ on slices with at least one non-background pixel while via $\ell_{AC}(\theta_t; \mathbf{x}_i, A_{i,1}, A_{i,2})$ otherwise. We notice that naive incorporation of **ACR** brings less observable boosts to the hard-thresholding methods. Therefore, the deliberate combination via reweighting in *AdaWAC* is essential for performance improvement.

Table 2: *AdaWAC* versus hard thresholding algorithms with TransUNet on Synapse.

Method	baseline	trim-train		trim-ratio		pseudo-*AdaWAC*	*AdaWAC*
			+ACR		+ACR		
DSC ↑	76.28	76.01	75.66	74.26	77.98	78.28	**79.12**
HD95 ↓	29.23	26.94	35.06	28.59	33.59	29.06	**25.79**

Segmentation on ACDC. Performance improvements granted by *AdaWAC* are also observed on the ACDC dataset (Table 3). We defer detailed visualization of ACDC segmentation to Appendix E.

Table 3: *AdaWAC* with TransUNet trained on ACDC.

Method	DSC ↑	HD95 ↓	RV	Myo	LV
TransUNet	89.36	3.02	88.36	83.84	95.87
AdaWAC (ours)	**90.41**	**1.29**	**89.50**	**85.78**	**95.95**

5.2 Ablation Study

On the influence of consistency regularization. To illustrate the role of consistency regularization in *AdaWAC*, we consider the **reweight-only** scenario with $\lambda_{AC} = 0$ such that $\ell_{AC}(\theta_t; \mathbf{x}_i, A_{i,1}, A_{i,2}) \equiv 0$ and therefore $\mathbf{b}_{[2]}$ (Algorithm 1 line 7) remains intact. With zero consistency regularization in *AdaWAC*, reweighting alone brings little improvement (Table 4).

On the influence of sample reweighting. We then investigate the effect of sample reweighting under different reweighting learning rates $\eta_{\boldsymbol{\beta}}$ (recall Algorithm 1): (i) **ACR-only** for $\eta_{\boldsymbol{\beta}} = 0$ (equivalent to the naive addition of $\ell_{AC}(\theta_t; \mathbf{x}_i, A_{i,1}, A_{i,2})$), (ii) ***AdaWAC*-0.01** for $\eta_{\boldsymbol{\beta}} = 0.01$, and (iii) ***AdaWAC*-1.0** for $\eta_{\boldsymbol{\beta}} = 1.0$. As Table 4 implies, when removing reweighting from *AdaWAC*, augmentation consistency regularization alone improves DSC slightly from 76.28 (baseline) to 77.89 (ACR-only), whereas *AdaWAC* boosts DSC to 79.12 (*AdaWAC*-1.0) with a proper choice of $\eta_{\boldsymbol{\beta}}$.

Table 4: Ablation study of *AdaWAC* with TransUNet trained on Synapse.

Method	DSC ↑	HD95 ↓	Aorta	Gallbladder	Kidney (L)	Kidney (R)	Liver	Pancreas	Spleen	Stomach
baseline	76.28	29.23	87.46	55.21	82.06	77.76	94.10	54.06	85.07	74.54
reweight-only	76.68	29.24	86.15	53.98	82.96	80.28	93.42	55.86	85.29	75.49
ACR-only	77.89	31.65	87.96	54.34	81.79	80.21	94.52	60.41	88.07	75.83
AdaWAC-0.01	77.94	27.81	87.58	52.75	82.29	80.22	94.90	55.92	91.63	78.23
AdaWAC-1.0	**79.12**	**25.79**	87.23	54.94	84.58	81.69	94.62	58.29	90.63	81.01

6 Discussion

In this paper, we exploit the non-uniformity in labels commonly observed in volumetric medical image segmentation via *AdaWAC*—a deliberate combination of adaptive weighting and augmentation consistency regularization. By casting the separation between sparse and dense segmentation labels as a subpopulation shift in the label distribution, we leverage the unsupervised consistency regularization on encoder layer outputs (of UNet architectures) as a natural reference to distinguish label-sparse and label-dense samples via comparisons against the supervised average cross-entropy losses. We formulate such comparisons as a weighted augmentation consistency (WAC) regularization problem and propose an adaptive weighting scheme—*AdaWAC*—for iterative and smooth separation of samples from different subpopulations with a convergence guarantee. Our experiments demonstrate empirical effectiveness of *AdaWAC* not only in improving segmentation performance and sample efficiency but also in enhancing robustness to the subpopulation shift in labels.

REFERENCES

Guillaume Alain, Alex Lamb, Chinnadhurai Sankar, Aaron Courville, and Yoshua Bengio. Variance reduction in sgd by distributed importance sampling. *arXiv preprint arXiv:1511.06481*, 2015.

Philip Bachman, Ouais Alsharif, and Doina Precup. Learning with pseudo-ensembles. *Advances in neural information processing systems*, 27:3365–3373, 2014.

Hritam Basak, Rajarshi Bhattacharya, Rukhshanda Hussain, and Agniv Chatterjee. An embarrassingly simple consistency regularization method for semi-supervised medical image segmentation. *arXiv preprint arXiv:2202.00677*, 2022.

Aharon Ben-Tal, Dick den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013. ISSN 00251909, 15265501.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019.

Dimitri Bertsekas. *Convex optimization theory*, volume 1. Athena Scientific, 2009.

Gerda Bortsova, Florian Dubost, Laurens Hogeweg, Ioannis Katramados, and Marleen de Bruijne. Semi-supervised medical image segmentation via learning consistency under transformations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 810–818. Springer, 2019.

Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv preprint arXiv:2105.05537*, 2021.

Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

John Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *arXiv preprint arXiv:1810.08750*, 2018.

John Duchi, Peter Glynn, and Hongseok Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *arXiv preprint arXiv:1610.03425*, 2016.

Siddharth Gopal. Adaptive sampling for sgd by exploiting side information. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pp. 364–372. JMLR.org, 2016.

Guy Hacohen and Daphna Weinshall. On the power of curriculum learning in training deep networks. In *International Conference on Machine Learning*, pp. 2535–2544. PMLR, 2019.

Jamie Haddock, Deanna Needell, Elizaveta Rebrova, and William Swartworth. Quantile-based Iterative Methods for Corrupted Systems of Linear Equations. *arXiv:2009.08089 [cs, math]*, September 2020. arXiv: 2009.08089.

Pavel Iakubovskii. Segmentation models pytorch. https://github.com/qubvel/segmentation_models.pytorch, 2019.

Angelos Katharopoulos and François Fleuret. Not all samples are created equal: Deep learning with importance sampling. In *International conference on machine learning*, pp. 2525–2534. PMLR, 2018.

Kenji Kawaguchi and Haihao Lu. Ordered sgd: A new stochastic optimization framework for empirical risk minimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 669–679. PMLR, 2020.

Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.

Xiaomeng Li, Lequan Yu, Hao Chen, Chi-Wing Fu, Lei Xing, and Pheng-Ann Heng. Transformation-consistent self-ensembling model for semisupervised medical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2):523–534, 2020.

Ilya Loshchilov and Frank Hutter. Online batch selection for faster training of neural networks. *arXiv preprint arXiv:1511.06343*, 2015.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pp. 565–571, 2016.

Deanna Needell, Rachel Ward, and Nati Srebro. Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. *Advances in neural information processing systems*, 27, 2014.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

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*, pp. 234–241. Springer, 2015.

Stuart J Russell and Peter Norvig. Artificial intelligence: a modern approach. malaysia, 2016.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020.

Vatsal Shah, Xiaoxia Wu, and Sujay Sanghavi. Choosing the sample with lowest loss makes sgd robust. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2120–2130, Online, 26–28 Aug 2020. PMLR.

Ruoqi Shen, Sebastien Bubeck, and Suriya Gunasekar. Data augmentation as feature manipulation. In *International Conference on Machine Learning*, pp. 19773–19808. PMLR, 2022.

Yanyao Shen and Sujay Sanghavi. Learning with bad training data via iterative trimmed loss minimization. In *International Conference on Machine Learning*, pp. 5739–5748. PMLR, 2019.

Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171 – 176, 1958.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

Saeid Asgari Taghanaki, Kumar Abhishek, Joseph Paul Cohen, Julien Cohen-Adad, and Ghassan Hamarneh. Deep semantic segmentation of natural and medical images: A review, 2019a.

