Appendix to
Rethinking Semi-Supervised Medical Image Segmentation: A Variance-Reduction Perspective

Sec. A provides additional theoretical analysis.

Sec. B provides additional details on the training datasets.

Sec. C provides additional details on the implementation.

Sec. D provides details on 2D/3D methods in comparision.

Sec. E provides additional framework details.

Sec. F provides model architecture details.

Sec. G provides more experimental results on the 2D LiTS [77] dataset.

Sec. H provides more experimental results on the 2D MMWHS [78] dataset.

Sec. I provides more experimental results on the 3D LA [79] dataset.

Sec. J provides more experimental results on the 3D MP-MRI dataset.

Sec. K provides more experimental results on three semantic segmentation benchmarks (Cityscapes [92], Pascal VOC 2012 [93], SUN RGB-D [94]).

Sec. L forms a better understanding of generalization across label ratios and frameworks.

Sec. M provides final checkpoint loss landscapes.

Sec. N provides additional details on the hyperparameter selection.

A Theoretical Results Details

A.1 Proof of Lemma 3.1

Proof. By definition of the estimate $\widehat{H}$, since the sampling from each group is independent, the variance can be written as

$$\mathrm{Var}[\widehat{H}_{\mathrm{SG}}] = \frac{1}{n^2}\sum_{m=1}^{M}\sum_{p\in\mathcal{D}_m}\sigma_m^2.$$

Since $|\mathcal{D}_m| = n_m$, it follows that

$$\mathrm{Var}[\widehat{H}_{\mathrm{SG}}] = \sum_{m=1}^{M}\frac{n_m^2}{n^2}\frac{\sigma_m^2}{n_m}.$$

Note that σ_m^2 is the variance of the $h(\mathbf{x};p)$ where $p \sim \mathcal{P}_m$ uniformly.

Analysis of SAG. We now consider SAG. Recall that, compared to SG, SAG first samples $n_m/2$ pixels from $\mathcal{P}_m$[6] from each $m = 1, \cdots, M$ in the same way as SG. SAG then deterministically picks the rest $n_m/2$ pixels to be the reflection pixels of the first half (see Figure 2(3)). Let's choose arbitrary group m and denote by p and p' the sampled pixel and its reflection from that group. Note that the variance of p satisfies $\mathrm{Var}_{\mathrm{SAG}}[h(\mathbf{x};p)] = \sigma_m^2$ since p is sampled in a same way as in the case of SG. Observe that, by symmetry, the variance of p and p' should be the same $\mathrm{Var}_{\mathrm{SAG}}[h(\mathbf{x};p)] = \mathrm{Var}_{\mathrm{SAG}}[h(\mathbf{x};p)] = \sigma_m^2$. It follows that:

$$\begin{aligned}&\mathrm{Var}_{\mathrm{SAG}}[h(\mathbf{x};p) + h(\mathbf{x};p')]\\&= \mathrm{Var}_{\mathrm{SAG}}[h(\mathbf{x};p)] + \mathrm{Var}_{\mathrm{SAG}}[h(\mathbf{x};p')] + 2\mathrm{Cov}_{\mathrm{SAG}}[h(\mathbf{x};p), h(\mathbf{x};p')]\\&\leq 4\sigma_m^2,\end{aligned}$$

[6] For simplicity, we ignore the rounding issue.

where the second step holds because the correlation between $h(\mathbf{x}; p)$ and $h(\mathbf{x}; p')$ is at most 1. It follows that:

$$\begin{aligned}
\mathrm{Var}[\widehat{H}_{\text{SAG}}] &= \sum_{m=1}^{M} \frac{|\mathcal{P}_m|^2}{|\mathcal{P}|^2} \frac{1}{|\mathcal{D}_m|^2} \mathrm{Var}\left[\sum_{p \in \mathcal{D}_m} h(\mathbf{x}; p)\right] \\
&= \sum_{m=1}^{M} \frac{n_m^2}{n^2} \frac{1}{n_m^2} \sum_{p,p'} \mathrm{Var}[h(\mathbf{x}; p) + h(\mathbf{x}; p')] \\
&\leq \sum_{m=1}^{M} \frac{n_m^2}{n^2} \frac{1}{n_m^2} \frac{n_m}{2} 4\sigma_m^2 \\
&= \sum_{m=1}^{M} \frac{2n_m}{n^2} \sigma_m^2 \\
&= 2\,\mathrm{Var}[\widehat{H}_{\text{SG}}].
\end{aligned}$$

□

A.2 Proof of Theorem 3.2

Proof of Theorem 3.2. Denote $w_m = n_m/n$, where $n_m = |\mathcal{D}_m|$. The unbiasedness is straightforward: by proportional sampling, the probability that an arbitrary pixel in group $\mathcal{P}_m$ being chosen is equal to $|\mathcal{D}_m|/|\mathcal{P}_m|$. Since $|\mathcal{D}_m| = |\mathcal{P}_m|$, the probability of being chosen is equal across all pixels, and hence an arbitrary pixel $p \in \mathcal{D} = \cup_m \mathcal{D}_m$ is equally likely to be any pixel in the population $\mathcal{P}$. As a result, we have

$$\begin{aligned}
\mathbb{E}[\widehat{H}_{\text{SG}}(\mathbf{x}; \mathcal{D})] = \mathbb{E}\left[\sum_{p \in \mathcal{D}} h(\mathbf{x}; p)/|\mathcal{D}|\right] &= \mathbb{E}_p[h(\mathbf{x}; p)] \\
&= H(\mathbf{x}),
\end{aligned}$$

where the first equality is by the construction of $\widehat{H}_{\text{SG}}(\mathbf{x}; \mathcal{D})$, the second equality is by symmetry of $p \in \mathcal{D}$, and the last inequality is by the above conclusion that p is equally possible to be any pixel in $\mathcal{P}$. This finishes the proof that $\widehat{H}_{\text{SG}}$ is unbiased.

The variance of $\widehat{H}_{\text{SG}}$ is equal to

$$\mathrm{Var}[\widehat{H}_{\text{SG}}] = \sum_{m=1}^{M} \frac{n_m^2}{n^2} \frac{\sigma_m^2}{n_m},$$

where σ_m^2 is the variance of the $f(\mathbf{x}; p)$ when $p \sim \mathcal{D}_m$ uniformly. Under the case of proportional sampling, i.e., $n_m \propto |\mathcal{D}_m|$ for all $m = 1, \cdots, M$ (for simplicity we assume such n_m's are all integers), the variance becomes

$$\mathrm{Var}[\widehat{H}_{\text{SG}}] = \sum_{m=1}^{M} \frac{(w_m \cdot n)^2}{n^2} \frac{\sigma_m^2}{w_m \cdot n} = \sum_{m=1}^{M} w_m \frac{\sigma_m^2}{n}.$$

We now consider NS (i.e. naïve sampling). By definition, the random sampling samples pixels uniformly from the set $\mathcal{P}$ of all pixels. Since n is the total number of pixels sampled, the variance of the $\widehat{H}_{\text{NS}}$ is given as $\mathrm{Var}[\widehat{H}_{NS}] = \sigma^2/n$ where σ^2 is the sampling variance of $h(\mathbf{x}; p)$ given $p \sim \mathcal{P}$ uniformly. To determine σ^2, we apply a variance decomposition trick via conditioning. Specifically, the sampling from NS is a two-step process: (1) sample the group index m from $[M]$, (2) sample the pixel uniformly from $\mathcal{P}_m$. Applying the law of total variance, we have

$$\begin{aligned}
&\mathrm{Var}_{p \overset{\text{unif.}}{\sim} \mathcal{P}}[h(\mathbf{x}; p)] \\
&= \mathbb{E}[\mathrm{Var}[h(\mathbf{x}; p)] | p \in \mathcal{P}_m] + \mathrm{Var}[\mathbb{E}[h(\mathbf{x}; p) | p \in \mathcal{P}_m]] \\
&= \sum_{m=1}^{M} w_m \sigma_m^2 + \sum_{m=1}^{M} \left(\mathbb{E}_{p \overset{\text{unif.}}{\sim} \mathcal{P}_m}[h(\mathbf{x}; p)] - \mathbb{E}_{p \overset{\text{unif.}}{\sim} \mathcal{P}}[h(\mathbf{x}; p)]\right)^2.
\end{aligned}$$

As a result, we conclude that, for any image $\mathbf{x}$, the sampling variance of SG estimate and NS estimate satisfies

$$\begin{aligned}\operatorname{Var}[\widehat{H}_{\text{NS}}] &= \operatorname{Var}[\widehat{H}_{\text{SG}}] \\ &\quad + \frac{1}{n}\sum_{m=1}^{M}\left(\mathbb{E}_{p \overset{\text{unif.}}{\sim} \mathcal{P}_m}[h(\mathbf{x};p)] - \mathbb{E}_{p \overset{\text{unif.}}{\sim} \mathcal{P}}[h(\mathbf{x};p)]\right)^2,\end{aligned}$$

which finishes the proof. □

A.3 Further Details on Convergence

We first make the following assumptions about the loss function and the gradient estimate in our learning problem.

Assumption A.1 (Smoothness)**.** *The objective function $\mathcal{L}(\cdot)$ is L-smooth, i.e., $\mathcal{L}$ is differentiable and $\|\nabla\mathcal{L}(\theta) - \nabla\mathcal{L}(\theta')\|_2 \leq L(\|\theta - \theta'\|_2)$ for all θ, θ' in the domain of $\nabla\mathcal{L}$.*

Assumption A.2 (Bounded variance)**.** *There exists some $\sigma_g^2 > 0$ such that for all θ in the domain,*

$$\mathbb{E}\left[\|g(\theta) - \nabla\mathcal{L}(\theta)\|^2\right] \leq \sigma_g^2.$$

These assumptions are standard in optimization theory and in various settings [82, 83, 84]. The first assumption allows us to characterize the loss landscape of our learning problem, and the second assumption ensures the gradient estimate does not deviate largely from the truth, which is essential for the training to converge. With these two assumptions, it is guaranteed that

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla\mathcal{L}(\theta)\|_2^2\right] \leq C\left(\frac{1}{T} + \frac{\sigma_g}{\sqrt{T}}\right).$$

Proposition 1. *Suppose the step sizes in SGD are taken to be $\min\left\{1/L, \alpha/(\sigma_g\sqrt{T})\right\}$ where T is the number of steps. Then under Assumption A.1 and A.2, there exists some constant C depending on $\mathcal{L}$, α, and L, such that*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla\mathcal{L}(\theta_t)\|_2^2\right] \leq C\left(\frac{1}{T} + \frac{\sigma_g}{\sqrt{T}}\right).$$

Proposition 1 shows that, in expectation, the convergence rate of the average gradient norm consists of a fast rate $1/T$ and a slow rate $\sigma_g/\sqrt{T}$, with the slow rate being the dominant term. Importantly, the slow rate depends on the variance of the gradient estimate. This suggests that a variance reduced gradient estimate allows SGD to reach an approximate local minimum with less iterations, leading to faster convergence of the training loss. This intuition is corroborated by Fig. 5, where we directly visualize the training trajectory of contrastive loss versus epoch for three sampling methods. We observe that SG features a faster loss decay and a narrower error bar, showing that it outperforms other methods in both the convergence speed and the stability. Furthermore, Proposition 1 indicates that our proposed sampling techniques are universal, since it applies to all scenarios as long as the mild assumptions A.1 and A.2 are satisfied. In other words, when it comes to scenarios that involve pixel/voxel-level sampling on 2D/3D images, utilizing SG/SAG instead of naive sampling can lead to enhanced stability and decreased variance.

B Datasets

Our experiments are conducted on five 2D/3D representative datasets in semi-supervised medical image segmentation literature, including 2D benchmarks (*i.e.*, ACDC [76], LiTS [77], and MMWHS [78]) and 3D benchmarks (*i.e.*, LA [79] and in-house MP-MRI).

- **The ACDC dataset** was collected from MICCAI 2017 ACDC challenge [76], consisting of 200 3D cardiac cine MRI scans with 3 classes – left ventricle (LV), myocardium (Myo), and right ventricle (RV). Following [49, 6, 7], we utilize 120, 40, and 40 scans for training, validation, and testing [7]. Note that 1%, 5%, and 10% label ratios are the ratio of patients. The splitting details are in the supplementary material. For pre-processing, we normalize the intensity of each 3D scan into $[0, 1]$ by using min-max normalization, and re-sample images and segmentation maps to 256×256 pixels.

- **The LiTS dataset** was collected from MICCAI 2017 Liver Tumor Segmentation Challenge [77], consisting of 131 contrast-enhanced 3D abdominal CT volumes with 2 classes – liver and tumor. Note that 1%, 5%, and 10% label ratios are the ratio of patients. We utilize 100 and 31 scans for training and testing, with random order. The splitting details are in the supplementary material. For pre-processing, we follow the setting in [3] by truncating the intensity of each 3D scan into $[-200, 250]$ HU for removing irrelevant and redundant features, normalizing each 3D scan into $[0, 1]$, and re-sample images and segmentation maps to 256×256 pixels.
- **The MMWHS dataset** was collected from MICCAI 2017 challenge [78], consisting of 20 3D cardiac MRI scans with 7 classes – left ventricle (LV), left atrium (LA), right ventricle (RV), right atrium (RA), myocardium (Myo), ascending aorta (AAo), and pulmonary artery (PA). Note that 1%, 5%, and 10% label ratios are based on the ratio of patients. Following [14], we utilize 15 and 5 scans for training and testing. The splitting details are in the supplementary material. For pre-processing, we normalize the intensity of each 3D scan into $[0, 1]$ by using min-max normalization, and re-sample images and segmentation maps to 256×256 pixels.
- **The LA dataset** [79] was a representative 3D benchmark, consisting of 100 gadolinium-enhanced MRI scans with one class – left atrium (LA), with an isotropic resolution of $0.625 \times 0.625 \times 0.625\text{mm}^3$. Note that 1%, 5%, and 10% label ratios are the ratio of patients. The fixed split (*i.e.*, 5%, and 10%) [5, 49] uses 80 and 20 scans for training and testing [8], and 1% label ratio is randomly split. The splitting details are in the supplementary material. For pre-processing, we crop all the scans at the heart region, normalize the intensities of each 3D scan into $[0, 1]$, and randomly crop all the training sub-volumes into $112 \times 112 \times 80\text{mm}^3$.
- **The Multi-phasic MRI (MP-MRI) dataset** was an in-house 3D dataset, consisting of 160 multi-phasic MRI scans with one class – liver, each of which includes T1 weighted DCE-MRI images at three-time points (*i.e.*, pre-contrast, arterial phase, and venous phases). Three images are mutually registered to the arterial phase images, with an isotropic voxel size of $1.00 \times 1.00 \times 1.00\text{mm}^3$. The dataset is randomly divided into 100 scans for training, 40 for validation, and 20 for testing. Note that 1%, 5%, and 10% label ratios are the ratio of patients. The splitting details are in the supplementary material. For pre-processing, we normalize the intensity of each 3D scan into $[0, 1]$ by using min-max normalization, and re-sample images and segmentation maps to 256×256 pixels.

