Exploiting a Joint Embedding Space for Generalized Zero-Shot Semantic Segmentation

Donghyeon Baek*, Youngmin Oh*, and Bumsub Ham
Yonsei University
(*: Equal contribution)

2021 ICCV OCTOBER 11-17 VIRTUAL

Summary

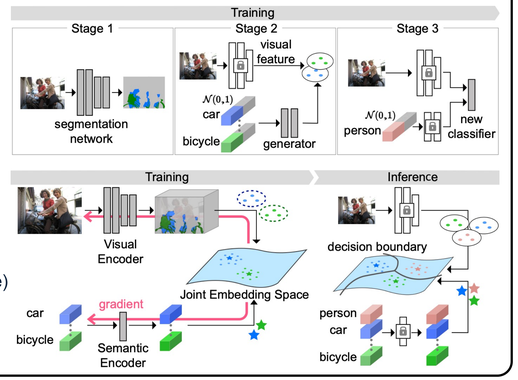
Problem statement

- Most generalized zero-shot semantic segmentation (GZS3) methods adopt a *generative* approach that synthesizes visual features of unseen classes from corresponding semantic ones (e.g., word2vec) to train novel classifiers for both seen and unseen classes
- Thus, they have two limitations: (1) the visual features of unseen classes are biased towards those of seen classes; (2) the classifier should be re-trained whenever novel unseen classes appear
- Goal: address these limitations in a unified framework

Contributions

- Introduce a simple yet effective *discriminative* approach for the task of GZS3, dubbed ***JoEm***
- Propose boundary-aware regression (BAR) and semantic consistency (SC) losses to learn a joint embedding space
- Introduce an effective inference technique, dubbed Apollonius calibration (AC), that modulates the decision boundary of the nearest-neighbor (NN) classifier adaptively

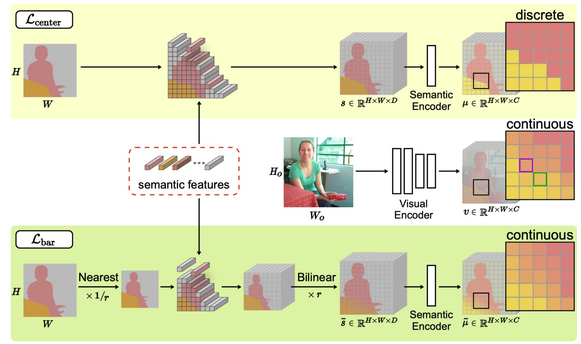
Comparison *JoEm* with SOTAs

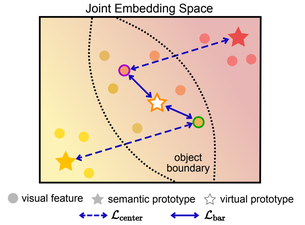


Training

$$\mathcal{L} = \mathcal{L}_{\text{ce}} + \mathcal{L}_{\text{bar}} + \lambda \mathcal{L}_{\text{sc}}$$

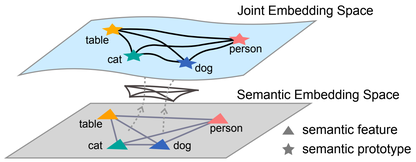
BAR and SC losses are complementary with each other, alleviating the seen bias problem





$$\mathcal{L}_{\text{center}} = \frac{1}{\sum_{c\in\mathcal{S}} |\mathcal{R}_c|} \sum_{c\in\mathcal{S}} \sum_{\mathbf{p}\in\mathcal{R}_c} d\left(v(\mathbf{p}), \mu(\mathbf{p})\right)$$

$$\mathcal{L}_{\text{bar}} = \frac{1}{\sum_{c\in\mathcal{S}} |\mathcal{R}_c|} \sum_{c\in\mathcal{S}} \sum_{\mathbf{p}\in\mathcal{R}_c} d\left(v(\mathbf{p}), \tilde{\mu}(\mathbf{p})\right)$$