Saeid Asgari Taghanaki, Yefeng Zheng, S Kevin Zhou, Bogdan Georgescu, Puneet Sharma, Daguang Xu, Dorin Comaniciu, and Ghassan Hamarneh. Combo loss: Handling input and output imbalance in multi-organ segmentation. *Computerized Medical Imaging and Graphics*, 75: 24–33, 2019b.

Yuxing Tang, Xiaosong Wang, Adam P Harrison, Le Lu, Jing Xiao, and Ronald M Summers. Attention-guided curriculum learning for weakly supervised classification and localization of thoracic diseases on chest radiographs. In *International Workshop on Machine Learning in Medical Imaging*, pp. 249–258. Springer, 2018.

Jonathan G Tullis and Aaron S Benjamin. On the effectiveness of self-paced learning. *Journal of memory and language*, 64(2):109–118, 2011.

Xi Wang, Hao Chen, Huiling Xiang, Huangjing Lin, Xi Lin, and Pheng-Ann Heng. Deep virtual adversarial self-training with consistency regularization for semi-supervised medical image classification. *Medical image analysis*, 70:102010, 2021a.

Xin Wang, Yudong Chen, and Wenwu Zhu. A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021b.

Yisen Wang, Weiyang Liu, Xingjun Ma, James Bailey, Hongyuan Zha, Le Song, and Shu-Tao Xia. Iterative Learning with Open-set Noisy Labels. *arXiv:1804.00092 [cs]*, March 2018. arXiv: 1804.00092.

Ken C. L. Wong, Mehdi Moradi, Hui Tang, and Tanveer Syeda-Mahmood. 3d segmentation with exponential logarithmic loss for highly unbalanced object sizes. In Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pp. 612–619, Cham, 2018. Springer International Publishing. ISBN 978-3-030-00931-1.

Xiaoxia Wu, Yuege Xie, Simon Shaolei Du, and Rachel Ward. Adaloss: A computationally-efficient and provably convergent adaptive gradient method. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(8):8691–8699, Jun. 2022.

Shuo Yang, Yijun Dong, Rachel Ward, Inderjit S Dhillon, Sujay Sanghavi, and Qi Lei. Sample efficiency of data augmentation consistency regularization. *arXiv preprint arXiv:2202.12230*, 2022.

Michael Yeung, Evis Sala, Carola-Bibiane Schönlieb, and Leonardo Rundo. Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation. *Computerized Medical Imaging and Graphics*, 95:102026, 2022. ISSN 0895-6111.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Yi Zhang, Bin Zhou, Lei Chen, Yulin Wu, and Hongchao Zhou. Multi-transformation consistency regularization for semi-supervised medical image segmentation. In *2021 4th International Conference on Artificial Intelligence and Big Data (ICAIBD)*, pp. 485–489. IEEE, 2021.

Amy Zhao, Guha Balakrishnan, Fredo Durand, John V Guttag, and Adrian V Dalca. Data augmentation using learned transformations for one-shot medical image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8543–8553, 2019.

Peilin Zhao and Tong Zhang. Stochastic optimization with importance sampling for regularized loss minimization. In *international conference on machine learning*, pp. 1–9. PMLR, 2015.

Hong-Yu Zhou, Chengdi Wang, Haofeng Li, Gang Wang, Shu Zhang, Weimin Li, and Yizhou Yu. Ssmd: semi-supervised medical image detection with adaptive consistency and heterogeneous perturbation. *Medical Image Analysis*, 72:102117, 2021.

A SEPARATION OF LABEL-SPARSE AND LABEL-DENSE SAMPLES

Proof of Proposition 1. We first observe that, since $\ell_{CE}(\theta; (\mathbf{x}, \mathbf{y}))$ and $\ell_{AC}(\theta; \mathbf{x}, A_1, A_2)$ are convex and continuous in θ for all $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$ and $A_1, A_2 \sim \mathcal{A}^2$, for all $i \in [n]$, $\widehat{L}_i^{WAC}(\theta, \boldsymbol{\beta})$ is continuous, convex in θ, and affine (thus concave) in $\boldsymbol{\beta}$; and therefore so is $\widehat{L}^{WAC}(\theta, \boldsymbol{\beta})$. Then with the compact and convex domains $\theta \in \mathcal{F}_{\theta^*}(\gamma)$ and $\boldsymbol{\beta} \in [0,1]^n$, Sion's minimax theorem (Sion, 1958) suggests the minimax equality,

$$\min_{\theta \in \mathcal{F}_{\theta^*}(\gamma)} \max_{\boldsymbol{\beta} \in [0,1]^n} \widehat{L}^{WAC}(\theta, \boldsymbol{\beta}) = \max_{\boldsymbol{\beta} \in [0,1]^n} \min_{\theta \in \mathcal{F}_{\theta^*}(\gamma)} \widehat{L}^{WAC}(\theta, \boldsymbol{\beta}), \tag{7}$$

where $\inf, \sup$ can be replaced by $\min, \max$ respectively due to compactness of the domains.

Further, by the continuity and convexity-concavity of $\widehat{L}^{WAC}(\theta, \boldsymbol{\beta})$, the pointwise maximum $\max_{\boldsymbol{\beta} \in [0,1]^n} \widehat{L}^{WAC}(\theta, \boldsymbol{\beta})$ is lower semi-continuous and convex in θ; while the pointwise minimum $\min_{\theta \in \mathcal{F}_{\theta^*}(\gamma)} \widehat{L}^{WAC}(\theta, \boldsymbol{\beta})$ is upper semi-continuous and concave in $\boldsymbol{\beta}$. Then via Weierstrass' theorem (Bertsekas (2009), Proposition 3.2.1), there exist $\widehat{\theta}^{WAC} \in \mathcal{F}_{\theta^*}(\gamma)$ and $\widehat{\boldsymbol{\beta}} \in [0,1]^n$ that achieve the minimax optimal by minimizing $\max_{\boldsymbol{\beta} \in [0,1]^n} \widehat{L}^{WAC}(\theta, \boldsymbol{\beta})$ and maximizing $\min_{\theta \in \mathcal{F}_{\theta^*}(\gamma)} \widehat{L}^{WAC}(\theta, \boldsymbol{\beta})$. Along with Equation (7), such $\left(\widehat{\theta}^{WAC}, \widehat{\boldsymbol{\beta}}\right)$ provides a saddle point for Equation (5) (Bertsekas (2009), Proposition 3.4.1).

Next, we show via contradiction that there exists a saddle point with $\widehat{\boldsymbol{\beta}}$ attained on a vertex $\widehat{\boldsymbol{\beta}} \in \{0,1\}^n$. Suppose the opposite, then for any saddle point $\left(\widehat{\theta}^{WAC}, \widehat{\boldsymbol{\beta}}\right)$, there must be an $i \in [n]$ with $\widehat{\boldsymbol{\beta}}_{[i]} \in (0,1)$, where we have the following contradictions:

(i) If $\ell_{CE}\left(\widehat{\theta}^{WAC}; (\mathbf{x}_i, \mathbf{y}_i)\right) < \ell_{AC}\left(\widehat{\theta}^{WAC}; \mathbf{x}_i, A_{i,1}, A_{i,2}\right)$, decreasing $\widehat{\boldsymbol{\beta}}_{[i]} > 0$ to $\widehat{\boldsymbol{\beta}}'_{[i]} = 0$ leads to $\widehat{L}^{WAC}\left(\widehat{\theta}^{WAC}, \widehat{\boldsymbol{\beta}}'\right) > \widehat{L}^{WAC}\left(\widehat{\theta}^{WAC}, \widehat{\boldsymbol{\beta}}\right)$, contradicting Equation (6).

(ii) If $\ell_{CE}\left(\widehat{\theta}^{WAC}; (\mathbf{x}_i, \mathbf{y}_i)\right) > \ell_{AC}\left(\widehat{\theta}^{WAC}; \mathbf{x}_i, A_{i,1}, A_{i,2}\right)$, increasing $\widehat{\boldsymbol{\beta}}_{[i]} < 1$ to $\widehat{\boldsymbol{\beta}}'_{[i]} = 1$ again leads to $\widehat{L}^{WAC}\left(\widehat{\theta}^{WAC}, \widehat{\boldsymbol{\beta}}'\right) > \widehat{L}^{WAC}\left(\widehat{\theta}^{WAC}, \widehat{\boldsymbol{\beta}}\right)$, contradicting Equation (6).