C Implementation Details

In our experiments, all of our evaluated methods have been trained using similar settings for simplicity in reproducing our results. All experiments are conducted with PyTorch [96] on an NVIDIA RTX 3090 Ti. We adopt an SGD optimizer with momentum 0.9 and weight decay 10^{-4}. The initial learning rate is set to 0.01. We use the 2D `UNet` [27] or 3D `VNet` [73] backbones as the segmentation network under different labeled ratio settings (*i.e.*, 1%, 5%, 10% labeled ratios). Following [14], we adopt the feature pyramid network (`FPN`) [97] architecture as the representation head ψ_r with a separate 512-dimension output layer. The momentum hyperparameter is set to 0.99.

For pre-training, the networks are trained for 100 epochs with a batch size of 6. We apply the *mined* views with $d = 5$, following [14]. As for fine-tuning, the networks are trained for 200 epochs with a batch size of 8. The learning rate decays by a factor of 10 every 2500 iterations during the training. We apply the temperature with $\tau_t = 0.01$, $\tau_s = 0.1$, and $\tau = 0.5$, respectively. The size of the memory bank is set to 36. For the CL training, we use the implementation from [14] and leave all parameters on their default settings, *e.g.*, we apply the hyperparameters with $\lambda_1 = 0.01$, $\lambda_2 = 1.0$, and $\lambda_3 = 1.0$. For other hyper-parameters in all the evaluated methods, we adopt the suggested settings in the original papers because they are not of direct interest to us.

[7] https://github.com/HiLab-git/SSL4MIS/tree/master/data/ACDC

[8] https://github.com/ycwu1997/SS-Net/tree/main/data/LA

Table 4: Ablation on data augmentation strategies: (1) random rotation; (2) random cropping; and (3) horizontal flipping, compared to our methods with two settings (*i.e.*, w/ no augmentation and w/ all three augmentation). Note that we use the identical data augmentation strategies (*i.e.*, random rotation, random cropping, and horizontal flipping), as [27, 85, 86, 87, 88, 5, 89, 6, 52, 90, 57, 46, 54, 91, 49, 13, 14] for fair comparisons.

Method	DSC[%]↑	ASD[voxel]↓
● ARCO-SAG (ours)	84.9	1.47
w/o random rotation	84.1	2.87
w/o random cropping	84.4	3.47
w/o horizontal flipping	73.7	6.70
w/ no augmentation	70.8	9.83
○ ARCO-SG (ours)	85.5	0.947
w/o random rotation	83.8	3.27
w/o random cropping	84.6	1.62
w/o horizontal flipping	78.8	4.15
w/ no augmentation	76.2	7.74

For 2D medical segmentation, we follow the same data augmentations [6, 49] to the *teacher*'s input and the *student*'s input, respectively. The augmentations include random rotation, random cropping, and horizontal flipping. For 3D medical segmentation, we follow the same data augmentations [5, 6, 91, 49] to the *teacher*'s input and the *student*'s input, respectively. The augmentations include random rotation and random flipping. We evaluate our methods on 3D segmentation results with two classical metrics: (1) Dice coefficient (DSC) and (2) Average Symmetric Surface Distance (ASD). Note that, for all the evaluated methods, we make no additional modifications during the training process for fair evaluations. We run all our experiments in the same environments with fixed random seeds (Hardware: Single NVIDIA GeForce RTX 3090 GPU; Software: PyTorch 1.10.2+cu113, and Python 3.8.11).

D 2D/3D Methods in Comparison

2D Medical Segmentation: For experiments, we use 2D UNet [27] as backbone, and compare ARCO with previous state-of-the-art medical segmentation methods: (1) *Baseline* (*i.e.*, UNet-F/UNet-L [27]) using both fully-supervised and limited-supervised settings; and (2) *SSL-based*: EM [85], CCT [86], DAN [87], URPC [52], DTC [6], DCT [88], ICT [90], MT [54], UAMT [5], SASSNet [89], CPS [57], GCL [46], MC-Net [91], SS-Net [49], ACTION [13], and MONA [14]. Note that among all the above evaluated methods, several methods use a contrastive objective, including GCL [46], SS-Net [49], ACTION [13], and MONA [14].

3D Medical Segmentation: For experiments, we use 3D VNet [73] as backbone, and compare ARCO with previous state-of-the-art medical segmentation methods: (1) *Baseline* (*i.e.*, VNet-F/VNet-L [73]) using both fully-supervised and limited-supervised settings; and (2) *SSL-based*: EM [85], CCT [86], DAN [87], URPC [52], DTC [6], DCT [88], ICT [90], MT [54], UAMT [5], SASSNet [89], CPS [57], GCL [46], MC-Net [91], SS-Net [49], ACTION [13], and MONA [14]. Note that among all the above evaluated methods, several methods use a contrastive objective, including GCL [46], SS-Net [49], ACTION [13], and MONA [14].

E Framework Overview

In the following, we provide a concise overview of ARCO, consisting of two training phases: (1) relational semi-supervised pre-training, and (2) anatomical contrastive fine-tuning. Note that, in this paper, sampling strategies in pixel-level CL framework are of direct interest, so we use a simplification of MONA [14] without using any additional complicated augmentation strategies.

(1) Relational Semi-Supervised Pretraining. Given an unlabeled sample, ARCO has two ways to define *augmented* and *mined* views: (1) ARCO augments the samples to be $\mathbf{x}^1$ and $\mathbf{x}^2$ as *augmented* views, with two separate data augmentation operators; and (2) ARCO randomly samples d *mined* views (*i.e.*, $\mathbf{x}^3$) from the unlabeled dataset with additional augmentation. The pairs $\left[\mathbf{x}^1, \mathbf{x}^2\right]$ are then processed by $[F_s, F_t]$, and in a similar way $\mathbf{x}^3$ is processed by F_t (See Figure 1(a) in main paper), outputting three global features $\left[\mathbf{h}^1, \mathbf{h}^2, \mathbf{h}^3\right]$ after E and their local features $\left[\mathbf{f}^1, \mathbf{f}^2, \mathbf{f}^3\right]$ after

Table 5: Quantitative comparisons (DSC[%] ↑ / ASD[voxel] ↓) across the three labeled ratio settings (1%, 5%, 10%) on the LiTS benchmark. All experiments are conducted as [27, 85, 86, 87, 88, 5, 89, 6, 52, 90, 57, 46, 54, 91, 49, 13, 14] in the identical setting for fair comparisons. Best and second-best results are coloured **blue** and red, respectively. UNet-F (fully-supervided) and UNet-L (semi-supervided) are considered as the upper bound and the lower bound for the performance comparison. Please refer to the text for discussion. We adopt the identical data augmentation (*i.e.*, random rotation, random cropping, and horizontal flipping) for fair comparisons.

	LiTS								
		1% Labeled			5% Labeled			10% Labeled	
Method	Average	Liver	Lesion	Average	Liver	Lesion	Average	Liver	Lesion
UNet-F [27]	68.2/ 16.9	90.6/8.14	45.8/25.6	68.2/16.9	90.6/8.14	45.8/25.6	68.2/16.9	90.6/8.14	45.8/25.6
UNet-L	50.8/30.4	76.5/15.2	25.0/45.6	63.4/30.4	90.5/9.84	36.4/50.9	64.6/28.3	88.4/18.6	40.9/37.9
EM [85]	56.6/38.4	86.4/26.3	26.9/50.5	61.2/33.3	87.6/9.47	34.7/57.1	62.9/38.5	87.4/21.3	38.3/55.7
CCT [86]	52.4/52.3	79.1/42.0	25.7/62.6	61.4/26.1	84.6/12.3	38.2/39.9	63.4/23.0	88.8/7.64	38.0/38.5
DAN [87]	57.1/28.3	84.5/19.2	29.6/37.4	62.3/25.8	88.6/9.64	36.1/42.1	63.2/30.7	87.3/15.4	39.1/46.1
URPC [52]	55.5/34.6	83.0/27.3	28.0/42.0	60.9/21.4	85.5/8.11	36.3/34.6	62.5/24.8	84.9/12.9	40.1/36.7
DTC [6]	39.3/37.5	68.1/10.7	10.5/64.3	59.2/18.6	85.0/7.54	33.3/29.7	62.5/24.8	84.9/12.9	40.1/36.7
DCT [88]	57.7/38.5	87.0/22.4	28.3/54.6	60.8/34.4	89.2/12.6	32.5/56.2	61.9/31.7	86.2/19.3	37.5/44.1
ICT [90]	58.3/32.2	86.5/22.8	30.1/41.5	60.1/39.1	86.8/12.6	33.3/65.6	62.5/32.4	88.1/16.7	36.9/48.2
MT [54]	54.7/24.7	83.1/10.2	26.3/39.1	60.9/23.7	87.5/**6.34**	34.4/41.1	62.3/23.7	88.5/9.32	36.1/38.1
UAMT [5]	55.5/34.6	83.0/27.3	28.0/42.0	61.5/24.7	84.5/10.6	38.6/38.8	62.9/23.6	87.4/7.78	38.4/39.6
SASSNet [89]	39.6/42.7	69.0/14.7	10.3/7.06	60.4/25.3	86.1/11.6	34.7/39.0	62.4/21.1	86.4/8.31	38.3/33.9
CPS [57]	57.7/39.6	87.0/22.4	28.3/54.6	59.5/26.3	84.0/9.01	34.9/43.5	62.1/30.8	88.6/18.3	35.6/43.3
GCL [46]	59.3/29.5	88.6/14.2	30.0/44.9	63.3/20.1	90.7/9.46	35.9/30.8	65.0/37.2	91.3/10.0	38.7/64.3
MC-Net [91]	60.9/32.1	87.1/17.8	34.8/46.5	61.6/19.8	86.3/8.21	36.9/31.4	63.4/29.9	89.0/13.1	38.0/46.8
SS-Net [49]	55.0/35.9	89.6/19.8	20.5/51.9	59.1/24.6	87.5/11.2	30.6/38.1	63.4/19.8	91.1/7.33	35.8/32.2
ACTION [13]	61.0/24.5	89.8/16.9	32.2/32.3	66.3/23.6	93.0/6.41	39.5/40.8	67.2/20.4	92.8/5.08	41.6/35.8
MONA [14]	62.2/24.9	91.3/13.9	34.0/35.9	66.6/16.6	93.1/7.74	40.1/25.4	68.3/18.0	93.4/8.88	43.3/27.0
● ARCO-SAG (ours)	64.1/20.6	91.5/7.63	36.8/**33.5**	67.3/13.2	93.0/7.10	41.7/19.3	69.4/14.9	93.3/7.60	45.5/22.2
○ ARCO-SG (ours)	**65.5**/**20.2**	**92.5**/**5.59**	**38.4**/34.7	**68.4**/**11.3**	**93.7**/6.63	**43.0**/**16.0**	**70.1**/**13.5**	**94.1**/**5.01**	**46.1**/**22.1**

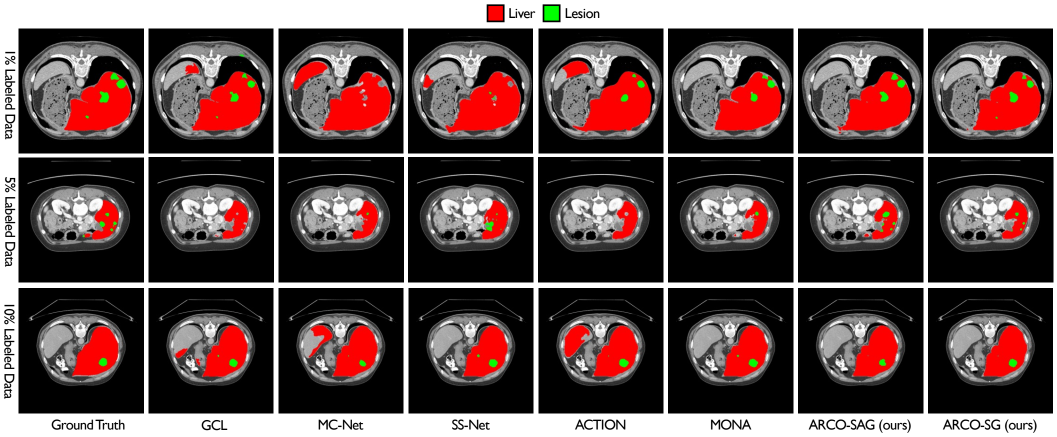


Figure 6: Visual results on LiTS with 1%, 5%, 10% label ratios. As is shown, ARCO-SG and ARCO-SAG consistently produce more accurate predictions on anatomical regions and boundaries compared to all other SSL methods.

D. These features are fed to the two-layer non-linear projectors for outputting global and local embeddings $\mathbf{v}_g$ and $\mathbf{v}_l$.

