Appendix

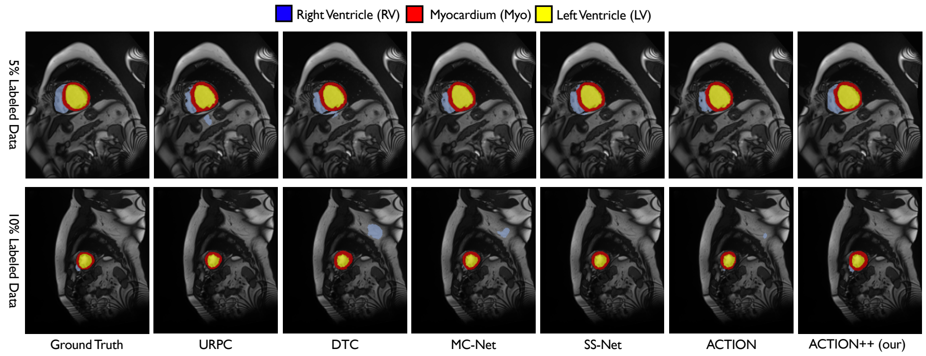

Fig. 3. Visualization results on ACDC with 5% and 10% labeled data. ACTION++ consistently outputs more accurate predictions, especially for small regions.

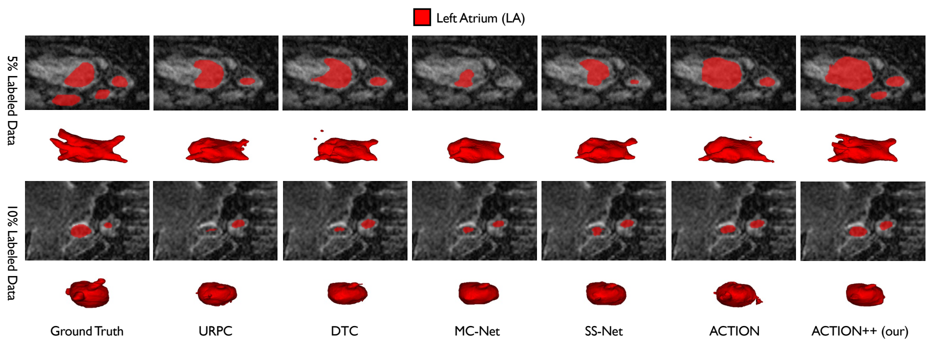

Fig. 4. Visualization results on LA with 5% and 10% labeled data. ACTION++ consistently achieves more sharper and accurate object boundaries.

Table 5. Ablation studies of different components (*i.e.*, ATS and SAACL).

Method	DSC[%]↑	ASD[voxel]↓
pre-training w/o ATS	86.2	2.69
pre-training w/ ATS	88.1	2.44
fine-tuning w/o SAACL/ATS	89.0	2.06
fine-tuning only w/ ATS	89.3	1.98
fine-tuning only w/ SAACL	89.5	1.96
fine-tuning w/ SAACL/ATS	**89.9**	**1.74**

Table 6. Effect of cosine period, different methods of varying τ, and λ_a.

T/#iterations	DSC	Method	DSC	λ_a	DSC
no/fixed τ	89.5	fixed	89.5	0.05	88.5
0.1	89.8	step	89.4	0.1	89.3
0.2	89.1	rand	88.9	0.2	**89.9**
0.5	89.2	oscil.	89.2	0.5	89.7
1.0	**89.9**	cos	**89.9**	1.0	89.1
2.0	89.7	-	-	10	87.9

A Theoretical Analysis

In this section, we discuss the performance guarantee of the proposed SAACL. For abstraction, we denote an image and its corresponding segmentation map as $\mathbf{x} = \{\omega_p\}_p$, $\mathbf{y} = \{y_p\}_p$, where ω_p is a pixel. We also denote the feature generator as f, such that $f(\omega_p; \mathbf{x}) = \boldsymbol{\phi}_p$ for any pixel p. Recent work [8] has shown that, to evaluate the performance of the representations learned via contrastive learning (CL), it suffices to consider a simplified nearest neighbour (NN) classifier[3] $g_f(\omega_p;\mathbf{x}) = \arg\min_{c\in[K]} \|f(\omega_p;\mathbf{x}) - \psi_c^\star\|_2$, where $\psi_c^\star$ denotes the center of class c in the latent representation space. To this end, we focus on the error rate of g_f defined as $\mathcal{E}(g_f) = \sum_{c=1}^{K} \mathbb{P}[g_f(\omega_p;\mathbf{x}) \neq c, \forall\omega_p \in Cla_c]$, where $\omega_p \in Cla_c$ refers to pixels in class c. Note that each class c, regardless of being head or tail class, has equal weight in the definition of $\mathcal{E}(g_f)$, indicating that a small $\mathcal{E}(g_f)$ implies good long-tail segmentation performance.