(iii) If $\ell_{CE}\left(\widehat{\theta}^{WAC}; (\mathbf{x}_i, \mathbf{y}_i)\right) = \ell_{AC}\left(\widehat{\theta}^{WAC}; \mathbf{x}_i, A_{i,1}, A_{i,2}\right)$, $\widehat{\boldsymbol{\beta}}_{[i]}$ can be replaced with any value in $[0,1]$, including $0, 1$.

Therefore, there must be a saddle point $\left(\widehat{\theta}^{WAC}, \widehat{\boldsymbol{\beta}}\right)$ with $\widehat{\boldsymbol{\beta}} \in \{0,1\}^n$ such that

$$\boldsymbol{\beta}_{[i]} = \mathbb{I}\left\{\ell_{CE}\left(\widehat{\theta}^{WAC}; (\mathbf{x}_i, \mathbf{y}_i)\right) > \ell_{AC}\left(\widehat{\theta}^{WAC}; \mathbf{x}_i, A_{i,1}, A_{i,2}\right)\right\}.$$

Finally, it remains to show that *w.h.p.* over $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i \in [n]} \sim P_\xi^n$ and $\{(A_{i,1}, A_{i,2})\}_{i \in [n]} \sim \mathcal{A}^{2n}$,

(i) $\ell_{CE}\left(\widehat{\theta}^{WAC}; (\mathbf{x}_i, \mathbf{y}_i)\right) \leq \ell_{AC}\left(\widehat{\theta}^{WAC}; \mathbf{x}_i, A_{i,1}, A_{i,2}\right)$ for all $(\mathbf{x}_i, \mathbf{y}_i) \sim P_0$; and

(ii) $\ell_{CE}\left(\widehat{\theta}^{WAC}; (\mathbf{x}_i, \mathbf{y}_i)\right) > \ell_{AC}\left(\widehat{\theta}^{WAC}; \mathbf{x}_i, A_{i,1}, A_{i,2}\right)$ for all $(\mathbf{x}_i, \mathbf{y}_i) \sim P_1$;

which leads to $\boldsymbol{\beta}_{[i]} = \mathbb{I}\{(\mathbf{x}_i, \mathbf{y}_i) \sim P_1\}$ *w.h.p.* as desired. To illustrate this, we begin by observing that when P_0 and P_1 are n-separated (Assumption 1), since $\widehat{\theta}^{WAC} \in \mathcal{F}_{\theta^*}(\gamma)$, there exists some $\omega > 0$ such that for each $i \in [n]$,

$$\mathbb{P}\left[\ell_{CE}\left(\widehat{\theta}^{WAC}; (\mathbf{x}_i, \mathbf{y}_i)\right) < \ell_{AC}\left(\widehat{\theta}^{WAC}; \mathbf{x}_i, A_{i,1}, A_{i,2}\right) \;\middle|\; (\mathbf{x}_i, \mathbf{y}_i) \sim P_0\right] \geq 1 - \frac{1}{\Omega\left(n^{1+\omega}\right)},$$

and

$$\mathbb{P}\left[\ell_{CE}\left(\widehat{\theta}^{WAC}; (\mathbf{x}_i, \mathbf{y}_i)\right) > \ell_{AC}\left(\widehat{\theta}^{WAC}; \mathbf{x}_i, A_{i,1}, A_{i,2}\right) \;\middle|\; (\mathbf{x}_i, \mathbf{y}_i) \sim P_1\right] \geq 1 - \frac{1}{\Omega\left(n^{1+\omega}\right)}.$$

Therefore by the union bound on a set of n samples $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i\in[n]} \sim P_\xi^n$,

$$\mathbb{P}\left[\ell_{CE}\left(\widehat{\theta}^{WAC}; (\mathbf{x}_i, \mathbf{y}_i)\right) < \ell_{AC}\left(\widehat{\theta}^{WAC}; \mathbf{x}_i, A_{i,1}, A_{i,2}\right) \ \forall\ (\mathbf{x}_i, \mathbf{y}_i) \sim P_0\right] \geq 1 - \frac{1}{\Omega\left(n^\omega\right)}, \quad (8)$$

and

$$\mathbb{P}\left[\ell_{CE}\left(\widehat{\theta}^{WAC}; (\mathbf{x}_i, \mathbf{y}_i)\right) > \ell_{AC}\left(\widehat{\theta}^{WAC}; \mathbf{x}_i, A_{i,1}, A_{i,2}\right) \ \forall\ (\mathbf{x}_i, \mathbf{y}_i) \sim P_1\right] \geq 1 - \frac{1}{\Omega\left(n^\omega\right)}. \quad (9)$$

Applying the union bound again on Equation (8) and Equation (9), we have the desired condition holds with probability $1 - \Omega\left(n^\omega\right)^{-1}$, *i.e.*, *w.h.p.*. □

B CONVERGENCE OF *AdaWAC*

For any $d \in \mathbb{N}$, let $\Delta_d^n \triangleq \left\{[\boldsymbol{\beta}_1; \ldots; \boldsymbol{\beta}_n] \in [0,1]^{n\times d} \;\middle|\; \|\boldsymbol{\beta}_i\|_1 = 1 \ \forall\, i \in [n]\right\}$. Then Equation (5) can be reformulated as:

$$\widehat{\theta}^{WAC}, \widehat{\boldsymbol{\beta}} = \operatorname*{argmin}_{\theta\in\mathcal{F}_{\theta^*}(\gamma)} \operatorname*{argmax}_{\mathbf{B}\in\Delta_2^n} \left\{\widehat{L}^{WAC}(\theta, \mathbf{B}) \triangleq \frac{1}{n}\sum_{i=1}^n \widehat{L}_i^{WAC}(\theta, \mathbf{B})\right\}, \quad (10)$$
$$\widehat{L}_i^{WAC}(\theta, \mathbf{B}) \triangleq \mathbf{B}_{[i,1]} \cdot \ell_{CE}(\theta; (\mathbf{x}_i, \mathbf{y}_i)) + \mathbf{B}_{[i,2]} \cdot \ell_{AC}(\theta; \mathbf{x}_i, A_{i,1}, A_{i,2}).$$

Proposition 3 (Convergence (formal restatement of Proposition 2))**.** *Assume that $\ell_{CE}(\theta; (\mathbf{x}, \mathbf{y}))$ and $\ell_{AC}(\theta; \mathbf{x}, A_1, A_2)$ are convex and continuous in θ for all $(\mathbf{x}, \mathbf{y}) \in \mathcal{X}\times\mathcal{Y}$ and $A_1, A_2 \sim \mathcal{A}^2$; $\mathcal{F}_{\theta^*}(\gamma) \subset \mathcal{F}_\theta$ is convex and compact. If there exist*

(i) $C_{\theta,*} > 0$ such that $\frac{1}{n}\sum_{i=1}^n \left\|\nabla_\theta \widehat{L}_i^{WAC}(\theta, \mathbf{B})\right\|_{\mathcal{F},*}^2 \leq C_{\theta,*}^2$ for all $\theta \in \mathcal{F}_{\theta^*}(\gamma)$, $\mathbf{B} \in \Delta_2^n$ and

(ii) $C_{\mathbf{B},*} > 0$ such that $\frac{1}{n}\sum_{i=1}^n \max\left\{\ell_{CE}(\theta; (\mathbf{x}_i, \mathbf{y}_i)), \ell_{AC}(\theta; \mathbf{x}_i, A_{i,1}, A_{i,2})\right\}^2 \leq C_{\mathbf{B},*}^2$ for all $\theta \in \mathcal{F}_{\theta^*}(\gamma)$,

then with $\eta_\theta = \eta_{\mathbf{B}} = 2\Big/\sqrt{5T\left(\gamma^2 C_{\theta,*}^2 + 2nC_{\mathbf{B},*}^2\right)}$, Algorithm 1 provides the convergence guarantee for the duality gap $\mathcal{E}\left(\overline{\theta}_T, \overline{\mathbf{B}}_T\right) \triangleq \max_{\mathbf{B}\in\Delta_2^n} \widehat{L}^{WAC}\left(\overline{\theta}_T, \mathbf{B}\right) - \min_{\theta\in\mathcal{F}_{\theta^*}(\gamma)} \widehat{L}^{WAC}\left(\theta, \overline{\mathbf{B}}_T\right)$:

$$\mathbb{E}\left[\mathcal{E}\left(\overline{\theta}_T, \overline{\mathbf{B}}_T\right)\right] \leq 2\sqrt{\frac{5\left(\gamma^2 C_{\theta,*}^2 + 2nC_{\mathbf{B},*}^2\right)}{T}},$$

where $\overline{\theta}_T = \frac{1}{T}\sum_{t=1}^T \theta_t$ and $\overline{\mathbf{B}}_T = \frac{1}{T}\sum_{t=1}^T \mathbf{B}_t$.