To alleviate the *collapse* issues [23, 24, 13], we make the architecture asymmetric between the *student* and *teacher* pipeline by further feeding both the global and local embeddings $\mathbf{v}$ with respect to the *student* branch into the non-linear predictor, producing $\mathbf{w}$ in both global and local manners [9]. After passing through the nonlinear projectors and predictor, the relational similarities between *augmented* and *mined* embeddings are computed using the softmax transform, which can be formulated as: $\mathbf{u}_s = \log \frac{\exp\left(\mathrm{sim}\left(\mathbf{w}^1, \mathbf{v}^3\right)/\tau_s\right)}{\sum_{n=1}^{N} \exp\left(\mathrm{sim}\left(\mathbf{w}^1, \mathbf{v}_n^3\right)/\tau_s\right)}$, $\mathbf{u}_t = \log \frac{\exp\left(\mathrm{sim}\left(\mathbf{w}^2, \mathbf{v}^3\right)/\tau_t\right)}{\sum_{n=1}^{N} \exp\left(\mathrm{sim}\left(\mathbf{w}^2, \mathbf{v}_n^3\right)/\tau_t\right)}$, where τ_s and τ_t are different temperature parameters. The unsupervised instance discrimination loss (*i.e.*, Kullback-Leibler divergence $\mathcal{KL}$) can be defined as:

$$\mathcal{L}_{\text{inst}} = \mathcal{KL}(\mathbf{u}_s || \mathbf{u}_t). \tag{E.1}$$

[9] We omit details of local instance discrimination setting for simplicity in following contexts.

The parameters of the *teacher* model are the exponential moving average (EMA) of the *student* model parameters that are updated by the stochastic gradient descent. For pretraining, the entire loss consists of the global and local instance discrimination loss, and supervised segmentation loss $\mathcal{L}_{\text{sup}}$ (*i.e.*, equal combination of Dice loss and cross-entropy loss), *i.e.*, $\mathcal{L} = \mathcal{L}_{\text{inst}}^{\text{global}} + \mathcal{L}_{\text{inst}}^{\text{local}} + \mathcal{L}_{\text{sup}}$.

(2) Anatomical Contrastive Finetuning (ACF). We use the pre-trained network weights as the initialization for subsequent fine-tuning (See Figure 1(b) in Main Paper). For semi-supervised training, we follow two principles described in [14] [10]: (1) *tailness*: giving more importance to tail class hard pixels; and (2) *diversity*: ensuring anatomical diversity in the set of different sampled images.

Following the abovementioned principles, we employ a two-step routine: (1) *Tailness*: we first perform anatomical contrastive formulation. Specifically, we additionally attach the representation head ψ_r[11], and generate a higher n-dimensional dense representation with the same spatial resolution as the input image. A pixel-level contrastive loss is designed to pull queries $\mathbf{r}_q \in \mathcal{R}$ to be similar to the positive keys $\mathbf{r}_k^+ \in \mathcal{R}$, and push apart the negative keys $\mathbf{r}_k^- \in \mathcal{R}$. The semi-supervised contrastive loss $\mathcal{L}_{\text{contrast}}$ is defined as:

$$\mathcal{L}_{\text{contrast}} = \sum_{c\in\mathcal{C}} \sum_{\mathbf{r}_q\sim\mathcal{R}_q^c} -\log \frac{\exp(\mathbf{r}_q\cdot\mathbf{r}_k^{c,+}/\tau)}{\exp(\mathbf{r}_q\cdot\mathbf{r}_k^{c,+}/\tau)+\sum_{\mathbf{r}_k^-\sim\mathcal{R}_k^c}\exp(\mathbf{r}_q\cdot\mathbf{r}_k^-/\tau)}, \tag{E.2}$$

where $\mathcal{C}$ is a set including all available classes in the current mini-batch, and τ is a temperature hyperparameter. We refer to $\mathcal{R}_q^c, \mathcal{R}_k^c, \mathbf{r}_k^{c,+}$ as a query set including all representations within this class c, a negative key set including all representations whose labels is not in class c, and the positive key which is the c-class mean representation, respectively. Consider $\mathcal{P}$ is a set including all pixel coordinates with the same resolution as R, these queries and keys are then defined as:

$$\mathcal{R}_q^c = \bigcup_{[i,j]\in\mathcal{A}} \mathbb{1}(\mathbf{y}_{[i,j]}=c)\,\mathbf{r}_{[i,j]},\ \mathcal{R}_k^c = \bigcup_{[i,j]\in\mathcal{A}} \mathbb{1}(\mathbf{y}_{[i,j]}\neq c)\,\mathbf{r}_{[i,j]},\ \mathbf{r}_k^{c,+} = \frac{1}{|\mathcal{R}_q^c|}\sum_{\mathbf{r}_q\in\mathcal{R}_q^c}\mathbf{r}_q\,. \tag{E.3}$$

(2) *Diversity*: we leverage the first-in-first-out (FIFO) memory bank [10] to search for K-nearest neighbors, and use the semi-supervised nearest neighbor loss $\mathcal{L}_{\text{nn}}$ in a way that maximizing cosine similarity, to exploit the inter-instance relationship.

For fine-tuning, the total loss includes contrastive loss $\mathcal{L}_{\text{contrast}}$, nearest neighbors loss $\mathcal{L}_{\text{nn}}$, unsupervised cross-entropy loss $\mathcal{L}_{\text{unsup}}$ and supervised segmentation loss $\mathcal{L}_{\text{sup}}$: $\mathcal{L}_{\text{sup}} + \lambda_1\mathcal{L}_{\text{contrast}} + \lambda_2\mathcal{L}_{\text{unsup}} + \lambda_3\mathcal{L}_{\text{nn}}$. See Appendix N for the ablation of hyperparameters.

F Model Architecture

Figure 1 provides an overview over our approach. Our semi-supervised segmentation model F takes an 2D/3D medical image x as input and outputs the segmentation map and the representation map. We leverage MONA pipeline [14] including two stages: (1) relational semi-supervised pre-training: on labeled data, the *student* network is trained by the ground-truth labels with the supervised loss $\mathcal{L}_{\text{sup}}$; while on unlabeled data, the *student* network takes the *augmened* and *mined* embeddings from the EMA teacher for instance discrimination $\mathcal{L}_{\text{inst}}$ in the global and local manner, (2) our proposed anatomical contrastive reconstruction fine-tuning: on labeled data, the student network is trained by the ground-truth labels with the supervised loss $\mathcal{L}_{\text{sup}}$; while on unlabeled data, the student model takes the representation maps and pseudo labels from the EMA teacher to give more importance to tail class $\mathcal{L}_{\text{contrast}}$, exploit the inter-instance relationship $\mathcal{L}_{\text{nn}}$, and compute unsupervised loss $\mathcal{L}_{\text{unsup}}$. See Appendix M for details of the visualization loss landscapes.

[10] In this paper, sampling strategies in pixel-level contrastive learning frameworks are of direct interest, so we use a simplification of `MONA` [14] – without using any additional complicated augmentation strategies.

[11] The representation head is only applied during training, and is removed during inference

[11] We adopt the identical data augmentation (*i.e.*, random rotation, random cropping, and horizontal flipping) for fair comparisons.

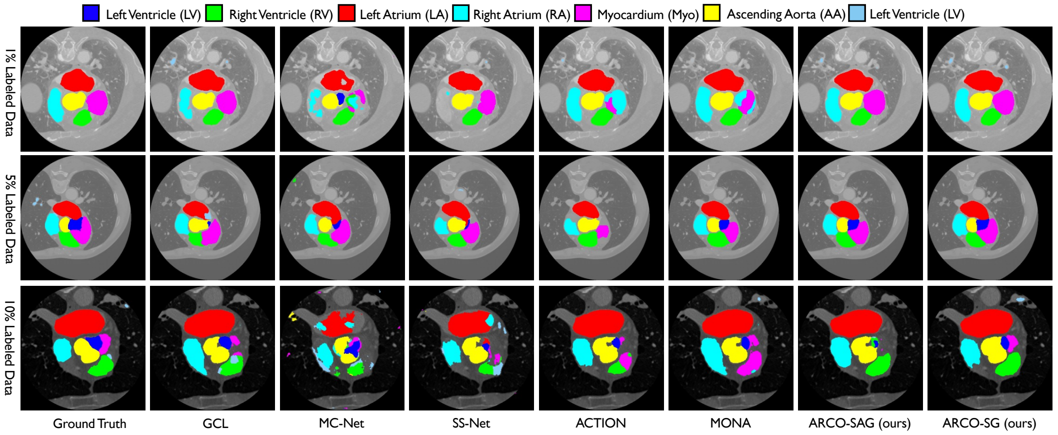


Figure 7: Visual results on MMWHS with 1%, 5%, 10% label ratios. As is shown, ARCO-SG and ARCO-SAG consistently produce more accurate predictions on anatomical regions and boundaries compared to all other SSL methods.

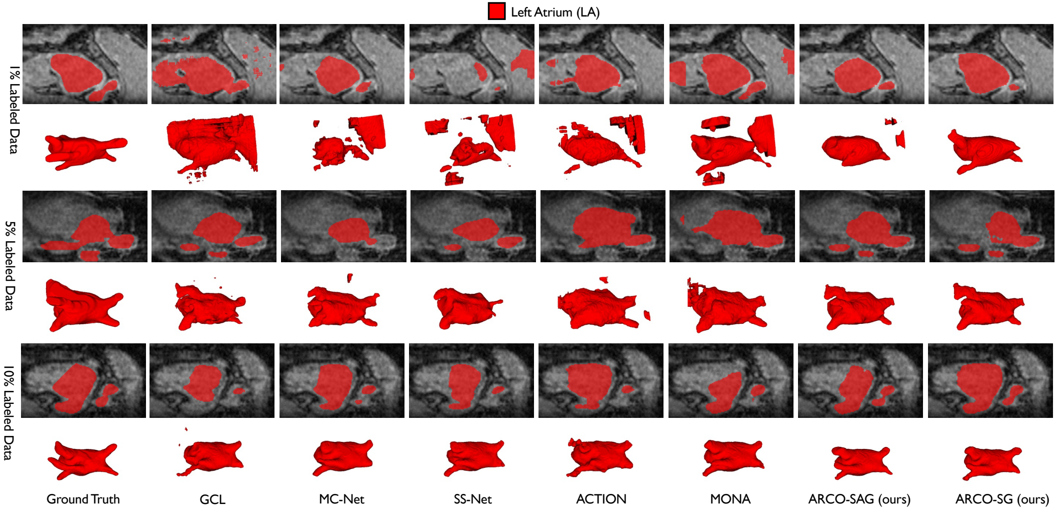


Figure 8: Visual results on LA with 1%, 5%, 10% label ratios. As is shown, ARCO-SG and ARCO-SAG consistently yield more accurate and sharper boundaries compared to all other SSL methods.

G More Experiment Results on LiTS

Figure 6 and Table 5 show qualitative and quantitative results, where our ARCO-SG and ARCO-SAG provide better segmentation results than all other methods. This clearly demonstrates the superiority of our models.

H More Experiment Results on MMWHS

We run the baselines and our methods on the third medical image segmentation dataset (*i.e.*, MMWHS [78]) under various label ratios (*i.e.*, 1%, 5%, 10%), reporting results in Table 6 and Figure 7. This clearly demonstrates the effectiveness of our models.

I More 3D Experiment Results on LA [79]

We run the baselines and our methods on the fourth 3D medical image segmentation dataset (*i.e.*, LA [79]) under various label ratios (*i.e.*, 1%, 5%, 10%), reporting results in Table 9 and Figure 8. This clearly demonstrates the robustness of our models.

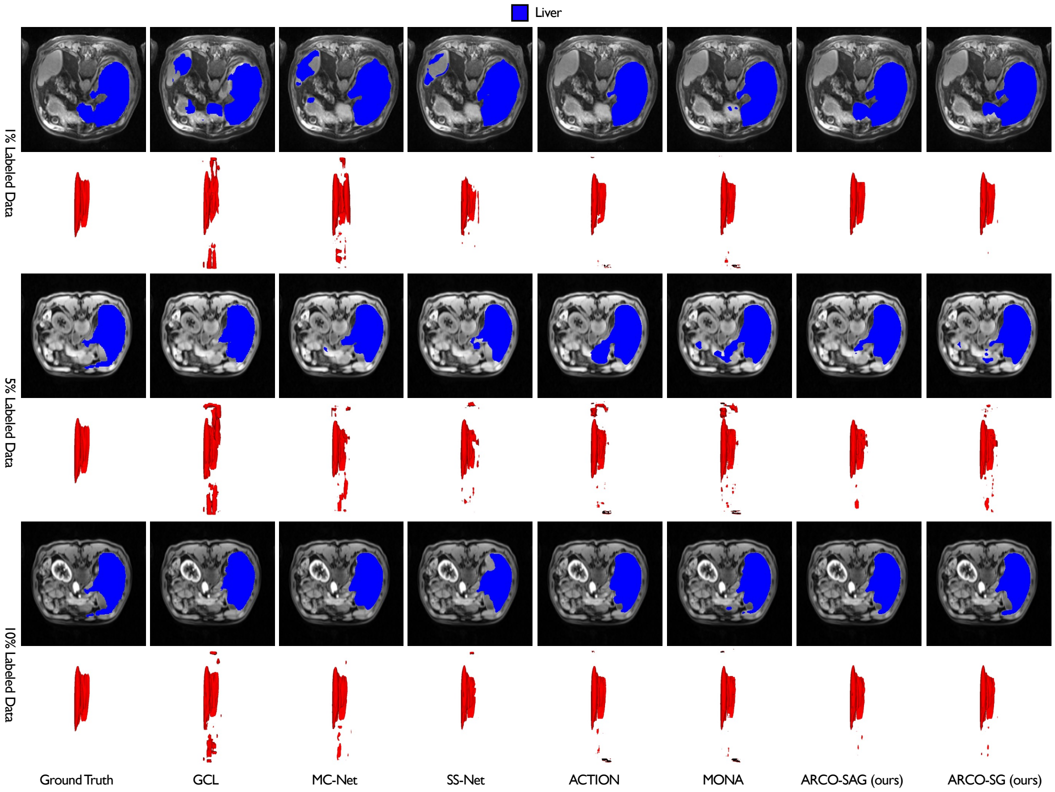


Figure 9: Visual results on MP-MRI with 1%, 5%, 10% label ratios. As is shown, ARCO-SG and ARCO-SAG consistently yield more accurate and sharper boundaries compared to all other SSL methods.

