Consistent Structural Relation Learning for Generalized Zero-Shot Segmentation

Peike Li, Yunchao Wei, and Yi Yang
ReLER Lab, AAII,University of Technology Sydney

Background

Q1: Why do we need semantic segmentation?

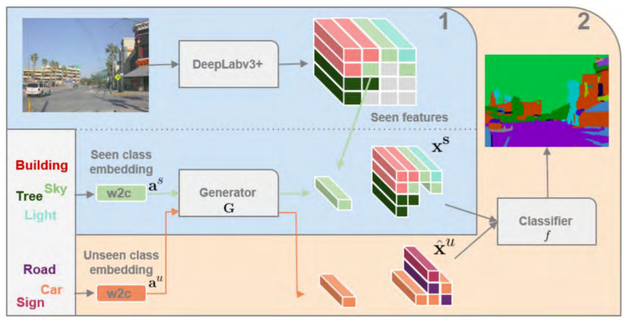
- Semantic segmentation aims to assign a semantic label to each pixel in the given image;
- Semantic segmentation is broadly applied in:
  - Video analysis;
  - Autonomous driving;
  - Virtual reality;
- To relieve the human effort in annotating accurate pixel-wise masks, there is an increasing trend of given less supervised signals, e.g.,
  - Weakly-supervised segmentation;
  - Few-shot segmentation;
  - Zero-shot segmentation;

Q2: What is the task of generalized zero-shot semantic segmentation (GZS3)?

- **"Generalized"** -> training with samples of seen categories, testing on both seen & unseen categories;
- **"Zero-shot"** -> no image of unseen categories given; we only have the word embeddings;
- **"Semantic Segmentation" ->** predict the category of each pixel in the image;

Q3: What is the common approach to achieve zero-shot segmentation?

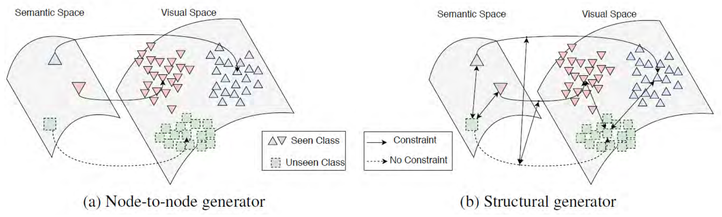
- **Step1:** train a segmentation model on images with only seen categories;
- **Step2**: train a generator which generates visual features from word embeddings on seen categories;
- **Step3:** generate visual features of unseen categories given the word embeddings;
- **Step4:** finetune the segmentation model on visual features from real seen and generated unseen features;



Method

Q4: What is the drawbacks of the common approach?

- Do not consider the relations between categories;
- No constraint on the generation of unseen visual features;
- Poor generalization ability;



(a) Node-to-node generator (b) Structural generator

Q5: What is your motivation?

- Different categories are roughly with similar relations in either semantic word embedding space or visual feature space;

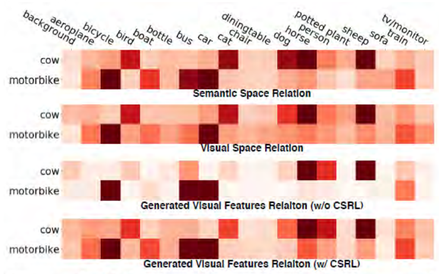
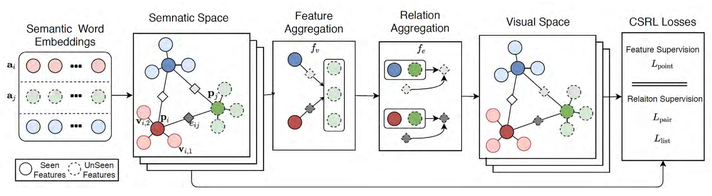


Figure 4: Relations between unseen (cow and motorbike) and seen categories.

Q6: What is your approach?

- Our CSRL incorporates the feature generating and relation learning into a unified architecture;
- Given the semantic word embedding, our CSRL generates visual features by alternately feature and relation aggregation;
- Our CSRL is trained under supervision from point-wise consistency on seen classes, pair-wise and list-wise consistency across seen and unseen classes.



Experiment

Q7: What is the performance of your method?

- We report the performance of generalized zero-shot semantic segmentation on Pascal-VOC dataset;
- Our method CSRL provides significant gains particularly on the unseen classes and achieve better results in terms of hIoU;

Table 1: Generalized zero-shot semantic segmentation performance on Pascal-VOC dataset.

Settings	Methods	Seen mIoU	Unseen mIoU	Overall mIoU	Overall hIoU
unseen-2	SegDevis	68.1%	3.2%	44.1%	6.1%
	SPNet	71.8%	34.7%	68.2%	46.8%
	ZS3Net	72.0%	35.4%	68.5%	47.5%
	CSRL	**73.4%**	**45.7%**	**70.7%**	**56.3%**
unseen-4	SegDevis	64.3%	2.9%	38.9%	5.5%
	SPNet	67.3%	21.8%	58.6%	32.9%
	ZS3Net	66.4%	23.2%	58.2%	34.4%
	CSRL	**69.8%**	**31.7%**	**62.5%**	**43.6%**
unseen-6	SegDevis	39.8%	2.7%	33.4%	5.1%
	SPNet	64.5%	20.1%	51.8%	30.6%
	ZS3Net	47.3%	24.2%	40.7%	32.0%
	CSRL	**66.2%**	**29.4%**	**55.6%**	**40.7%**
unseen-8	SegDevis	35.7%	2.0%	24.3%	3.8%
	SPNet	61.2%	19.9%	45.5%	30.0%
	ZS3Net	29.2%	22.9%	26.8%	25.7%
	CSRL	**62.4%**	**26.9%**	**48.8%**	**37.6%**
unseen-10	SegDevis	31.7%	1.9%	16.9%	3.6%
	SPNet	59.0%	18.1%	39.5%	27.7%
	ZS3Net	33.9%	18.1%	26.3%	23.6%
	CSRL	**59.2%**	**21.0%**	**50.0%**	**31.0%**

Visualization

Q8: Can you show me some visualization results?

- We compare the results with former state-of-the-art ZS3Net on Pascal-VOC (`cow` & `motorbike` are unseen categories).