Proof of Proposition 3. The proof is an application of the standard convergence guarantee for the online mirror descent on saddle point problems, as recapitulated in Lemma 1.

Specifically, for $\mathbf{B} \in \Delta_2^n$, we use the norm $\|\mathbf{B}\|_{1,2} \triangleq \sqrt{\sum_{i=1}^n \left(\sum_{j=1}^2 \left|\mathbf{B}_{[i,j]}\right|\right)^2}$ with its dual norm $\|\mathbf{B}\|_{1,2,*} \triangleq \sqrt{\sum_{i=1}^n \left(\max_{j\in[2]} \left|\mathbf{B}_{[i,j]}\right|\right)^2}$. We consider a mirror map $\varphi_{\mathbf{B}} : [0,1]^{n\times 2} \to \mathbb{R}$ such that $\varphi_{\mathbf{B}}(\mathbf{B}) = \sum_{i=1}^n \sum_{j=1}^2 \mathbf{B}_{[i,j]} \log \mathbf{B}_{[i,j]}$. We observe that, since $\mathbf{B}_{[i,:]}, \mathbf{B'}_{[i,:]} \in \Delta_2$ for all $i \in [n]$,

$$D_{\varphi_{\mathbf{B}}}(\mathbf{B}, \mathbf{B}') = \sum_{i=1}^n \sum_{j=1}^2 \mathbf{B}_{[i,j]} \log \frac{\mathbf{B}_{[i,j]}}{\mathbf{B'}_{[i,j]}} \geq \frac{1}{2}\sum_{i=1}^n \left(\sum_{j=1}^2 \left|\mathbf{B}_{[i,j]} - \mathbf{B'}_{[i,j]}\right|\right)^2 = \frac{1}{2}\|\mathbf{B} - \mathbf{B}'\|_{1,2}^2,$$

and therefore $\varphi_{\mathbf{B}}$ is 1-strongly convex with respect to $\|\cdot\|_{1,2}$. With such $\varphi_{\mathbf{B}}$, we have the associated Fenchel dual $\varphi_{\mathbf{B}}^*(\mathbf{G}) = \sum_{i=1}^n \log\left(\sum_{j=1}^2 \exp\left(\mathbf{G}_{[i,j]}\right)\right)$, along with the gradients

$$\nabla\varphi_{\mathbf{B}}(\mathbf{B})_{[i,j]} = 1 + \log \mathbf{B}_{[i,j]}, \quad \nabla\varphi_{\mathbf{B}}^*(\mathbf{G})_{[i,j]} = \frac{\exp\left(\mathbf{G}_{[i,j]}\right)}{\sum_{j=1}^2 \exp\left(\mathbf{G}_{[i,j]}\right)},$$

such that the mirror descent update on $\mathbf{B}$ is given by

$$\begin{aligned}(\mathbf{B}_{t+1})_{[i,j]} =& \nabla\varphi^*_{\mathbf{B}}\left(\nabla\varphi_{\mathbf{B}}(\mathbf{B}_t) - \eta_{\mathbf{B}} \cdot \nabla_{\mathbf{B}}\widehat{L}^{\textit{WAC}}_{i_t}(\theta_t, \mathbf{B}_t)\right) \\ =& \frac{(\mathbf{B}_t)_{[i,j]} \exp\left(\eta_{\mathbf{B}} \cdot \left(\nabla_{\mathbf{B}}\widehat{L}^{\textit{WAC}}_{i_t}(\theta_t, \mathbf{B}_t)\right)_{[i,j]}\right)}{\sum_{j=1}^{2}(\mathbf{B}_t)_{[i,j]} \exp\left(\eta_{\mathbf{B}} \cdot \left(\nabla_{\mathbf{B}}\widehat{L}^{\textit{WAC}}_{i_t}(\theta_t, \mathbf{B}_t)\right)_{[i,j]}\right)}.\end{aligned}$$

For $i_t \sim [n]$ uniformly, the stochastic gradient with respect to $\mathbf{B}$ satisfies that

$$\begin{aligned}&\mathbb{E}_{i_t\sim[n]}\left[\left\|\nabla_{\mathbf{B}}\widehat{L}^{\textit{WAC}}_{i_t}(\theta_t, \mathbf{B}_t)\right\|^2_{1,2,*}\right] \\ =& \frac{1}{n}\sum_{i_t=1}^{n}\max\left\{\ell_{\textit{CE}}(\theta_t; (\mathbf{x}_{i_t}, \mathbf{y}_{i_t})), \ell_{\textit{AC}}(\theta_t; \mathbf{x}_{i_t}, A_{i_t,1}, A_{i_t,2})\right\}^2 \leq C^2_{\mathbf{B},*}.\end{aligned}$$

Further, measuring in the distance induced by $\varphi_{\mathbf{B}}$, we have

$$R^2_{\Delta_2^n} \triangleq \max_{\mathbf{B}\in\Delta_2^n}\varphi_{\mathbf{B}}(\mathbf{B}) - \min_{\mathbf{B}\in\Delta_2^n}\varphi_{\mathbf{B}}(\mathbf{B}) = 0 - \sum_{i=1}^{n}\sum_{j=1}^{2}\frac{1}{2}\log\frac{1}{2} = n.$$

Meanwhile, for $\theta \in \mathcal{F}_{\theta^*}(\gamma)$, we consider the norm $\|\theta\|_{\mathcal{F}} \triangleq \sqrt{\langle\theta,\theta\rangle_{\mathcal{F}}}$ induced by the inner product that characterizes $\mathcal{F}_\theta$, with the associated dual norm $\|\cdot\|_{\mathcal{F},*}$. We use a mirror map $\varphi_\theta : \mathcal{F}_\theta \to \mathbb{R}$ such that $\varphi_\theta(\theta) = \frac{1}{2}\|\theta - \theta^*\|^2_{\mathcal{F}}$. By observing that

$$D_{\varphi_\theta}(\theta, \theta') = \frac{1}{2}\|\theta - \theta'\|^2_{\mathcal{F}} \quad \forall\, \theta, \theta' \in \mathcal{F}.$$

we have φ_θ being 1-strongly convex with respect to $\|\cdot\|_{\mathcal{F}}$. With the gradient of φ_θ, $\nabla\varphi_\theta(\theta) = \theta - \theta^*$, and that of its Fenchel dual $\nabla\varphi^*_\theta(g) = g + \theta^*$, at the $(t+1)$-th iteration, we have

$$\theta_{t+1} = \nabla\varphi^*_\theta\left(\nabla\varphi_\theta(\theta_t) - \eta_\theta \cdot \nabla_\theta\widehat{L}^{\textit{WAC}}_{i_t}(\theta_t, \mathbf{B}_{t+1})\right) = \theta_t - \eta_\theta \cdot \nabla_\theta\widehat{L}^{\textit{WAC}}_{i_t}(\theta_t, \mathbf{B}_{t+1}).$$

For $i_t \sim [n]$ uniformly, the stochastic gradient with respect to f satisfies that

$$\mathbb{E}_{i_t\sim[n]}\left[\left\|\nabla_\theta\widehat{L}^{\textit{WAC}}_{i_t}(\theta_t, \mathbf{B}_{t+1})\right\|^2_{\mathcal{F},*}\right] = \frac{1}{n}\sum_{i_t=1}^{n}\left\|\nabla_\theta\widehat{L}^{\textit{WAC}}_{i_t}(\theta_t, \mathbf{B}_{t+1})\right\|^2_{\mathcal{F},*} \leq C^2_{\theta,*}.$$

Further, in light of the definition of $\mathcal{F}_{\theta^*}(\gamma)$, since $\theta^* \in \mathcal{F}_{\theta^*}(\gamma)$, with $\theta^* = \operatorname{argmin}_{\theta\in\mathcal{F}_{\theta^*}(\gamma)}\varphi_\theta(\theta)$ and $\theta' = \operatorname{argmax}_{\theta\in\mathcal{F}_{\theta^*}(\gamma)}\varphi_\theta(\theta)$, we have

$$R^2_{\mathcal{F}_{\theta^*}(\gamma)} \triangleq \max_{\theta\in\mathcal{F}_{\theta^*}(\gamma)}\varphi_\theta(\theta) - \min_{\theta\in\mathcal{F}_{\theta^*}(\gamma)}\varphi_\theta(\theta) = \frac{1}{2}\|\theta' - \theta^*\|^2_{\mathcal{F}} \leq \frac{\gamma^2}{2}.$$