J More 3D Experiment Results on MP-MRI

We run the baselines and our methods on the five 3D medical image segmentation dataset (*i.e.*, MP-MRI) under various label ratios (*i.e.*, 1%, 5%, 10%), reporting results in Table 9 and Figure 9. This clearly shows the superiority of our models.

K More Experiment Results on Semantic Segmentation

To further validate the effectiveness, we experiment on three popular segmentation benchmarks (*i.e.*, Cityscapes [92], Pascal VOC 2012 [93], indoor scene segmentation dataset – SUN RGB-D [94]) in the semi-supervised full-label settings. We follow the identical setting [95] to sample labelled images to ensure that every class appears sufficiently in our three datasets, *i.e.*, CityScapes, Pascal VOC, and SUN RGB-D. Specifically, we aim to have the least frequent class appear in at least 5, 15, and 50 images in each dataset, respectively. In CityScapes, we have at least 12 semantic classes represented in our labeled images, while in Pascal VOC and SUN RGB-D we have at least 3 and 1 semantic classes, respectively. We compare our ARCO-SG and ARCO-SAG under various grid settings (*i.e.*, 9, 16, 25) to baselines (supervised and semi-supervised ReCo). Table 7 shows the mean IoU validation performance. We can see that for all cases, ARCO-SG and ARCO-SAG consistently improve performance, compared to ReCo, in all the semi-supervised settings. For example, under the fewest label setting in each dataset, compared to ReCo, applying stratified group sampling (SG) can improve results by an especially significant margin, with up to 2.4 – 7.8% relative gains. As shown in Pascal VOC 2012 [93] – 60 labels (Figure 10), 200 labels (Figure 11), 600 labels (Figure 12), Cityscapes [92] – 20 labels (Figure 13), 50 labels (Figure 14), 150 labels (Figure 15), SUN RGB-D [94] – 50 labels (Figure 16), 150 labels (Figure 17), 500 labels (Figure 18), ARCO-SG and ARCO-SAG consistently yield more accurate and sharper boundaries compared to ReCo. All those clearly demonstrate the superiority of our models.

L Generalization across Label Ratios and Frameworks

Besides generalizing well across different datasets and diverse labeled settings, we additionally evaluate the performance of SG and SAG sampling coupled with different CL frameworks (*i.e.*, `MoCov2` [98], k`NN-MoCo` [99], `SimCLR` [11], `BYOL` [23], `ISD` [24], `VICReg` [25]) on ACDC under various label ratios (*i.e.*, 1%, 5%, 10%). In this work, we mainly study the state-of-the-art CL frameworks from the computer vision domain hereinafter for ablations, considering their superior performance in the task of image classification. Note that, for each fair comparison, we strictly follow the default setting in these CL frameworks [98, 99, 11, 23, 24, 25] for pretraining, and fine-tune each networks using the same settings in Appendix C. Full experimental details are in Appendix C.

Since they work orthogonally to our pretraining strategy, we conduct a comprehensive comparison of these CL-based frameworks in Table 8. Clearly, our proposed SG and SAG sampling help significantly in improving the segmentation performance across all the CL-based frameworks and is capable of being integrated with previous frameworks for further enhanced **model robustness**. Moreover, it is important to observe that the performance benefits of our methods increase significantly with a lower label setting. This observation augments the necessity of our proposed methods while training networks with high **label efficiency**.

M Final Checkpoint Loss Landscapes

From visualizations in Figure 19, `ARCO-SG` converges to much flatter loss valleys, which evidences their robustness in learning anatomical features.

N Ablation on Different Training Settings

Hyperparameter Selection. For grid search, we detail the tuning steps here. The tuning is done in sequential order. λ_1 is chosen from $\{0.001, 0.005, 0.01, 0.05, 0.1, 1.0\}$, and λ_2, λ_3 are chosen from $\{0.1, 0.2, 0.5, 1.0, 2.0, 10.0\}$. We use the validation set to search over hyperparameters and find the best hyperparameter on ACDC with 1% labeled ratio. As shown in Figure 20, with a carefully tuned hyperparameters λ_1=0.01, λ_2=1.0, and λ_3=1.0, such setting achieves superior performance compared to others.

Table 6: Quantitative comparisons (DSC[%] ↑ / ASD[voxel] ↓) across the three labeled ratio settings (1%, 5%, 10%) on the MMWHS benchmark. All experiments are conducted as [27, 85, 86, 87, 88, 5, 89, 6, 52, 90, 57, 46, 54, 91, 49, 13, 14] in the identical setting for fair comparisons. Best and second-best results are coloured **blue** and red, respectively. `UNet-F` (fully-supervided) and `UNet-L` (semi-supervided) are considered as the upper bound and the lower bound for the performance comparison. Please refer to the text for discussion. Note that, Left Ventricle → LV, Right Ventricle → RV, Left Atrium → LA, Right Atrium → RA, Myocardium → Myo, Ascending Aorta → AA, Pulmonary Artery → PA. We adopt the identical data augmentation (*i.e.*, random rotation, random cropping, and horizontal flipping) for fair comparisons.

Method (1% Labeled)	Average	LV	RV	LA	RA	Myo	AA	PA
`UNet-F` [27]	85.8/8.01	87.0/4.11	79.5/14.7	92.7/4.96	81.6/13.1	83.9/9.32	95.0/3.33	81.1/6.46
`UNet-L`	58.3/33.9	73.5/28.2	57.9/37.6	74.9/30.2	47.2/65.4	61.9/27.8	74.0/18.7	18.6/29.4
`EM` [85]	59.5/63.2	71.7/53.0	64.7/26.3	66.6/51.5	61.7/39.4	56.8/66.0	81.9/48.8	12.9/157.2
`CCT` [86]	62.8/27.5	78.1/18.3	62.8/45.1	83.0/18.4	45.3/69.9	67.0/18.9	76.3/10.6	26.7/11.2
`DAN` [87]	63.8/39.0	76.3/15.4	62.2/58.3	68.0/25.3	52.3/48.2	57.0/54.3	89.0/28.0	42.0/43.4
`URPC` [52]	65.7/29.7	77.0/29.5	65.1/33.4	87.6/18.5	52.1/56.3	65.8/28.5	85.5/22.7	27.3/29.7
`DTC` [6]	62.9/32.3	78.0/29.6	63.6/29.4	74.6/30.8	49.4/53.5	67.3/28.1	89.0/10.0	18.2/44.4
`DCT` [88]	60.0/35.3	72.1/31.6	55.6/42.9	79.0/24.0	54.6/65.7	62.7/31.4	67.3/26.7	29.2/35.3
`ICT` [90]	59.9/32.8	77.1/15.5	53.4/41.8	79.7/25.4	44.1/69.0	62.9/28.5	74.1/20.8	28.4/28.8
`MT` [54]	61.3/36.0	70.7/37.5	62.8/29.0	74.0/49.6	52.8/58.5	58.7/27.0	85.3/13.0	24.9/37.5
`UAMT` [5]	61.1/37.6	75.1/31.9	60.3/49.2	79.5/30.7	50.8/81.8	62.5/33.1	76.6/11.7	23.0/25.0
`SASSNet` [89]	62.5/33.9	76.6/27.6	62.8/29.0	74.0/49.6	52.8/58.5	58.7/27.0	85.3/13.0	24.9/37.5
`CPS` [57]	69.2/22.6	77.9/20.1	67.9/25.3	87.8/18.0	67.6/27.7	66.7/14.2	91.5/20.3	24.8/32.5
`GCL` [46]	71.6/20.3	83.5/10.9	66.0/31.7	91.2/5.29	58.9/48.0	73.2/16.9	89.8/4.59	38.6/24.8
`MC-Net` [91]	65.4/56.9	67.6/59.0	59.6/45.8	68.8/29.6	61.7/62.8	53.8/56.8	84.4/80.5	61.9/63.6
`SS-Net` [49]	60.6/30.7	67.5/15.5	66.0/42.5	73.2/25.1	53.7/52.9	59.5/15.0	81.0/29.5	23.2/34.5
`ACTION` [13]	78.2/10.5	88.2/3.95	71.0/11.3	89.4/**3.00**	74.1/32.5	79.0/8.52	83.2/3.82	62.3/10.2
`MONA` [14]	82.2/8.05	89.2/3.64	75.8/6.98	89.9/3.31	77.6/27.4	79.5/3.70	94.4/**2.72**	69.4/8.61
● `ARCO-SAG` (ours)	86.1/8.78	89.9/3.54	80.9/8.79	92.1/3.22	83.1/24.1	84.0/7.36	94.8/5.41	77.5/9.02
○ `ARCO-SG` (ours)	**87.3/5.79**	**90.7/3.43**	**82.5/5.58**	**92.9**/3.13	**86.2/15.0**	**85.6/2.89**	**95.2**/2.85	**77.9/7.62**
Method (5% Labeled)	Average	LV	RV	LA	RA	Myo	AA	PA
`UNet-F` [27]	85.8/8.01	87.0/4.11	79.5/14.7	92.7/4.96	81.6/13.1	83.9/9.32	95.0/3.33	81.1/6.46
`UNet-L`	72.3/31.1	78.5/31.2	69.0/25.6	78.1/23.8	57.0/45.3	69.4/34.5	90.2/14.5	63.9/42.5
`EM` [85]	80.6/17.3	82.0/22.7	75.3/26.5	87.9/19.2	72.8/19.4	74.8/19.1	94.1/6.52	77.6/7.68
`CCT` [86]	79.0/21.9	82.9/15.3	73.5/31.6	83.3/14.4	75.5/34.0	77.9/24.3	92.7/17.6	67.0/16.1
`DAN` [87]	79.4/22.7	80.1/37.0	77.2/30.0	83.1/12.3	74.4/13.9	78.9/27.4	92.2/28.9	69.5/22.7
`URPC` [52]	76.3/25.5	84.2/18.8	71.3/25.2	78.5/18.2	63.4/39.4	72.0/24.9	93.5/17.3	71.0/34.9
`DTC` [6]	76.4/21.3	82.3/17.2	72.4/24.4	76.1/15.2	65.0/31.8	75.2/20.9	92.8/13.1	70.8/26.7
`DCT` [88]	80.8/23.0	84.0/45.4	75.7/26.0	87.9/12.1	73.9/31.7	77.2/34.1	94.6/5.26	72.5/6.45
`ICT` [90]	77.9/18.8	84.1/13.2	76.7/26.3	79.2/14.0	66.5/24.9	74.1/18.8	94.2/6.21	70.3/28.3
`MT` [54]	77.5/24.2	83.5/15.4	72.8/29.0	78.0/16.2	68.9/39.2	74.7/24.4	93.3/11.8	71.1/33.4
`UAMT` [5]	76.2/21.1	83.4/20.7	71.5/24.4	77.0/14.6	62.8/30.6	75.8/22.1	93.0/8.91	69.7/26.0
`SASSNet` [89]	75.2/25.0	80.9/25.0	70.8/31.2	80.0/17.0	61.4/40.0	70.0/31.3	92.4/18.9	71.0/21.9
`CPS` [57]	78.3/22.5	83.0/29.6	68.8/27.7	85.0/20.4	73.1/15.5	71.9/35.2	94.7/9.05	71.9/20.2
`GCL` [46]	83.5/7.41	86.4/4.72	78.5/9.79	88.6/4.34	79.8/12.1	81.4/8.07	93.5/3.91	76.4/8.95
`MC-Net` [91]	78.5/23.9	83.7/23.8	74.4/25.7	81.9/16.8	70.8/38.4	74.4/23.9	93.3/14.2	71.1/25.0
`SS-Net` [49]	78.0/25.2	83.0/7.76	74.8/32.0	82.3/20.3	69.6/42.6	71.1/15.8	92.4/25.5	73.2/32.1
`ACTION` [13]	85.4/6.71	88.2/3.09	78.8/9.66	90.5/2.84	80.6/15.6	84.4/7.37	94.0/2.56	81.3/5.86
`MONA` [14]	87.3/6.62	90.2/2.92	80.9/9.62	92.8/2.65	82.5/15.2	87.0/8.53	95.3/1.86	82.7/5.60
● `ARCO-SAG` (ours)	88.6/6.73	91.1/**2.70**	83.4/13.0	92.9/2.84	84.3/16.1	89.0/4.68	95.9/1.57	83.3/6.19
○ `ARCO-SG` (ours)	**89.3/4.80**	**91.2/2.70**	**84.6/8.30**	**93.7/2.49**	**85.6/10.4**	**89.2/3.41**	**96.0/1.42**	**84.7/4.95**
Method (10% Labeled)	Average	LV	RV	LA	RA	Myo	AA	PA
`UNet-F` [27]	85.8/8.01	87.0/4.11	79.5/14.7	92.7/4.96	81.6/13.1	83.9/9.32	95.0/3.33	81.1/6.46
`UNet-L`	77.8/19.7	82.8/8.92	77.3/16.1	75.9/22.8	74.6/24.7	75.3/11.6	90.8/21.6	67.8/32.3
`EM` [85]	82.1/15.1	86.7/19.8	78.4/24.5	88.1/7.46	77.6/15.9	75.8/25.1	95.0/4.13	73.2/8.86
`CCT` [86]	79.4/16.3	85.4/5.65	73.5/30.0	89.1/7.10	68.2/31.2	70.2/24.3	92.4/5.90	77.1/9.77
`DAN` [87]	80.2/15.0	81.6/22.6	74.2/21.2	85.0/10.1	75.5/17.1	76.9/20.2	94.0/4.40	74.4/9.15
`URPC` [52]	81.9/12.3	88.1/9.41	68.3/20.7	88.1/6.73	76.6/14.3	80.4/19.6	94.5/4.26	77.2/11.0
`DTC` [6]	79.5/20.6	82.8/10.8	75.8/18.7	85.9/13.8	75.2/41.5	74.4/13.9	90.7/25.1	71.7/20.1
`DCT` [88]	82.8/12.5	85.4/11.6	78.0/23.3	89.0/4.30	79.0/16.5	75.5/19.1	94.3/4.42	78.4/8.08
`ICT` [90]	82.2/12.0	88.4/5.11	75.0/13.5	89.0/6.98	75.2/26.4	79.6/20.4	94.9/4.29	73.3/7.30
`MT` [54]	79.4/19.8	80.4/24.1	70.3/21.3	86.0/18.0	80.0/17.0	73.3/28.7	92.3/20.9	73.8/8.92
`UAMT` [5]	83.7/14.2	86.7/12.3	80.3/20.6	89.6/8.10	79.5/19.2	79.2/19.6	93.9/10.3	73.8/8.92
`SASSNet` [89]	81.8/15.5	84.9/8.01	78.3/15.9	84.4/12.5	79.3/27.3	79.0/14.6	93.4/8.30	73.3/22.3
`CPS` [57]	82.0/13.1	84.4/9.85	78.5/21.1	85.9/6.61	81.0/18.7	76.4/18.3	93.2/7.04	74.9/10.3
`GCL` [46]	86.7/8.76	90.5/2.95	81.3/19.6	90.4/4.31	83.1/18.1	86.7/5.84	94.8/2.00	80.3/8.53
`MC-Net` [91]	81.9/15.4	85.4/5.78	80.1/17.2	81.5/11.1	79.7/34.1	79.8/10.9	93.1/6.28	73.7/22.4
`SS-Net` [49]	82.3/13.9	85.7/8.80	79.5/17.6	84.1/12.1	80.2/20.0	81.0/14.0	93.6/8.60	72.0/16.1
`ACTION` [13]	86.1/5.93	88.9/3.25	81.3/6.99	89.4/3.13	81.6/14.1	87.8/3.76	94.4/2.53	79.4/7.78
`MONA` [14]	87.6/6.83	90.7/2.89	82.8/8.99	91.8/3.48	85.2/15.7	87.2/5.32	94.9/4.32	80.4/7.13
● `ARCO-SAG` (ours)	89.3/**4.42**	**91.1**/2.97	84.8/**6.30**	**94.1/2.38**	86.1/**9.73**	89.2/2.89	**96.0**/1.68	**83.6/5.02**
○ `ARCO-SG` (ours)	**89.4**/4.80	**91.6/2.56**	**85.0**/6.79	93.9/2.53	**86.3**/11.5	**89.6/2.71**	95.9/**1.66**	**83.6**/5.88