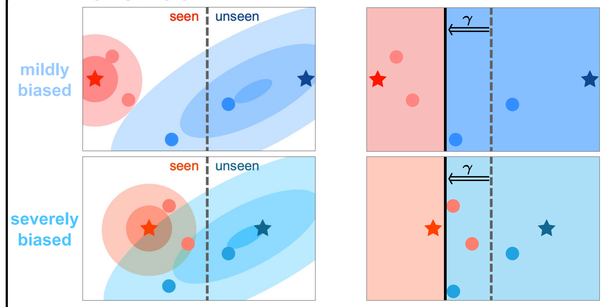


$$\hat{r}_{ij} = \frac{e^{-\tau_\mu d(\mu_i, \mu_j)}}{\sum_{j\in\mathcal{S}} e^{-\tau_\mu d(\mu_i, \mu_j)}}$$

$$r_{ij} = \frac{e^{-\tau_s d(s_i, s_j)}}{\sum_{j\in\mathcal{S}} e^{-\tau_s d(s_i, s_j)}}$$

$$\mathcal{L}_{\text{sc}} = -\sum_{i\in\mathcal{S}} \sum_{j\in\mathcal{S}} r_{ij} \log \frac{\hat{r}_{ij}}{r_{ij}}$$

Inference



$$\hat{y}_{\text{nn}}(\mathbf{p}) = \underset{c\in\mathcal{S}\cup\mathcal{U}}{\operatorname{argmin}}\, d(v(\mathbf{p}), \mu_c)$$

$$\hat{y}_{\text{cs}}(\mathbf{p}) = \underset{c\in\mathcal{S}\cup\mathcal{U}}{\operatorname{argmin}}\, d(v(\mathbf{p}), \mu_c) - \gamma \mathbb{1}[c \in \mathcal{U}]$$

$$\hat{y}_{\text{ac}}(\mathbf{p}) = \begin{cases} c_{12}(\mathbf{p}) \ , & c_{1\text{st}} \in \mathcal{S} \text{ and } c_{2\text{nd}} \in \mathcal{U} \\ c_{1\text{st}} \ , & \text{otherwise} \end{cases}$$

$$c_{12}(\mathbf{p}) = c_{1\text{st}} \mathbb{1}\left[\frac{d_1(\mathbf{p})}{d_2(\mathbf{p})} \le \sigma\right] + c_{2\text{nd}} \mathbb{1}\left[\frac{d_1(\mathbf{p})}{d_2(\mathbf{p})} > \sigma\right]$$

AC makes the NN classifier less susceptible to the seen bias problem, even without re-training the classifier

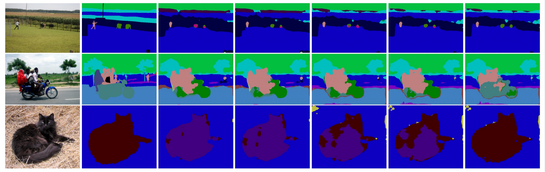
Results on the experimental settings provided by ZS3Net

Performance on standard GZS3 benchmarks (PASCAL VOC and PASCAL Context)