Finally, leveraging Lemma 1 completes the proof. □

We recall the standard convergence guarantee for online mirror descent on saddle point problems. In general, we consider a stochastic function $F : \mathcal{U}\times\mathcal{V}\times\mathcal{I} \to \mathbb{R}$ with the randomness of $F(u, v; i)$ on $i \in \mathcal{I}$. Overloading and notation $\mathcal{I}$ both as the distribution of i and as the support, we are interested in solving the saddle point problem on the expectation function

$$\min_{u\in\mathcal{U}}\max_{v\in\mathcal{V}} f(u, v) \quad \text{where} \quad f(u, v) \triangleq \mathbb{E}_{i\sim\mathcal{I}}[F(u, v; i)]. \tag{11}$$

Assumption 2. Assuming that the stochastic objective satisfies the following:

(i) For every $i \in \mathcal{I}$, $F(\cdot, v, i)$ is convex for all $v \in \mathcal{V}$ and $F(u, \cdot, i)$ is concave for all $u \in \mathcal{U}$.
(ii) The stochastic subgradients $G_u(u, v; i) \in \partial_u F(u, v; i)$ and $G_v(u, v; i) \in \partial_v F(u, v; i)$ with respect to u and v evaluated at any $(u, v) \in \mathcal{U}\times\mathcal{V}$ provide unbiased estimators for some respective subgradients of the expectation function: for any $(u, v) \in \mathcal{U}\times\mathcal{V}$, there exist some $g_u(u, v) \triangleq \mathbb{E}_{i\sim\mathcal{I}}[G_u(u, v; i)] \in \partial_u f(u, v)$ and $g_v(u, v) \triangleq \mathbb{E}_{i\sim\mathcal{I}}[G_v(u, v; i)] \in \partial_v f(u, v)$.

(iii) Let $\|\cdot\|_{\mathcal{U}}$ and $\|\cdot\|_{\mathcal{V}}$ be arbitrary norms that are well-defined on $\mathcal{U}$ and $\mathcal{V}$, while $\|\cdot\|_{\mathcal{U},*}$ and $\|\cdot\|_{\mathcal{V},*}$ be their respective dual norms. There exist constants $C_{u,*}, C_{v,*} > 0$ such that

$$\mathbb{E}_{i\sim\mathcal{I}}\left[\|G_u(u,v;i)\|_{\mathcal{U},*}^2\right] \le C_{u,*}^2 \quad \text{and} \quad \mathbb{E}_{i\sim\mathcal{I}}\left[\|G_v(u,v;i)\|_{\mathcal{V},*}^2\right] \le C_{v,*}^2 \quad \forall\ (u,v)\in\mathcal{U}\times\mathcal{V}.$$

For the online mirror descent, we further introduce two mirror maps that induce distances on $\mathcal{U}$ and $\mathcal{V}$, respectively.

Assumption 3. Let $\varphi_u : \mathcal{D}_u \to \mathbb{R}$ and $\varphi_v : \mathcal{D}_v \to \mathbb{R}$ satisfy the following:

(i) $\mathcal{U} \subseteq \mathcal{D}_u \cup \partial\mathcal{D}_u, \mathcal{U}\cap\mathcal{D}_u \neq \emptyset$ and $\mathcal{V} \subseteq \mathcal{D}_v \cup \partial\mathcal{D}_v, \mathcal{V}\cap\mathcal{D}_v \neq \emptyset$.
(ii) φ_u is ρ_u-strongly convex with respect to $\|\cdot\|_{\mathcal{U}}$; φ_v is ρ_v-strongly convex with respect to $\|\cdot\|_{\mathcal{V}}$.
(iii) $\lim_{u\to\partial\mathcal{D}_u}\|\nabla\varphi_u(u)\|_{\mathcal{U},*} = \lim_{v\to\partial\mathcal{D}_v}\|\nabla\varphi_v(v)\|_{\mathcal{V},*} = +\infty$.

Given the learning rates η_u, η_v, in each iteration $t = 1,\dots,T$, the online mirror descent samples $i_t \sim \mathcal{I}$ and updates

$$\begin{aligned} v_{t+1} &= \operatorname*{argmin}_{v\in\mathcal{V}} \ -\eta_v \cdot G_v(u_t, v_t; i_t)^\top v + D_{\varphi_v}(v, v_t), \\ u_{t+1} &= \operatorname*{argmin}_{u\in\mathcal{U}} \ \eta_u \cdot G_u(u_t, v_{t+1}; i_t)^\top u + D_{\varphi_u}(u, u_t), \end{aligned} \tag{12}$$

where $D_\varphi(w,w') = \varphi(w) - \varphi(w') - \nabla\varphi(w')^\top(w-w')$ denotes the Bregman divergence.

We measure the convergence of the saddle point problem in the duality gap:

$$\mathcal{E}(\overline{u}_T, \overline{v}_T) \triangleq \max_{v\in\mathcal{V}} f(\overline{u}_T, v) - \min_{u\in\mathcal{U}} f(u, \overline{v}_T)$$

such that, with

$$R_{\mathcal{U}} \triangleq \sqrt{\max_{u\in\mathcal{U}\cap\mathcal{D}_u}\varphi_u(u) - \min_{u\in\mathcal{U}\cap\mathcal{D}_u}\varphi_u(u)} \quad \text{and} \quad R_{\mathcal{V}} \triangleq \sqrt{\max_{v\in\mathcal{V}\cap\mathcal{D}_v}\varphi_v(v) - \min_{v\in\mathcal{V}\cap\mathcal{D}_v}\varphi_v(v)},$$

the online mirror descent converges as following.

Lemma 1 ((Nemirovski et al., 2009) (3.11))**.** *Under Assumption 2 and Assumption 3, when taking constant learning rates* $\eta_u = \eta_v = 2\Big/\sqrt{5T\left(\frac{2R_{\mathcal{U}}^2}{\rho_u}C_{u,*}^2 + \frac{2R_{\mathcal{V}}^2}{\rho_v}C_{v,*}^2\right)}$*, with* $\overline{u}_T = \frac{1}{T}\sum_{t=1}^T u_t$ *and* $\overline{v}_T = \frac{1}{T}\sum_{t=1}^T v_t$*,*

$$\mathbb{E}\left[\mathcal{E}(\overline{u}_T, \overline{v}_T)\right] \le 2\sqrt{\frac{10\left(\rho_v R_{\mathcal{U}}^2 C_{u,*}^2 + \rho_u R_{\mathcal{V}}^2 C_{v,*}^2\right)}{\rho_u\rho_v\cdot T}}.$$

Example 1 (Binary linear pixel-wise classifiers with convex and continuous objectives)**.** We consider a pixel-wise binary classification problem with $\mathcal{X} = [0,1]^d$, augmentations $A : \mathcal{X} \to \mathcal{X}$ for all $A \sim \mathcal{A}$, and a class of linear "UNets",

$$\mathcal{F} = \left\{ f_\theta : \mathcal{X} \to [0,1]^d \;\middle|\; f_\theta(\mathbf{x}) = \sigma\left(\boldsymbol{\theta}_d\boldsymbol{\theta}_e^\top\mathbf{x}\right) = \psi_\theta\left(\phi_\theta(\mathbf{x})\right),\ \phi_\theta(\mathbf{x}) = \frac{1}{\sqrt{d}}\boldsymbol{\theta}_e^\top\mathbf{x} \right\},$$

where the parameter space $\theta = (\boldsymbol{\theta}_e, \boldsymbol{\theta}_d) \in \mathcal{F}_\theta = \mathbb{S}^{d-1}\times\mathbb{S}^{d-1}$ is equipped with the ℓ_2 norm $\|\theta\|_{\mathcal{F}} = \left(\|\boldsymbol{\theta}_e\|_2^2 + \|\boldsymbol{\theta}_d\|_2^2\right)^{1/2}$; $\sigma : \mathbb{R}^d \to [0,1]^d$ denotes entry-wise application of the sigmoid function $\sigma(z) = (1+e^{-z})^{-1}$; and the latent space of encoder outputs $(\mathcal{Z}, \varrho)$ is simply the real line. Given the data distribution P_ξ, we recall that $\theta^* \triangleq \operatorname{argmin}_{\theta\in\mathcal{F}_\theta}\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim P_\xi}\left[\ell_{CE}(\theta;(\mathbf{x},\mathbf{y}))\right]$ and let $\mathcal{F}_{\theta^*}(\gamma) = \{\theta\in\mathcal{F}_\theta \mid \|\theta-\theta^*\|_{\mathcal{F}} \le \gamma\}$ for some $\gamma = O\left(1/\sqrt{d}\right)$. We assume that $\left|\mathbf{x}^\top\boldsymbol{\theta}_e^*\right| = O(1)$ for all $\mathbf{x}\in\mathcal{X}$. Then, $\ell_{CE}(\theta;(\mathbf{x},\mathbf{y}))$ and $\ell_{AC}(\theta;\mathbf{x},A_1,A_2)$ are convex and continuous in θ for all $(\mathbf{x},\mathbf{y})\in\mathcal{X}\times[K]^d$, $A_1, A_2 \sim \mathcal{A}^2$; while $C_{\theta,*} \le \max\left(2\sqrt{2}, 2\lambda_{AC}\right)$ and $C_{\boldsymbol{\beta},*} \le \max\left(O(1), 2\lambda_{AC}\right)$.