Table 7: Quantitative comparisons (Intersection Over Union (IoU) ↑) for Pascal VOC, CityScapes and SUN RGB-D datasets. All experiments are conducted as [95] in the identical setting for fair comparisons. Best and second-best results are colored **blue** and red, respectively.

	Pascal VOC				CityScapes				SUN RGB-D			
Method	60 labels	200 labels	600 labels	all labels	20 labels	50 labels	150 labels	all labels	50 labels	150 labels	500 labels	all labels
Supervised	39.4	55.5	64.6	77.8	38.2	45.9	55.4	70.9	20.0	29.2	38.9	51.8
ReCo [95] + ClassMix	57.1	69.4	73.2	-	49.9	57.9	65.0	-	30.5	40.4	44.6	-
● ARCO-SAG (9 Grid) + ClassMix	58.3	70.5	75.4	-	50.2	60.2	66.5	-	31.5	40.9	45.7	-
● ARCO-SAG (16 Grid) + ClassMix	58.7	70.9	75.1	-	50.1	60.6	66.3	-	37.8	40.2	45.7	-
● ARCO-SAG (25 Grid) + ClassMix	59.1	70.9	74.9	-	49.8	60.6	66.7	-	**38.5**	40.5	45.5	-
○ ARCO-SG (9 Grid) + ClassMix	59.2	**71.8**	75.3	-	52.5	60.9	**66.8**	-	32.4	41.4	46.6	-
○ ARCO-SG (16 Grid) + ClassMix	**59.6**	71.7	**75.5**	-	**53.7**	61.2	66.2	-	37.7	41.0	46.4	-
○ ARCO-SG (25 Grid) + ClassMix	59.5	71.7	75.2	-	51.5	**61.8**	66.4	-	38.3	**41.5**	**47.3**	-

Table 8: Ablation on different contrastive learning frameworks with respect to various labeled ratio settings (1%, 5%, 10%). Experiments are conducted on ACDC using UNet [27] as the backbone. Here we report the segmentation performance in terms of DSC[%] and ASD[mm]. On all three labeled settings, incorporating our methods (*i.e.*, SG and SAG) consistently achieve superior model robustness gains compared to naïve sampling across different state-of-the-art CL frameworks.

		1% Labeled		5% Labeled		10% Labeled	
Category	Method	DSC ↑	ASD ↓	DSC ↑	ASD ↓	DSC ↑	ASD ↓
MoCov2 [98]	NS	76,4	5.64	83.3	3.78	83.8	3.17
	● SAG	80.9	4.16	84.7	3.53	84.6	3.84
	○ SG	81.4	4.01	85.1	1.48	84.9	3.08
kNN-MoCo [99]	NS	78.3	4.54	83.8	3.74	84.1	2.97
	● SAG	81.8	4.08	84.6	3.41	84.7	2.58
	○ SG	83.9	3.17	85.4	3.17	85.0	2.47
SimCLR [11]	NS	74.9	4.89	80.9	3.19	84.1	2.78
	● SAG	78.5	4.01	83.8	2.68	85.8	2.01
	○ SG	79.1	3.49	84.3	2.31	86.0	1.76
BYOL [23]	NS	77.3	5.01	82.9	3.14	84.8	1.67
	● SAG	79.8	4.12	85.0	2.81	85.7	1.36
	○ SG	80.2	3.79	85.8	1.67	85.9	1.13
ISD [24]	NS	79.3	3.54	81.2	2.86	85.6	1.90
	● SAG	81.4	3.50	83.1	2.53	86.3	1.61
	○ SG	82.2	2.04	83.7	1.97	86.7	1.34
VICReg [25]	NS	64.0	10.6	79.1	4.18	82.9	3.89
	● SAG	81.1	3.49	83.8	3.02	86.3	2.14
	○ SG	81.6	3.12	84.1	2.78	86.8	2.01
ARCO (ours)	NS	82.6	1.43	86.9	1.07	87.7	1.33
	● SAG	84.9	1.47	87.1	0.848	88.5	1.40
	○ SG	85.5	0.947	88.7	0.841	89.4	0.776

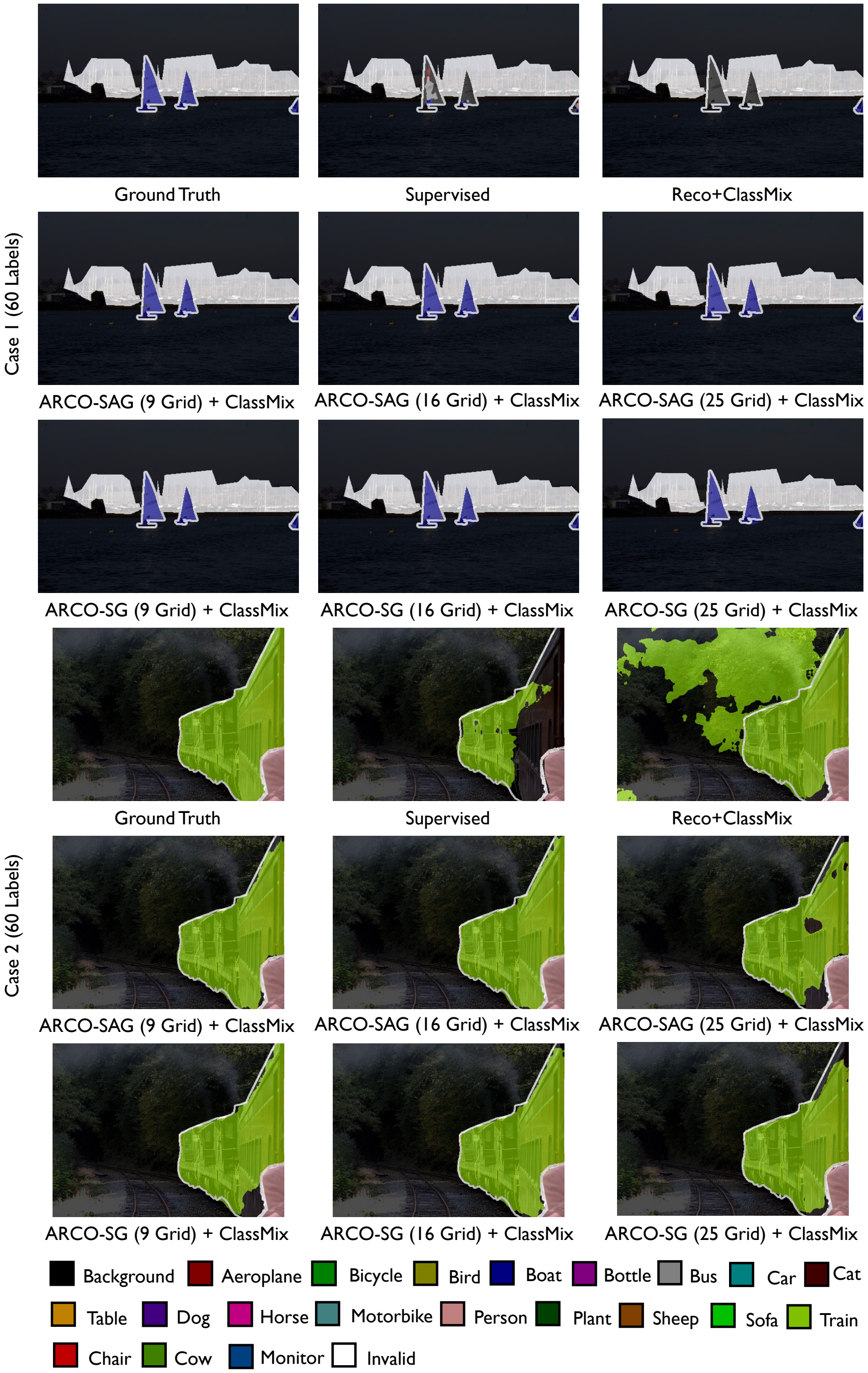

Case 1 (60 Labels)
Ground Truth
Supervised
Reco+ClassMix
ARCO-SAG (9 Grid) + ClassMix
ARCO-SAG (16 Grid) + ClassMix
ARCO-SAG (25 Grid) + ClassMix
ARCO-SG (9 Grid) + ClassMix
ARCO-SG (16 Grid) + ClassMix
ARCO-SG (25 Grid) + ClassMix
Case 2 (60 Labels)
Ground Truth
Supervised
Reco+ClassMix
ARCO-SAG (9 Grid) + ClassMix
ARCO-SAG (16 Grid) + ClassMix
ARCO-SAG (25 Grid) + ClassMix
ARCO-SG (9 Grid) + ClassMix
ARCO-SG (16 Grid) + ClassMix
ARCO-SG (25 Grid) + ClassMix
Background
Aeroplane
Bicycle
Bird
Boat
Bottle
Bus
Car
Cat
Table
Dog
Horse
Motorbike
Person
Plant
Sheep
Sofa
Train
Chair
Cow
Monitor
Invalid

Figure 10: Visual results on Pascal validation set with 60 labels. As is shown, ARCO-SG and ARCO-SAG consistently yield more accurate and sharper boundaries compared to all other SSL methods.