Datasets	Methods	unseen-2			unseen-4			unseen-6			unseen-8			unseen-10		
		mIoU$_\mathcal{S}$	mIoU$_\mathcal{U}$	hIoU	mIoU$_\mathcal{S}$	mIoU$_\mathcal{U}$	hIoU	mIoU$_\mathcal{S}$	mIoU$_\mathcal{U}$	hIoU	mIoU$_\mathcal{S}$	mIoU$_\mathcal{U}$	hIoU	mIoU$_\mathcal{S}$	mIoU$_\mathcal{U}$	hIoU
VOC	DeViSE [12]	68.1	3.2	6.1	64.3	2.9	5.5	39.8	2.7	5.1	35.7	2.0	3.8	31.7	1.9	3.6
	SPNet [51]	71.8	34.7	46.8	67.3	21.8	32.9	64.5	20.1	30.6	61.2	19.9	30.0	59.0	18.1	27.7
	ZS3Net [3]	72.0	35.4	47.5	66.4	23.2	34.4	47.3	24.2	32.0	29.2	22.9	25.7	33.9	18.1	23.6
	CSRL [32]	73.4	45.7	**56.3**	69.8	31.7	43.6	66.2	29.4	40.7	62.4	26.9	37.6	59.2	21.0	31.0
	Ours	68.9 (1.0)	43.2 (0.9)	53.1 (0.4)	67.0 (1.2)	33.4 (0.4)	**44.6** (0.3)	63.2 (0.4)	30.5 (0.3)	**41.1** (0.2)	58.5 (0.9)	29.0 (0.8)	**38.8** (0.6)	63.5 (0.4)	22.5 (0.4)	**33.2** (0.4)
Context	DeViSE [12]	35.8	2.7	5.0	33.4	2.5	4.7	31.9	2.1	3.9	22.0	1.7	3.2	17.5	1.3	2.4
	SPNet [51]	38.2	16.7	23.2	36.3	18.1	24.2	31.9	19.9	24.5	28.6	14.3	19.1	27.1	9.8	14.4
	ZS3Net [3]	41.6	21.6	28.4	37.2	24.9	29.8	32.1	20.7	25.2	20.9	16.0	18.1	20.8	12.7	15.8
	CSRL [32]	41.9	27.8	33.4	39.8	23.9	29.9	35.5	22.0	27.2	31.7	18.1	23.0	29.4	14.6	19.5
	Ours	38.2 (1.2)	32.9 (1.4)	**35.3** (0.9)	36.9 (0.8)	30.7 (1.5)	**33.5** (0.7)	36.2 (0.6)	23.2 (0.4)	**28.3** (0.4)	32.4 (0.9)	20.2 (0.4)	**24.9** (0.3)	33.0 (0.6)	14.9 (0.7)	**20.5** (0.6)

Ablation studies on the unseen-4 split of PASCAL Context

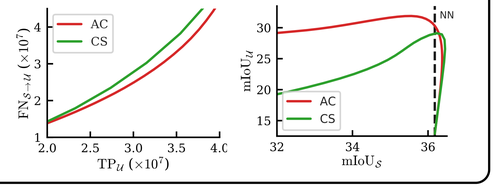
$\mathcal{L}_{\text{ce}}$	$\mathcal{L}_{\text{center}}$	$\mathcal{L}_{\text{bar}}$	$\mathcal{L}_{\text{sc}}$	CS	AC	mIoU$_\mathcal{S}$	mIoU$_\mathcal{U}$	hIoU
✓	✓					37.7	10.0	15.8
✓		✓				37.9	10.7	16.7
✓	✓		✓			36.1	11.8	17.8
✓		✓	✓			36.2	12.9	19.0
✓		✓	✓	✓		36.2	29.1	32.3
✓		✓	✓		✓	35.7	31.8	**33.7**



Analysis of embedding spaces on the unseen-4 split of PASCAL VOC

Methods	mIoU$_\mathcal{S}$	mIoU$_\mathcal{U}$	hIoU
S→V: $\mathcal{L}_{\text{center}}$	61.7	20.9	31.2
S→V: $\mathcal{L}_{\text{bar}}$ + $\mathcal{L}_{\text{sc}}$	65.7	30.3	41.5
ZS3Net [3]	66.4	23.2	34.4
ZS3Net†	68.8	28.8	40.6
ZS3Net‡	68.5	31.8	**43.4**

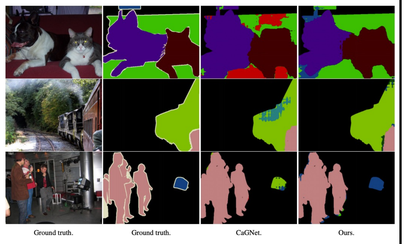
Comparison AC with calibrated stacking



Results on the experimental setting provided by SPNet

Note that this setting uses pixel-wise annotations for the background class during inference.

Methods	mIoU$_\mathcal{S}$	mIoU$_\mathcal{U}$	hIoU
SPNet [12]	78.0	15.6	26.1
ZS3Net [2]	77.3	17.7	28.7
CaGNet [7]	78.4	26.6	39.7
Ours w/o AC	78.9	30.6	44.1
Ours	77.7	32.5	**45.9**



Project site