Rationale for Example 1. Let $\mathbf{y}_k = \mathbb{I}\{\mathbf{y} = k\}$ entry-wisely for $k = 0, 1$. We would like to show that, for any given $(\mathbf{x},\mathbf{y})\in\mathcal{X}\times[K]^d$, $A_1, A_2 \sim \mathcal{A}^2$,

$$\begin{aligned} \ell_{CE}(\theta) &= -\frac{1}{d}\left(\mathbf{y}_1^\top\log\sigma\left(\boldsymbol{\theta}_d\boldsymbol{\theta}_e^\top\mathbf{x}\right) + \mathbf{y}_0^\top\log\sigma\left(-\boldsymbol{\theta}_d\boldsymbol{\theta}_e^\top\mathbf{x}\right)\right), \\ \ell_{AC}(\theta) &= \frac{\lambda_{AC}}{\sqrt{d}}\cdot\left(A_1(\mathbf{x}) - A_2(\mathbf{x})\right)^\top\boldsymbol{\theta}_e \end{aligned}$$

are convex and continuous in $\theta = (\boldsymbol{\theta}_e, \boldsymbol{\theta}_d)$.

First, we observe that $\ell_{AC}(\theta)$ is linear (and therefore convex and continuous) in θ for all $\mathbf{x} \in \mathcal{X}$, $A_1, A_2 \sim \mathcal{A}^2$, with

$$\nabla_{\boldsymbol{\theta}_e} \ell_{AC}(\theta) = \frac{\lambda_{AC}}{\sqrt{d}} \cdot (A_1(\mathbf{x}) - A_2(\mathbf{x})), \quad \nabla_{\boldsymbol{\theta}_d} \ell_{AC}(\theta) = \mathbf{0}$$

such that $\|\nabla_\theta \ell_{AC}(\theta)\|_{\mathcal{F},*} \leq 2\lambda_{AC}$.

Meanwhile, with $\mathbf{z}(\theta) = \boldsymbol{\theta}_d \boldsymbol{\theta}_e^\top \mathbf{x}$, we have $\ell_{CE}(\theta) = -\frac{1}{d}\left(\mathbf{y}_1^\top \log \sigma(\mathbf{z}(\theta)) + \mathbf{y}_0^\top \log \sigma(-\mathbf{z}(\theta))\right)$ being convex and continuous in $\mathbf{z}(\theta)$:

$$\nabla_{\mathbf{z}}^2 \ell_{CE}(\theta) = \frac{1}{d} \operatorname{diag}(\sigma(\mathbf{z}(\theta))) \operatorname{diag}(1 - \sigma(\mathbf{z}(\theta))) \succcurlyeq 0.$$

Therefore, $\ell_{CE}(\theta)$ is convex and continuous in θ for all $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times [K]^d$:

$$\underbrace{\nabla_\theta^2 \ell_{CE}(\theta)}_{2d \times 2d} = \begin{bmatrix} \mathbf{x}\boldsymbol{\theta}_d^\top \\ \left(\boldsymbol{\theta}_e^\top \mathbf{x}\right) \mathbf{I}_d \end{bmatrix} \left(\frac{1}{d} \operatorname{diag}(\sigma(\mathbf{z}(\theta))) \operatorname{diag}(1 - \sigma(\mathbf{z}(\theta))) \right) \begin{bmatrix} \mathbf{x}\boldsymbol{\theta}_d^\top & \left(\boldsymbol{\theta}_e^\top \mathbf{x}\right) \mathbf{I}_d \end{bmatrix} \succcurlyeq 0,$$

where $\mathbf{I}_d$ denotes the $d \times d$ identity matrix. Further, from the derivation, we have

$$\nabla_{\boldsymbol{\theta}_e} \ell_{CE}(\theta) = \frac{1}{d} \boldsymbol{\theta}_d^\top \left(\sigma\left(\boldsymbol{\theta}_d \boldsymbol{\theta}_e^\top \mathbf{x}\right) - \mathbf{y}\right) \mathbf{x}, \quad \nabla_{\boldsymbol{\theta}_d} \ell_{CE}(\theta) = \frac{\boldsymbol{\theta}_e^\top \mathbf{x}}{d} \left(\sigma\left(\boldsymbol{\theta}_d \boldsymbol{\theta}_e^\top \mathbf{x}\right) - \mathbf{y}\right)$$

such that $\|\nabla_\theta \ell_{CE}(\theta)\|_{\mathcal{F},*} = \sqrt{\|\nabla_{\boldsymbol{\theta}_e} \ell_{CE}(\theta)\|_2^2 + \|\nabla_{\boldsymbol{\theta}_d} \ell_{CE}(\theta)\|_2^2} \leq 2\sqrt{2}$.

Finally, knowing $\|\nabla_\theta \ell_{CE}(\theta)\|_{\mathcal{F},*} \leq 2\sqrt{2}$ and $\|\nabla_\theta \ell_{AC}(\theta)\|_{\mathcal{F},*} \leq 2\lambda_{AC}$, we have

$$\left\|\nabla_\theta \widehat{L}_i^{WAC}(\theta, \boldsymbol{\beta})\right\|_{\mathcal{F},*} \leq \boldsymbol{\beta}_{[i]} \|\nabla_\theta \ell_{CE}(\theta)\|_{\mathcal{F},*} + (1 - \boldsymbol{\beta}_{[i]}) \|\nabla_\theta \ell_{AC}(\theta)\|_{\mathcal{F},*} \leq \max\left(2\sqrt{2}, 2\lambda_{AC}\right)$$

for all $i \in [n]$, and therefore,

$$C_{\theta,*} \leq \max\left(2\sqrt{2}, 2\lambda_{AC}\right).$$

Besides, with

$$\ell_{AC}(\theta) \leq \frac{\lambda_{AC}}{\sqrt{d}} \|A_1(\mathbf{x}) - A_2(\mathbf{x})\|_2 \|\boldsymbol{\theta}_e\|_2 \leq 2\lambda_{AC},$$

and since

$$\begin{aligned} \left(\boldsymbol{\theta}_d \boldsymbol{\theta}_e^\top \mathbf{x}\right)_{[j]} \leq& \left|\mathbf{x}^\top \boldsymbol{\theta}_e\right| \leq \left|\mathbf{x}^\top (\boldsymbol{\theta}_e - \boldsymbol{\theta}_e^*)\right| + \left|\mathbf{x}^\top \boldsymbol{\theta}_e^*\right| \leq \|\mathbf{x}\|_2 \|\boldsymbol{\theta}_e - \boldsymbol{\theta}_e^*\|_2 + O(1) \\ \leq& \gamma\sqrt{d} + O(1) = O(1) \end{aligned}$$

for all $j \in [d]$, $\ell_{CE}(\theta) \leq \log\left(1 + e^{O(1)}\right) = O(1)$, we have

$$C_{\boldsymbol{\beta},*} \leq \max(O(1), 2\lambda_{AC}).$$

□

C Dice Loss for Pixel-wise Class Imbalance

With finite samples in practice, since the averaged cross-entropy loss (Equation (2)) weights each pixel in the image label equally, the pixel-wise class imbalance can become a problem. For example, the background pixels can be dominant in most of the segmentation labels, making the classifier prone to predict pixels as background.