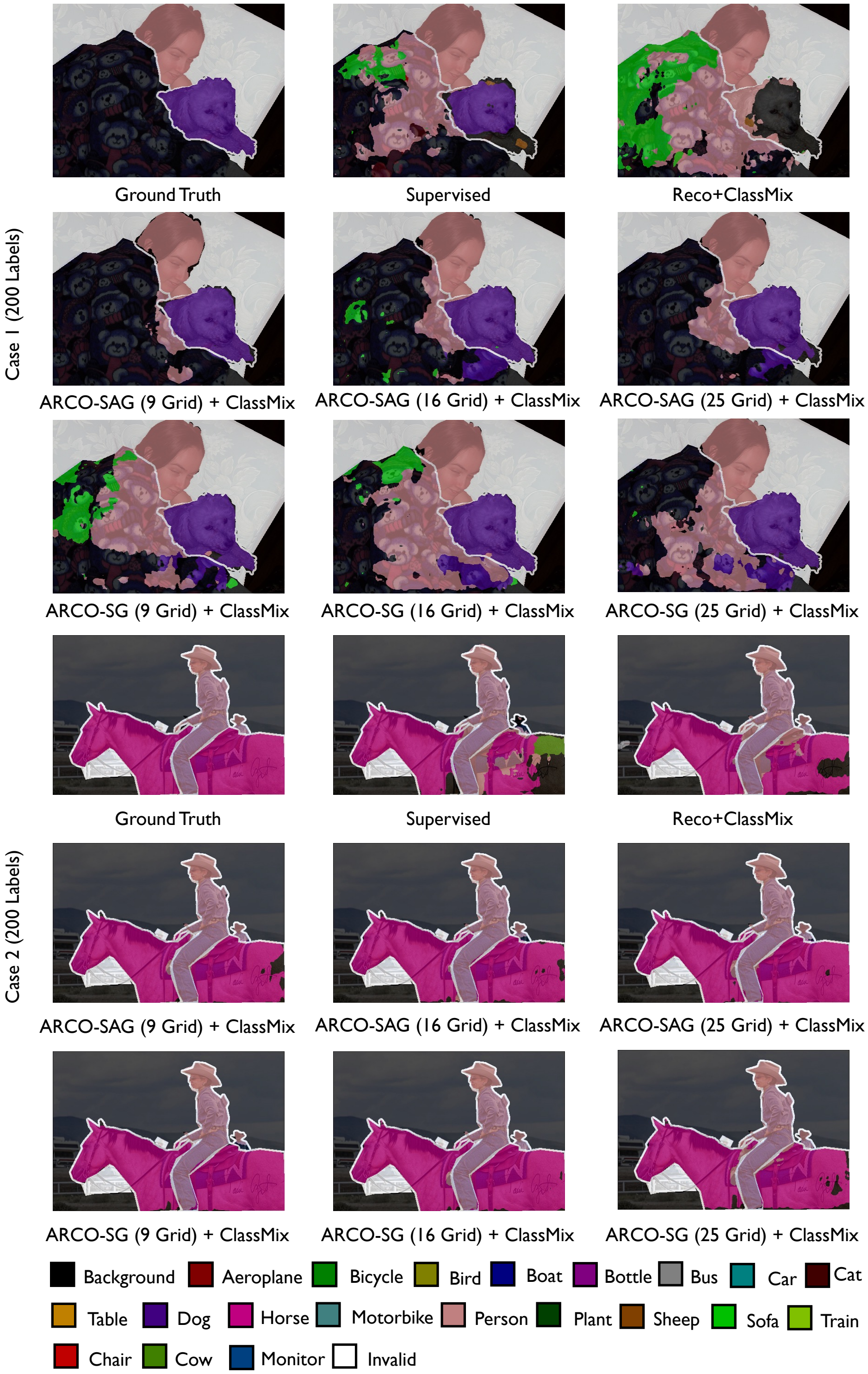

Case 1 (200 Labels)
Ground Truth
Supervised
Reco+ClassMix
ARCO-SAG (9 Grid) + ClassMix
ARCO-SAG (16 Grid) + ClassMix
ARCO-SAG (25 Grid) + ClassMix
ARCO-SG (9 Grid) + ClassMix
ARCO-SG (16 Grid) + ClassMix
ARCO-SG (25 Grid) + ClassMix
Case 2 (200 Labels)
Ground Truth
Supervised
Reco+ClassMix
ARCO-SAG (9 Grid) + ClassMix
ARCO-SAG (16 Grid) + ClassMix
ARCO-SAG (25 Grid) + ClassMix
ARCO-SG (9 Grid) + ClassMix
ARCO-SG (16 Grid) + ClassMix
ARCO-SG (25 Grid) + ClassMix
Background
Aeroplane
Bicycle
Bird
Boat
Bottle
Bus
Car
Cat
Table
Dog
Horse
Motorbike
Person
Plant
Sheep
Sofa
Train
Chair
Cow
Monitor
Invalid

Figure 11: Visual results on Pascal validation set with 200 labels. As is shown, `ARCO-SG` and `ARCO-SAG` consistently yield more accurate and sharper boundaries compared to all other SSL methods.

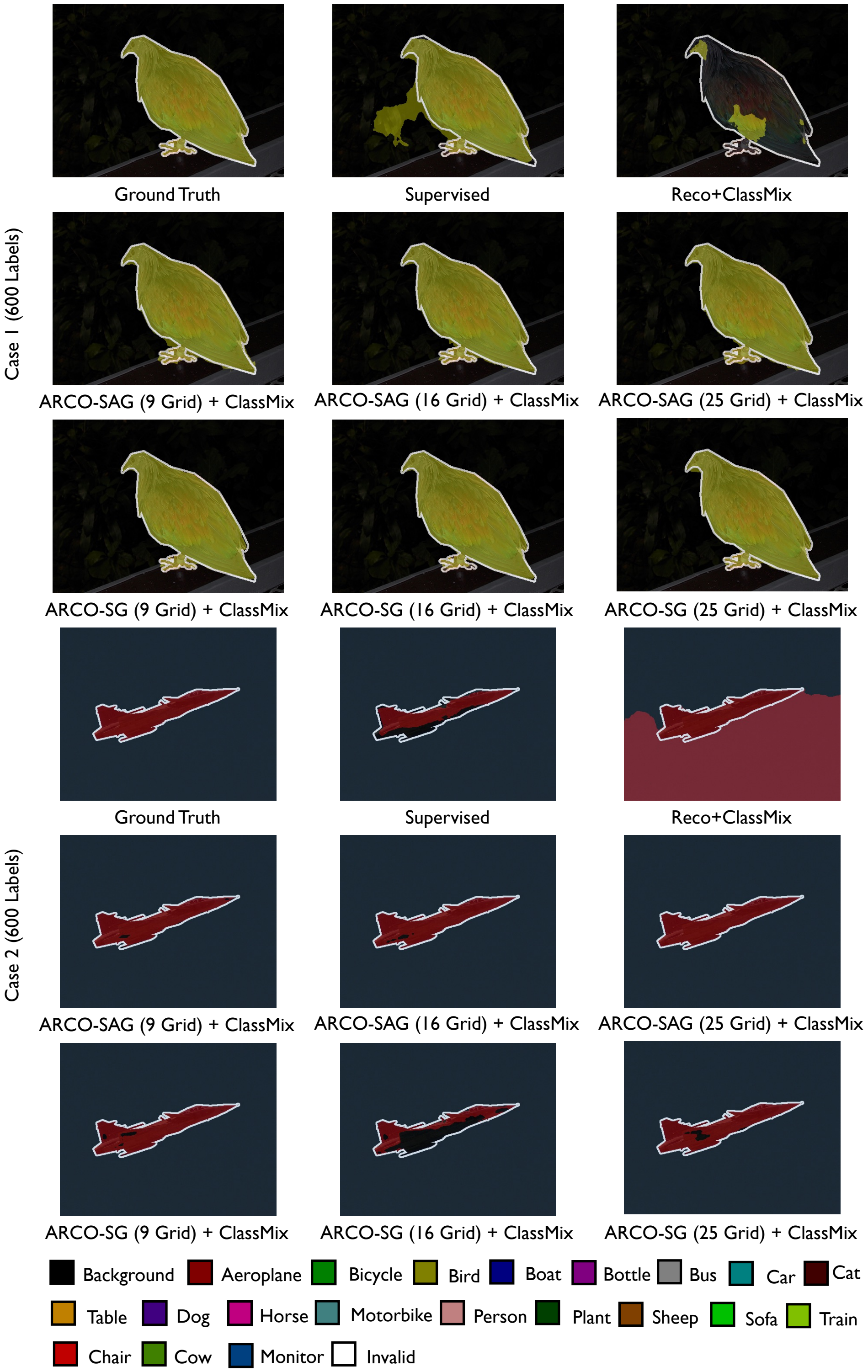

Case 1 (600 Labels)
Ground Truth
Supervised
Reco+ClassMix
ARCO-SAG (9 Grid) + ClassMix
ARCO-SAG (16 Grid) + ClassMix
ARCO-SAG (25 Grid) + ClassMix
ARCO-SG (9 Grid) + ClassMix
ARCO-SG (16 Grid) + ClassMix
ARCO-SG (25 Grid) + ClassMix
Case 2 (600 Labels)
Ground Truth
Supervised
Reco+ClassMix
ARCO-SAG (9 Grid) + ClassMix
ARCO-SAG (16 Grid) + ClassMix
ARCO-SAG (25 Grid) + ClassMix
ARCO-SG (9 Grid) + ClassMix
ARCO-SG (16 Grid) + ClassMix
ARCO-SG (25 Grid) + ClassMix
Background
Aeroplane
Bicycle
Bird
Boat
Bottle
Bus
Car
Cat
Table
Dog
Horse
Motorbike
Person
Plant
Sheep
Sofa
Train
Chair
Cow
Monitor
Invalid

Figure 12: Visual results on Pascal validation set with 600 labels. As is shown, ARCO-SG and ARCO-SAG consistently yield more accurate and sharper boundaries compared to all other SSL methods.

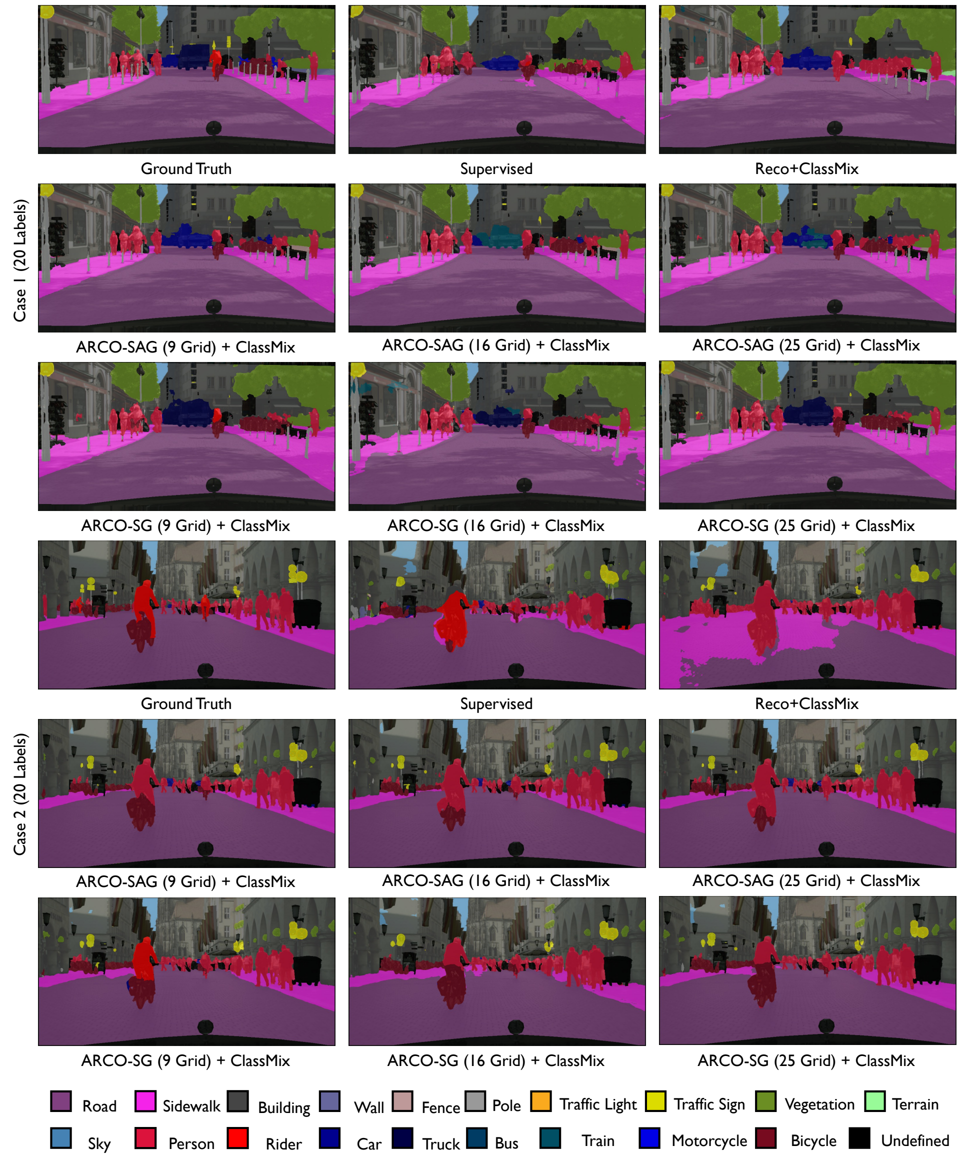
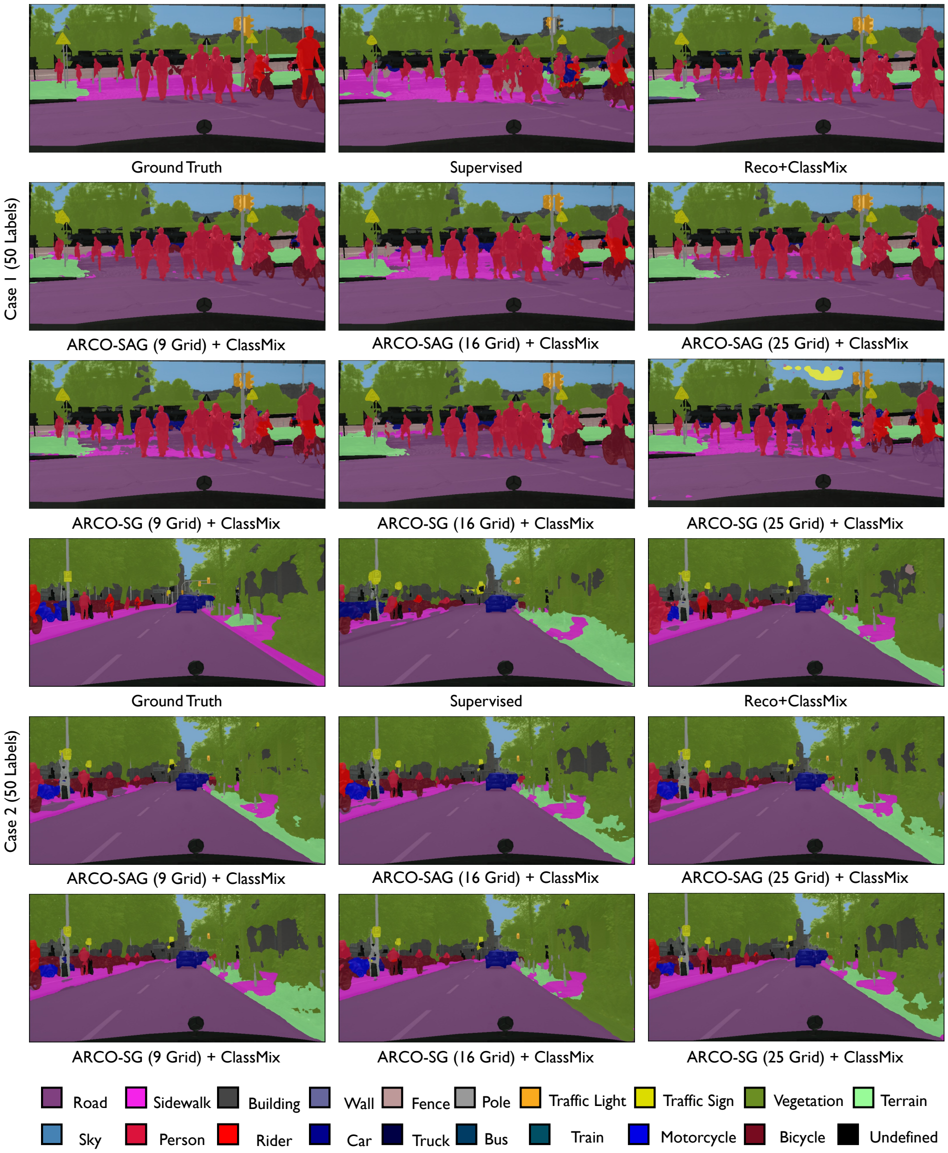


Figure 13: Visual results on Cityscapes validation set with 20 labels. As is shown, ARCO-SG and ARCO-SAG consistently yield more accurate and sharper boundaries compared to all other SSL methods.