We now demonstrate that SAACL helps achieve a small error $\mathcal{E}(g_f)$. The success of contrastive learning mainly depends on two aspects: positive alignment and class divergence [8]. Specifically, the positive alignment is defined as follows:

$$A = \sqrt{\mathbb{E}_{\mathbf{x},\tilde{\mathbf{x}}}\mathbb{E}_{c\in[K]}\mathbb{E}_{\omega_p\in Cla_c}[\|f(\omega_p;\mathbf{x}) - f(\omega_p;\tilde{\mathbf{x}})\|^2]}, \tag{7}$$

where $\mathbf{x}$ and $\tilde{\mathbf{x}}$ are two augmentations from the same input sample (*i.e.*, positive sample pairs). The class divergence is defined as $D = \max_{c\neq c'} \overline{\boldsymbol{\phi}}_c \cdot \overline{\boldsymbol{\phi}}_{c'}$, which computes the distances between class centers. The following theorem discloses the link between the error rate and the alignment A and class divergence D.

Theorem 1 ([8]). *There exist some constant $\rho(\sigma,\delta,\epsilon)$ and Δ whose value depends on the data augmentation method and Lipschitzness of the model f. Let $\zeta(\sigma,\delta,\epsilon) = r^2[1-\rho(\sigma,\delta,\epsilon) - \sqrt{2\rho(\sigma,\delta,\epsilon)} - \Delta/2]$. If for any class $c, c' \in [K]$, it holds that $\overline{\boldsymbol{\phi}}_c \cdot \overline{\boldsymbol{\phi}}_{c'} \leq \zeta(\sigma,\delta,\epsilon)$, then $\mathcal{E}(g_f) \leq 1-\sigma+\mathcal{O}(1/\epsilon)A$.*

Due to space limit, please refer to Theorem 1 in [8] for the detailed mathematical form of $\rho(\sigma,\delta,\epsilon)$, Δ and the problem-related parameters σ, δ and ϵ. For our purpose, we observe that: (1) good positive alignment (small A) directly indicates low error according to the error upper bound; (2) a large class divergence (small D) can help satisfy the condition on $\overline{\boldsymbol{\phi}}_c \cdot \overline{\boldsymbol{\phi}}_{c'}$. Therefore, both A are D are crucial to improving the representation learning.

From (5), both the alignment and the diversity are captured by the objective $\mathcal{L}_{\text{aaco}}$. We rewrite (5) as $\mathcal{L}_{\text{aaco}} = \sum_{i=1}^{n}(\mathcal{L}_{i,1} + \lambda_a\mathcal{L}_{i,2})/n$, where $\mathcal{L}_{i,1}$ equals:

$$-\sum_{\boldsymbol{\phi}_i^+}\log\frac{\exp(\boldsymbol{\phi}_i\cdot\boldsymbol{\phi}_i^+/\tau_{sa})}{\sum_{\boldsymbol{\phi}_j}\exp(\boldsymbol{\phi}_i\cdot\boldsymbol{\phi}_j/\tau_{sa})} = -\sum_{\boldsymbol{\phi}_i^+}\boldsymbol{\phi}_i\cdot\boldsymbol{\phi}_i^+/\tau_{sa} - \sum_{\boldsymbol{\phi}_i^+}\log\sum_{\boldsymbol{\phi}_j}\exp(\boldsymbol{\phi}_i\cdot\boldsymbol{\phi}_j/\tau_{sa}).$$

Here the first term in the above can be rewrite as $\sum_{\boldsymbol{\phi}_i^+}\|\boldsymbol{\phi}_i - \boldsymbol{\phi}_i^+\|^2/(2\tau_{sa}) - 1$ given the normalization $\|\boldsymbol{\phi}_p\| = 1$ for all pixels p. Then by the definition $f(\omega_p;\mathbf{x}) = \boldsymbol{\phi}_p$ and (7), it is clear that $\mathcal{L}_{i,1}$ induces small A (i.e. good alignment).

Similar analysis shows $\mathcal{L}_{i,2}$ encourages $\boldsymbol{\phi}_i$ to be close to the pre-computed optimal class center $\boldsymbol{\nu}_i$ (small $\|\boldsymbol{\phi}_i - \boldsymbol{\nu}_i\|$). The class centers computed from solving (3) induces large distance $\|\boldsymbol{\nu}_i - \boldsymbol{\nu}_j\|$ between centers. Furthermore, since (3) does not involve any data yet, it is immune to long-tailness and can guarantee well-separeted centers for the representation of tail classes. Together it holds that $\mathcal{L}_{i,2}$ encourages large $\|\overline{\boldsymbol{\phi}}_c - \overline{\boldsymbol{\phi}}_{c'}\|$ for $c \neq c'$, or equivalently small $\overline{\boldsymbol{\phi}}_c \cdot \overline{\boldsymbol{\phi}}_{c'}$, which is exactly the class divergence.

[3] This is because an NN classifier is a special case of a linear classifier, which can be approximated by a neural network. See Sec. 2 of [8].