To cope with such vulnerability, (Chen et al., 2021; Cao et al., 2021; Wong et al., 2018; Taghanaki et al., 2019b; Yeung et al., 2022) propose to combine the cross-entropy loss with the *dice loss*—a

popular segmentation loss based on the overlap between true labels and their corresponding predictions in each class:

$$\ell_{DICE}\left(\theta;(\mathbf{x},\mathbf{y})\right)=1-\frac{1}{K}\sum_{k=1}^{K}DSC\left(f_\theta\left(\mathbf{x}\right)_{[:,k]},\mathbb{I}\left\{\mathbf{y}=k\right\}\right), \tag{13}$$

where for any $\mathbf{p}\in[0,1]^d$, $\mathbf{q}\in\{0,1\}^d$, $DSC\left(\mathbf{p},\mathbf{q}\right)=\frac{2\mathbf{p}^\top\mathbf{q}}{\|\mathbf{p}\|_1+\|\mathbf{q}\|_1}\in[0,1]$ denotes the dice coefficient (Milletari et al., 2016; Taghanaki et al., 2019a). Notice that by measuring the bounded dice coefficient for each of the K classes individually, the dice loss tends to be robust to class imbalance.

Taghanaki et al. (2019b) merges both dice and averaged cross-entropy losses via a convex combination. It is also a common practice to add a smoothing term in both the nominator and denominator of the DSC (Russell & Norvig, 2016).

Combining the dice loss (Equation (13)) with the weighted augmentation consistency regularization formulation (Equation (5)), in practice, we solve

$$\begin{aligned}&\widehat{\theta}^{WAC},\widehat{\boldsymbol{\beta}}\in\underset{\theta\in\mathcal{F}_{\theta^*}(\gamma)}{\operatorname{argmin}}\ \underset{\boldsymbol{\beta}\in[0,1]^n}{\operatorname{argmax}}\left\{\widehat{L}^{WAC}\left(\theta,\boldsymbol{\beta}\right)\triangleq\frac{1}{n}\sum_{i=1}^{n}\widehat{L}_i^{WAC}\left(\theta,\boldsymbol{\beta}\right)\right\} \\ &\widehat{L}_i^{WAC}\left(\theta,\boldsymbol{\beta}\right)\triangleq\ell_{DICE}\left(\theta;(\mathbf{x}_i,\mathbf{y}_i)\right)+\boldsymbol{\beta}_{[i]}\cdot\ell_{CE}\left(\theta;(\mathbf{x}_i,\mathbf{y}_i)\right)+(1-\boldsymbol{\beta}_{[i]})\cdot\ell_{AC}\left(\theta;\mathbf{x}_i,A_{i,1},A_{i,2}\right)\end{aligned} \tag{14}$$

with a slight modification in Algorithm 1 line 9:

$$\begin{aligned}\theta_t\leftarrow\theta_{t-1}-\eta_\theta\cdot\Big(&\nabla_\theta\ell_{DICE}\left(\theta_{t-1};(\mathbf{x}_{i_t},\mathbf{y}_{i_t})\right)+\left(\boldsymbol{\beta}_t\right)_{[i_t]}\cdot\nabla_\theta\ell_{CE}\left(\theta_{t-1};(\mathbf{x}_{i_t},\mathbf{y}_{i_t})\right)\\&+\Big(1-\left(\boldsymbol{\beta}_t\right)_{[i_t]}\Big)\cdot\nabla_\theta\ell_{AC}\left(\theta_{t-1};\mathbf{x}_{i_t},A_{i_t,1},A_{i_t,2}\right)\Big).\end{aligned}$$

On the influence of incorporating dice loss in experiments. We note that, in the experiments, the dice loss ℓ_{DICE} is treated independently of *AdaWAC* in Algorithm 1 via the standard stochastic gradient descent. In particular for the comparison with hard thresholding algorithms in Table 2, we keep the updating on ℓ_{DICE} of the original untrimmed batch intact for both **trim-train** and **trim-ratio** to exclude the potential effect of ℓ_{DICE} that is not involved in reweighting.

D Implementation Details and Datasets

We follow the official implementation of TransUNet[9] for model training. We use the same optimizer (SGD with learning rate 0.01, momentum 0.9, and weight decay 1e-4). For the Synapse dataset, we train TransUNet for 150 epochs on the training dataset and evaluate the last-iteration model on the test dataset. For the ACDC dataset, we train TransUNet for 360 epochs in total, while validating models on the ACDC validation dataset for every 10 epochs and testing on the best model selected by the validation. The total number of training iterations (*i.e.*, total number of batches) is set to be the same as that in the vanilla TransUNet (Chen et al., 2021) experiments. In particular, the results in Table 1 are averages (and standard deviations) over 3 arbitrary random seeds. The results in Table 2, Table 3, and Table 4 are given by the original random seed used in the TransUNet experiments.

Synapse multi-organ segmentation dataset (Synapse). The Synapse dataset[10] is multi-organ abdominal CT scans for medical image segmentation in the MICCAI 2015 Multi-Atlas Abdomen Labelling Challenge (Chen et al., 2021). There are 30 volumetric CT scans with variable volume sizes ($512\times512\times85-512\times512\times198$), and slice thickness ranges from 2.5mm to 5.0mm. We use the pre-processed data provided by Chen et al. (2021) and follow their train/test split to use 18 volumes for training and 12 volumes for testing on 8 abdominal organs—aorta, gallbladder, left kidney (L), right kidney (R), liver, pancreas, spleen, and stomach. The abdominal organs were labeled by experience undergraduates and verified by a radiologist using MIPAV software according to the information from Synapse wiki page.

[9] https://github.com/Beckschen/TransUNet

[10] See detailed description at https://www.synapse.org/#!Synapse:syn3193805/wiki/217789

Automated cardiac diagnosis challenge dataset (ACDC). The ACDC dataset[11] is cine-MRI scans in the MICCAI 2017 Automated Cardiac Diagnosis Challenge. There are 200 scans from 100 patients, and each patient has two volumetric frames with slice thickness from 5mm to 8mm. We use the pre-processed data also provided by Chen et al. (2021) and follow their train/validate/test split to use 70 patients' scans for training, 10 patients' scans for validation, and 20 patients' scans for testing on three cardiac structures—left ventricle (LV), myocardium (MYO), and right ventricle (RV). The data were labeled by one clinical expert according to the description on ACDC dataset website.

E Additional Experimental Results

E.1 Sample Efficiency and Robustness of *AdaWAC* with UNet

In addition to the empirical evidence on TransUNet presented in Table 1, here, we demonstrate that the sample efficiency and distributional robustness of *AdaWAC* extend to the more widely used UNet architecture. In Table 5, analogous to Table 1, the experiments on the **full** and **half-slice** datasets provide evidence for the *sample efficiency* of *AdaWAC* compared to the baseline (ERM+SGD) on UNet. Meanwhile, the *distributional robustness* of *AdaWAC* with UNet is well illustrated by the **half-vol** and **half-sparse** experiments.

Table 5: *AdaWAC* with UNet trained on the full Synapse and its subsets

Training	Method	DSC ↑	HD95 ↓	Aorta	Gallbladder	Kidney (L)	Kidney (R)	Liver	Pancreas	Spleen	Stomach
full	baseline	74.04 ± 1.52	36.65 ± 0.33	84.93	55.59	77.59	70.92	92.21	55.01	82.87	73.21
	AdaWAC	$\mathbf{76.71 \pm 0.62}$	$\mathbf{30.67 \pm 2.85}$	85.68	55.19	80.15	75.45	94.11	56.19	87.54	81.39
half-slice	baseline	73.09 ± 0.10	40.05 ± 4.99	83.23	53.18	74.69	71.51	92.74	52.81	83.85	72.71
	AdaWAC	$\mathbf{75.12 \pm 0.78}$	$\mathbf{29.26 \pm 2.16}$	85.15	55.77	79.29	72.47	93.71	54.93	86.09	73.53
half-vol	baseline	63.21 ± 2.53	64.20 ± 4.46	79.46	45.79	55.79	54.91	88.65	41.61	71.68	67.77
	AdaWAC	$\mathbf{71.09 \pm 1.14}$	$\mathbf{39.95 \pm 7.76}$	83.15	49.14	75.74	70.33	90.47	44.81	82.34	72.75
half-sparse	baseline	37.30 ± 1.32	69.67 ± 2.89	61.57	8.33	57.45	50.44	60.28	23.51	17.83	18.99
	AdaWAC	$\mathbf{44.85 \pm 1.03}$	$\mathbf{62.40 \pm 5.17}$	71.56	8.40	65.42	62.73	74.02	24.16	36.65	15.88

Implementation details of UNet experiments. For the backbone architecture of experiments in Table 5, we use a UNet with a ResNet-34 encoder initialized with ImageNet pre-trained weights. We leverage the implementation of UNet and load the pre-trained model via the PyTorch API for segmentation models (Iakubovskii, 2019). For training, we use the same optimizer (SGD with learning rate 0.01, momentum 0.9, and weight decay 1e-4) and the total number of epochs (150 epochs on Synapse training set) as the TransUNet experiments, evaluating the last-iteration model on the test dataset. As before, the results in Table 5 are averages (and standard deviations) over 3 arbitrary random seeds.