Ground Truth

Supervised

Reco+ClassMix

ARCO-SAG (9 Grid) + ClassMix

ARCO-SAG (16 Grid) + ClassMix

ARCO-SAG (25 Grid) + ClassMix

ARCO-SG (9 Grid) + ClassMix

ARCO-SG (16 Grid) + ClassMix

ARCO-SG (25 Grid) + ClassMix

Ground Truth

Supervised

Reco+ClassMix

ARCO-SAG (9 Grid) + ClassMix

ARCO-SAG (16 Grid) + ClassMix

ARCO-SAG (25 Grid) + ClassMix

ARCO-SG (9 Grid) + ClassMix

ARCO-SG (16 Grid) + ClassMix

ARCO-SG (25 Grid) + ClassMix

Figure 14: Visual results on Cityscapes validation set with 50 labels. As is shown, ARCO-SG and ARCO-SAG consistently yield more accurate and sharper boundaries compared to all other SSL methods.

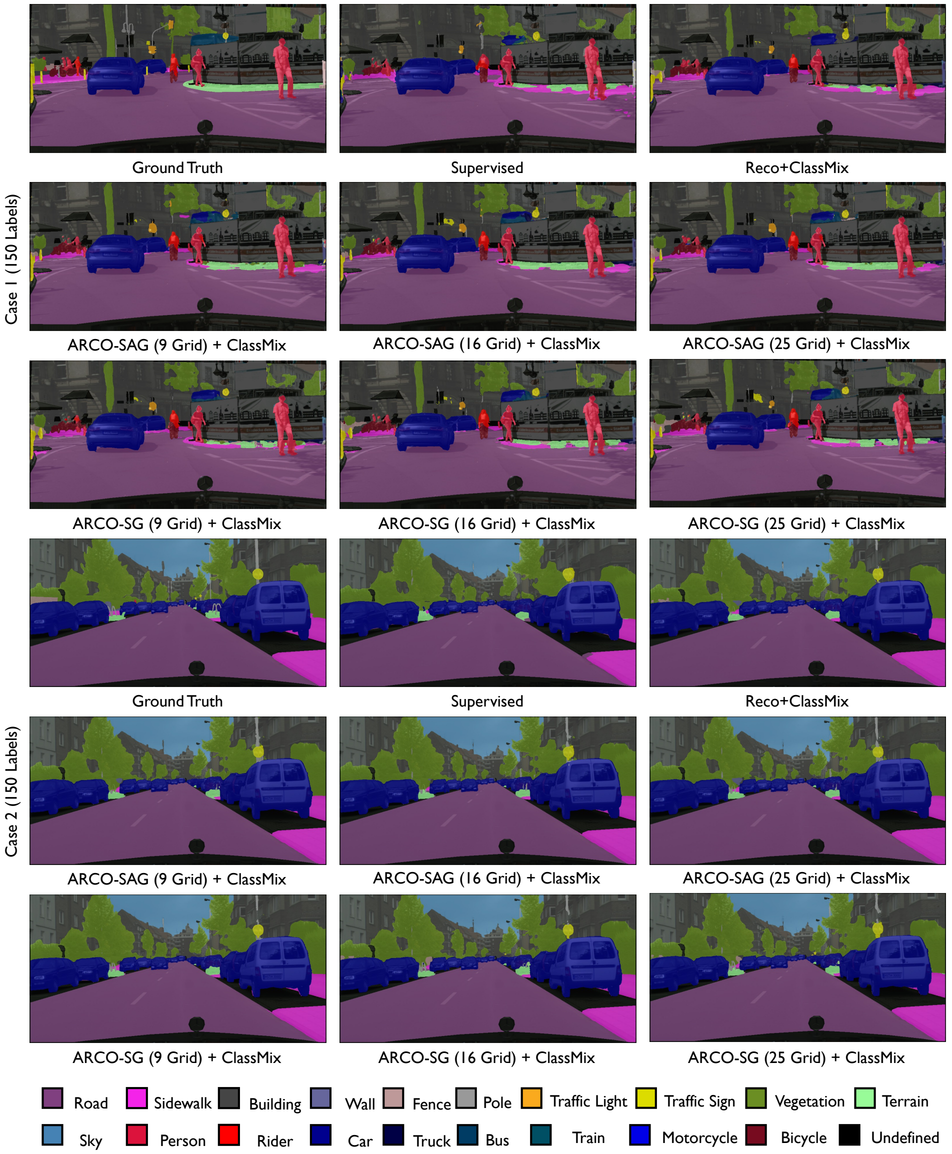

Ground Truth
Supervised
Reco+ClassMix
Case 1 (150 Labels)
ARCO-SAG (9 Grid) + ClassMix
ARCO-SAG (16 Grid) + ClassMix
ARCO-SAG (25 Grid) + ClassMix
ARCO-SG (9 Grid) + ClassMix
ARCO-SG (16 Grid) + ClassMix
ARCO-SG (25 Grid) + ClassMix
Ground Truth
Supervised
Reco+ClassMix
Case 2 (150 Labels)
ARCO-SAG (9 Grid) + ClassMix
ARCO-SAG (16 Grid) + ClassMix
ARCO-SAG (25 Grid) + ClassMix
ARCO-SG (9 Grid) + ClassMix
ARCO-SG (16 Grid) + ClassMix
ARCO-SG (25 Grid) + ClassMix
Road
Sidewalk
Building
Wall
Fence
Pole
Traffic Light
Traffic Sign
Vegetation
Terrain
Sky
Person
Rider
Car
Truck
Bus
Train
Motorcycle
Bicycle
Undefined

Figure 15: Visual results on Cityscapes validation set with 150 labels. As is shown, `ARCO-SG` and `ARCO-SAG` consistently yield more accurate and sharper boundaries compared to all other SSL methods.

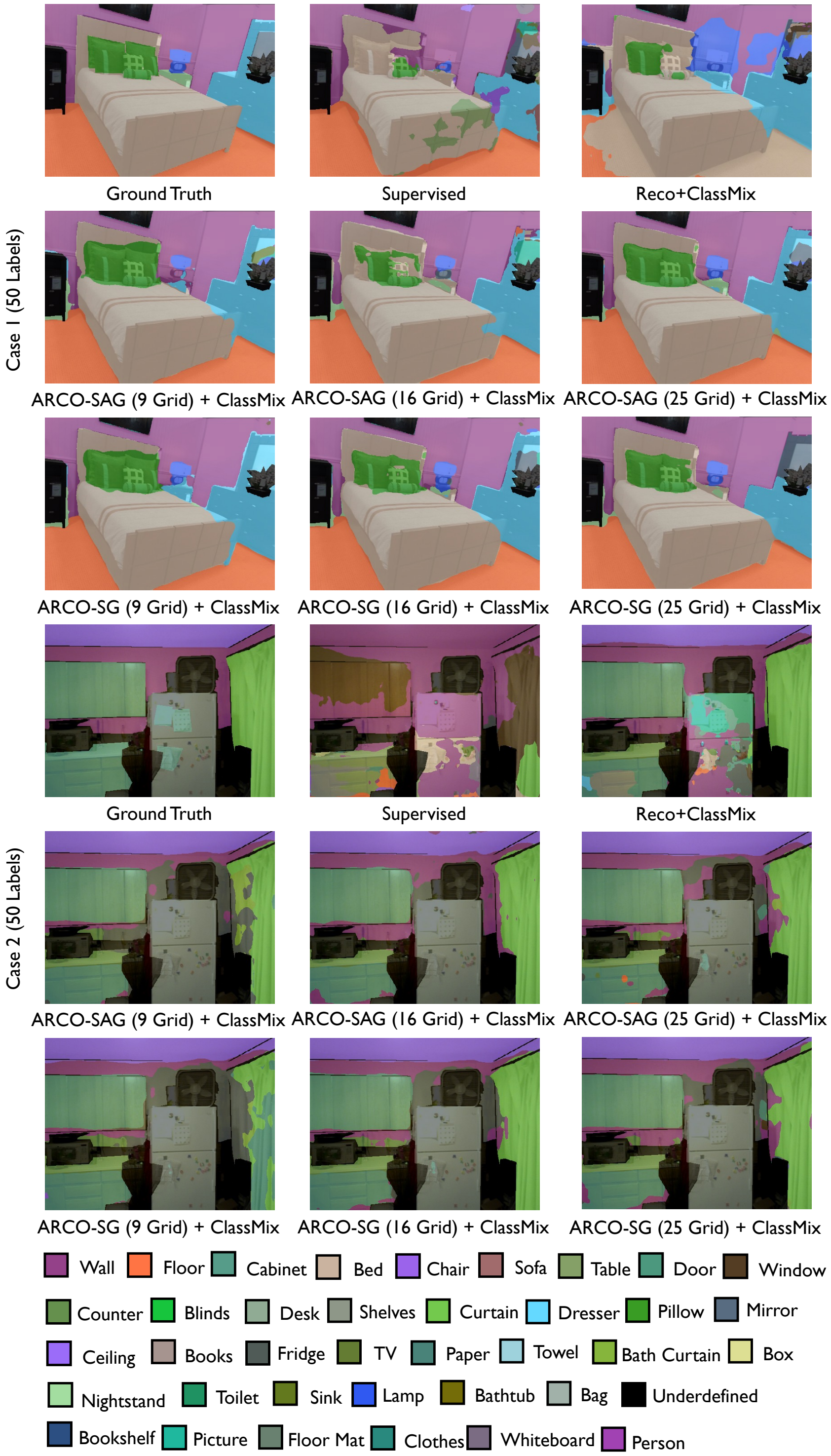

Case 1 (50 Labels)
Ground Truth
Supervised
Reco+ClassMix
ARCO-SAG (9 Grid) + ClassMix
ARCO-SAG (16 Grid) + ClassMix
ARCO-SAG (25 Grid) + ClassMix
ARCO-SG (9 Grid) + ClassMix
ARCO-SG (16 Grid) + ClassMix
ARCO-SG (25 Grid) + ClassMix
Case 2 (50 Labels)
Ground Truth
Supervised
Reco+ClassMix
ARCO-SAG (9 Grid) + ClassMix
ARCO-SAG (16 Grid) + ClassMix
ARCO-SAG (25 Grid) + ClassMix
ARCO-SG (9 Grid) + ClassMix
ARCO-SG (16 Grid) + ClassMix
ARCO-SG (25 Grid) + ClassMix
Wall
Floor
Cabinet
Bed
Chair
Sofa
Table
Door
Window
Counter
Blinds
Desk
Shelves
Curtain
Dresser
Pillow
Mirror
Ceiling
Books
Fridge
TV
Paper
Towel
Bath Curtain
Box
Nightstand
Toilet
Sink
Lamp
Bathtub
Bag
Underdefined
Bookshelf
Picture
Floor Mat
Clothes
Whiteboard
Person

Figure 16: Visual results on SUN RGB-D validation set with 50 labels. As is shown, `ARCO-SG` and `ARCO-SAG` consistently yield more accurate and sharper boundaries compared to all other SSL methods.

