Global-local View Training

Resize & crop

Input image

Seed mask (CLIP generated)

CLIP 🔒

Segmentation network

CARB w/ local

CARB w/ global

Consistency-Aware Region Balancing (CARB)

k : Iteration
n : Queue size

Pseudo-mask (P)

$\mathbf{P} \cap \mathbf{S}$

Consistent region ($M^{\mathbf{c}}$)

$\mathcal{L}_{\mathbf{c}}^{k}$

Loss of $M^{\mathbf{c}}$

enque Queue (Q^c) deque

| $\mathcal{L}_{\mathbf{c}}^{k}$ | $\mathcal{L}_{\mathbf{c}}^{k-1}$ | $\mathcal{L}_{\mathbf{c}}^{k-2}$ | ... | $\mathcal{L}_{\mathbf{c}}^{k-n}$ |

$\boldsymbol{Avg(Q^c)}$ $\boldsymbol{Avg(Q^i)}$

Prediction of Seg. (S)

$\mathbf{P} - \mathbf{S}$

Inconsistent region ($M^{\mathbf{i}}$)

$\mathcal{L}_{\mathbf{i}}^{k}$

Loss of $M^{\mathbf{i}}$

enque Queue (Q^i) deque

| $\mathcal{L}_{\mathbf{i}}^{k}$ | $\mathcal{L}_{\mathbf{i}}^{k-1}$ | $\mathcal{L}_{\mathbf{i}}^{k-2}$ | ... | $\mathcal{L}_{\mathbf{i}}^{k-n}$ |

Segmentation output (f)

$$\mathcal{L} = \mathcal{L}_{\mathbf{c}}^{k} + \boldsymbol{\omega} \cdot \mathcal{L}_{\mathbf{i}}^{k}$$

Training loss

$$\boldsymbol{\omega} = \frac{Avg\,(Q^{\mathbf{c}})}{Avg\,(Q^{\mathbf{i}})}$$

Adaptive weight for $\mathcal{L}_{\mathbf{i}}^{k}$