E.2 Visualization of Segmentation on ACDC dataset

As shown in Figure 4, the model trained by *AdaWAC* segments cardiac structures with more accurate shapes (column 1), identifies organs missed by baseline TransUNet (column 2-3) and circumvents the false-positive pixel classifications (*i.e.*, fake predictions of background pixels as organs) suffered by the TransUNet baseline (column 4-6).

E.3 Visualization of Segmentation on Synapse with Distributional Shift

Figure 5 visualizes the segmentation predictions on 6 Synapse test slices made by models trained via *AdaWAC* (ours) and via the baseline (ERM+SGD) with TransUNet (Chen et al., 2021) on the **half-sparse** subset of the Synapse training set. We observe that, although the segmentation performances of both the baseline and *AdaWAC* are compromised by the extreme scarcity of label-dense samples and the severe distributional shift, *AdaWAC* provides more accurate predictions on the relative positions of organs, as well as less misclassification of organs (*e.g.*, the baseline tends to misclassify other organs and the background as the left kidney). Nevertheless, due to the scarcity of labels, both the model trained with *AdaWAC* and that trained with the baseline fail to make good predictions on the segmentation boundaries.

[11] See detailed description at https://www.creatis.insa-lyon.fr/Challenge/acdc/

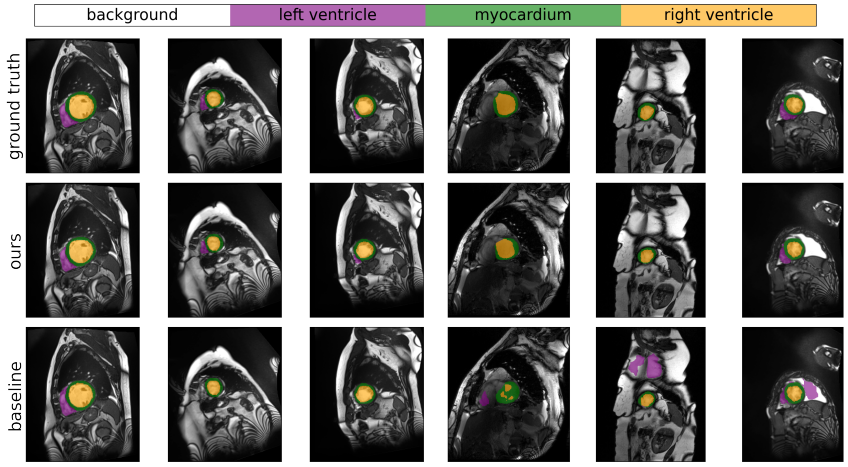

Figure 4: Visualization of segmentation results on ACDC dataset. From top to bottom: ground truth, ours, and baseline method.

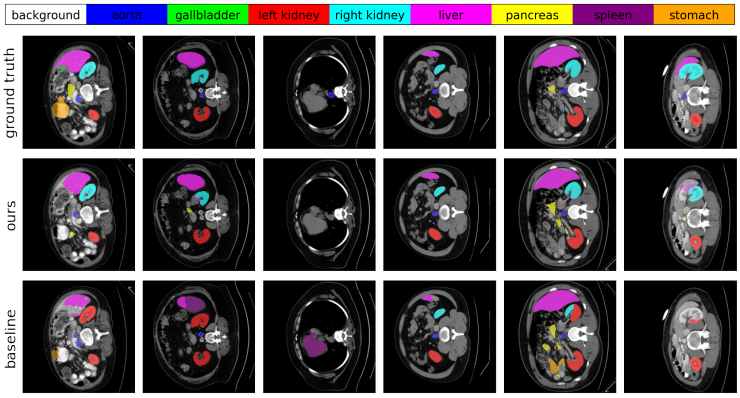

Figure 5: Visualization of segmentation predictions made by models trained via *AdaWAC* (ours) and via the baseline (ERM+SGD) with TransUNet (Chen et al., 2021) on the **half-sparse** subset of the Synapse training set. Top to bottom: ground truth, ours (*AdaWAC*), baseline.

E.4 Experimental Results on Previous Metrics

In this section, we include the results of experiments on Synapse[12] dataset with metrics defined in TransUNet (Chen et al., 2021) for reference. In TransUNet (Chen et al., 2021), DSC is 1 when the sum of ground truth labels is zero (i.e., gt.sum() == 0) while the sum of predicted labels is nonzero (i.e., pred.sum() > 0). However, according to the definition of dice scores, $DSC = 2|A \cap B|/(|A| + |B|), \forall A, B$, the DSC for the above case should be 0 since the intersection is 0 and the denominator is non-zero. In our evaluation, we change the special condition for DSC as 1 to pred.sum == 0 and gt.sum() == 0 instead, in which case the denominator is 0.

[12]Note that the numbers of correct metrics and metrics in TransUNet (Chen et al., 2021) on ACDC dataset are the same.

Table 6: *AdaWAC* with TransUNet trained on the full Synapse and its subsets, measured by metrics in TransUNet (Chen et al., 2021).

Training	Method	DSC ↑	HD95 ↓	Aorta	Gallbladder	Kidney (L)	Kidney (R)	Liver	Pancreas	Spleen	Stomach
full	baseline	77.32	29.23	87.46	63.54	82.06	77.76	94.10	54.06	85.07	74.54
	AdaWAC	80.16	25.79	87.23	63.27	84.58	81.69	94.62	58.29	90.63	81.01
half-slice	baseline	76.24	24.66	86.26	57.61	79.32	76.55	94.34	54.04	86.20	75.57
	AdaWAC	78.14	29.75	86.66	62.28	81.36	78.84	94.60	57.95	85.38	78.01
half-vol	baseline	72.65	35.86	83.29	43.70	78.25	77.25	92.92	51.32	83.80	70.66
	AdaWAC	75.93	34.95	84.45	60.40	79.59	76.06	93.19	54.46	84.91	74.37
half-sparse	baseline	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	AdaWAC	39.68	80.93	76.59	0.00	66.53	62.11	49.69	31.09	12.30	19.11

Table 7: AdaWAC versus hard thresholding algorithms with TransUNet on Synapse, measured by metrics in TransUNet (Chen et al., 2021).

Method	DSC ↑	HD95↓	Aorta	Gallbladder	Kidney (L)	Kidney (R)	Liver	Pancreas	Spleen	Stomach
baseline	77.32	29.23	87.46	63.54	82.06	77.76	94.10	54.06	85.07	74.54
trim-train	77.05	26.94	86.70	60.65	80.02	76.64	94.25	54.20	86.44	77.49
trim-ratio	75.30	28.59	87.35	57.29	78.70	72.22	94.18	52.32	86.31	74.03
trim-train+ACR	76.70	35.06	87.11	62.22	74.19	75.25	92.19	57.16	88.21	77.30
trim-ratio+ACR	79.02	33.59	86.82	61.67	83.52	81.22	94.07	59.06	88.08	77.71
AdaWAC (ours)	80.16	25.79	87.23	63.27	84.58	81.69	94.62	58.29	90.63	81.01

Table 8: Ablation study of AdaWAC with TransUNet trained on Synapse, measured by metrics in TransUNet (Chen et al., 2021).

Method	DSC ↑	HD95↓	Aorta	Gallbladder	Kidney (L)	Kidney (R)	Liver	Pancreas	Spleen	Stomach
baseline	77.32	29.23	87.46	63.54	82.06	77.76	94.10	54.06	85.07	74.54
reweight-only	77.72	29.24	86.15	62.31	82.96	80.28	93.42	55.86	85.29	75.49
ACR-only	78.93	31.65	87.96	62.67	81.79	80.21	94.52	60.41	88.07	75.83
AdaWAC-0.01	78.98	27.81	87.58	61.09	82.29	80.22	94.90	55.92	91.63	78.23
AdaWAC-1.0	80.16	25.79	87.23	63.27	84.58	81.69	94.62	58.29	90.63	81.01