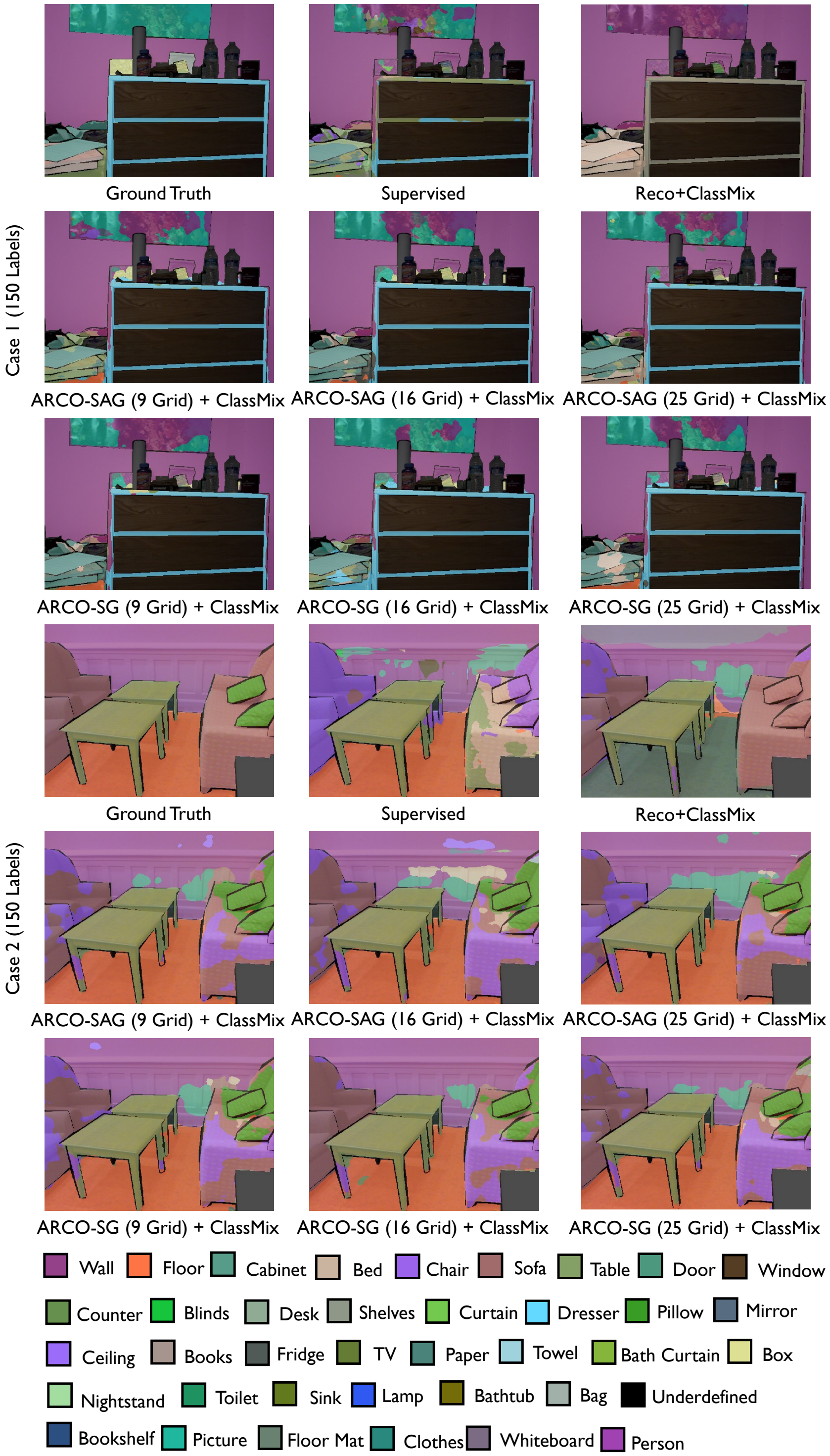


Figure 17: Visual results on SUN RGB-D validation set with 150 labels. As is shown, ARCO-SG and ARCO-SAG consistently yield more accurate and sharper boundaries compared to all other SSL methods.

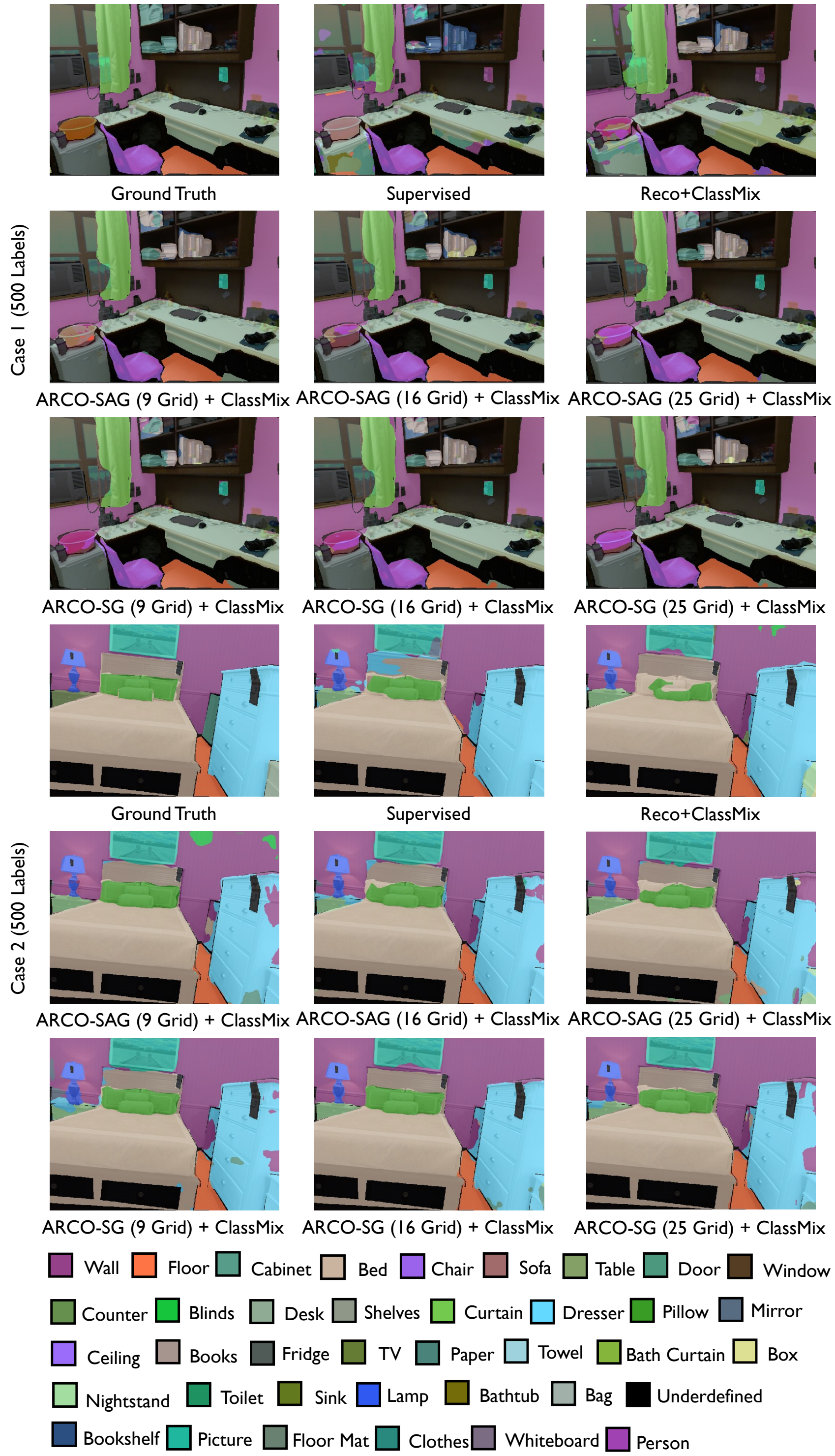


Figure 18: Visual results on SUN RGB-D validation set with 500 labels. As is shown, ARCO-SG and ARCO-SAG consistently yield more accurate and sharper boundaries compared to all other SSL methods.

Table 9: Quantitative comparisons (DSC[%] ↑ / ASD[voxel] ↓) across the three labeled ratio settings (1%, 5%, 10%) on the 3D MP-MRI and 3D LA benchmarks. All 3D experiments are conducted as [73, 85, 86, 87, 88, 5, 89, 6, 52, 90, 57, 46, 54, 91, 49, 13, 14] in the identical setting for fair comparisons. Best and second-best results are coloured **blue** and red, respectively. `VNet-F` (fully-supervided) and `VNet-L` (semi-supervided) are considered as the upper bound and the lower bound for the performance comparison. Please refer to the text for discussion.

	MP-MRI			LA		
	1% Labeled	5% Labeled	10% Labeled	1 Labeled (1%)	4 Labeled (5%)	8 Labeled (10%)
3D Method	Liver	Liver	Liver	Left Atrium (LA)	Left Atrium (LA)	Left Atrium (LA)
`VNet-F` [73]	93.1/5.73	93.1/5.73	93.1/5.73	91.5/1.51	91.5/1.51	91.5/1.51
`VNet-L`	68.6/33.4	81.6/12.5	87.9/7.55	40.0/21.2	52.6/9.87	82.7/3.26
`EM` [85]	73.2/30.1	86.0/15.8	91.9/6.89	48.3/21.3	81.1/4.68	82.7/4.77
`CCT` [86]	74.2/24.5	86.0/15.8	90.9/9.54	40.3/13.8	70.8/8.31	82.0/5.25
`DAN` [87]	69.4/31.4	88.3/10.2	91.1/8.76	38.5/22.0	78.8/6.53	80.2/5.37
`URPC` [52]	72.7/29.9	89.8/9.00	91.7/7.41	65.0/8.97	80.2/5.48	83.1/4.57
`DTC` [6]	78.6/18.8	89.6/10.1	90.5/11.2	36.2/11.7	83.6/2.81	87.1/2.23
`DCT` [88]	74.1/31.1	88.2/12.4	90.1/11.1	42.9/19.1	80.1/9.06	80.4/9.18
`ICT` [90]	72.3/30.5	88.6/11.2	91.1/8.46	47.7/16.0	78.4/6.96	85.4/4.14
`MT` [54]	73.8/29.4	87.7/12.8	92.0/7.15	58.1/17.8	77.0/8.15	82.8/5.90
`UAMT` [5]	71.6/31.2	87.1/12.8	91.3/9.71	60.3/11.3	82.3/3.82	87.8/2.12
`SASSNet` [89]	78.8/19.6	88.4/13.1	88.7/13.1	51.5/14.6	81.6/3.58	87.5/2.59
`CPS` [57]	80.0/17.1	89.2/10.8	91.0/9.16	45.1/22.0	79.7/9.28	80.7/5.16
`GCL` [46]	78.9/14.2	87.9/8.29	90.4/5.68	52.6/12.8	75.5/7.60	84.8/4.22
`MC-Net` [91]	79.7/20.2	90.1/8.27	92.4/**4.34**	44.3/14.1	83.6/2.70	87.6/1.82
`SS-Net` [49]	88.0/8.93	90.9/9.94	92.0/5.67	43.4/14.8	86.3/2.31	88.6/1.90
`ACTION` [13]	86.5/13.6	90.3/12.3	90.9/10.0	71.1/6.23	86.6/2.24	88.7/2.10
`MONA` [14]	91.3/**5.31**	92.2/9.46	92.3/8.16	72.8/10.7	87.0/2.81	89.5/2.40
● `ARCO-SAG` (ours)	91.5/6.82	**92.5**/6.95	92.6/7.54	73.2/6.47	86.9/2.73	89.1/2.30
○ `ARCO-SG` (ours)	**91.6**/6.60	**92.5/6.31**	**92.8**/6.64	**75.0/4.06**	**87.8/1.66**	**89.9/1.47**

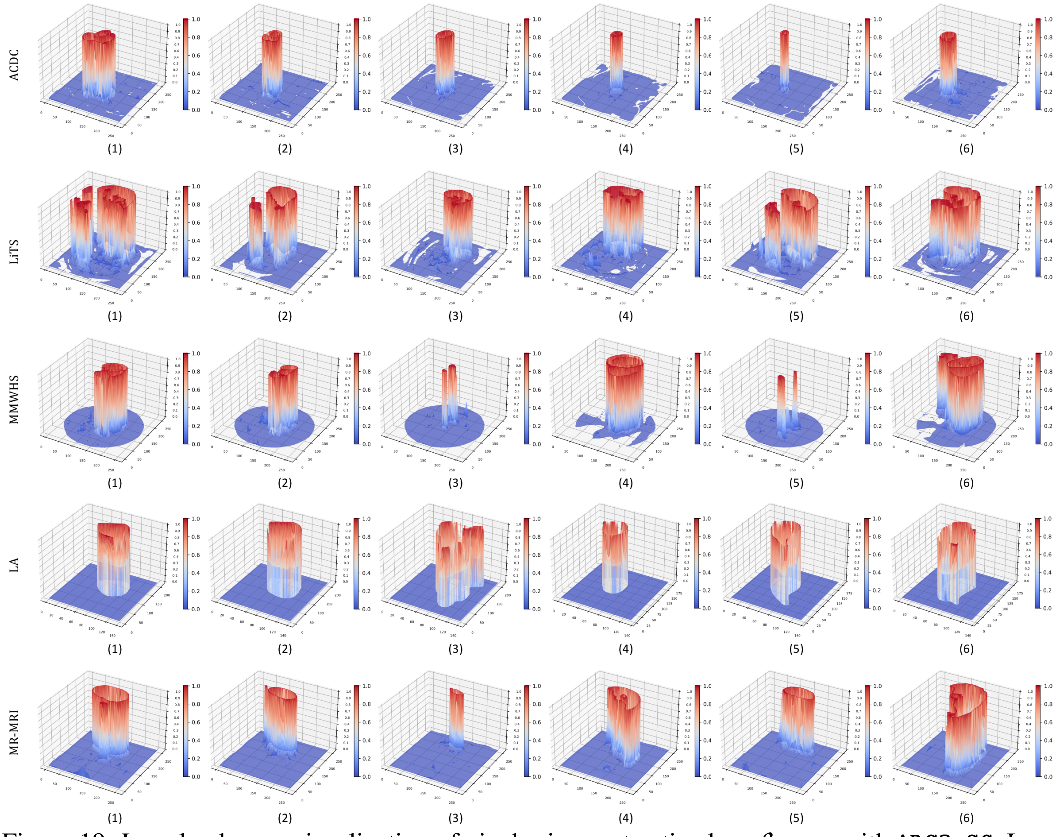


Figure 19: Loss landscape visualization of pixel-wise contrastive loss $\mathcal{L}_{\text{contrast}}$ with `ARCO-SG`. Loss plots are generated with same original images randomly chosen from ACDC [76], LiTS [77], MMWHS [78], LA [79], and MP-MRI, respectively. z-axis denotes the loss value at each pixel.

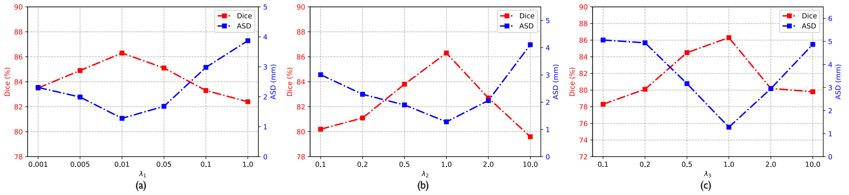


Figure 20: Effects of hyperparameters $\lambda_1, \lambda_2, \lambda_3$. We report Dice and ASD of ARCO-SG on ACDC with 1% labeled ratio.