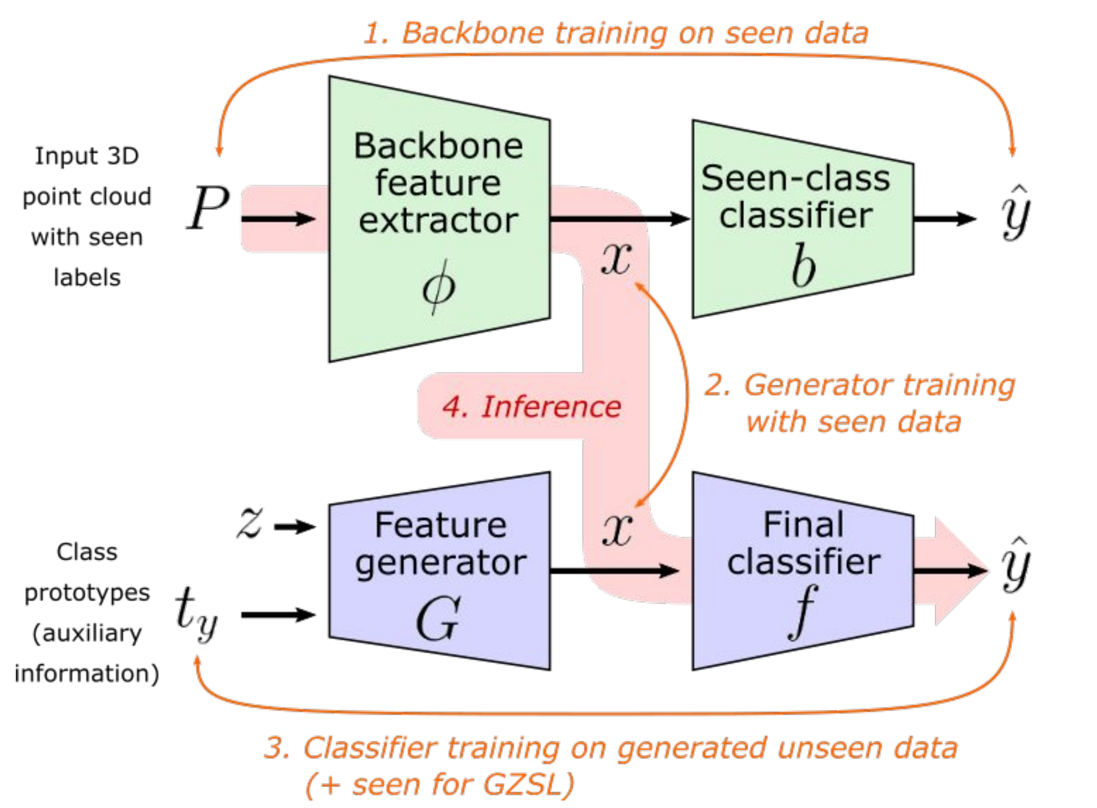
Generative Zero-Shot Learning for Semantic Segmentation of 3D Point Clouds

Björn Michele[1], Alexandre Boulch[1], Gilles Puy[1], Maxime Bucher[1], Renaud Marlet[1,2]
[1]valeo.ai, France. [2]LIGM, Ecole des Ponts, Univ Gustave Eiffel, CNRS, France.

Zero-shot learning for 3D point clouds (PCs)

- Zero-shot learning (ZSL): detect, at inference time, objects of classes which have not been seen during training.

- We use a generative approach based on [1,2] and adapt it to 3D PCs:
  - A *backbone* $\phi(\cdot)$ extracting a meaningful representation x of 3D point clouds.
  - A *feature generator* $G(\cdot)$ learning to generate representations x based on class prototypes. The generated representations are used to train a classifier for unseen classes.



- Existing ZSL methods for 3D point clouds did not make use of generative approaches and do not tackle semantic segmentation.

- **Contributions:**
  - A generative framework handling both ZSL and Generalized ZSL (GZSL) for 3D point clouds, for semantic segmentation and classification.
  - 3 benchmarks for 3D semantic segmentation based on SemanticKITTI [8] (outdoor), S3DIS [9] and ScanNet [10] (indoor).
  - 2 additional baselines for 3D GZSL segmentation.

References

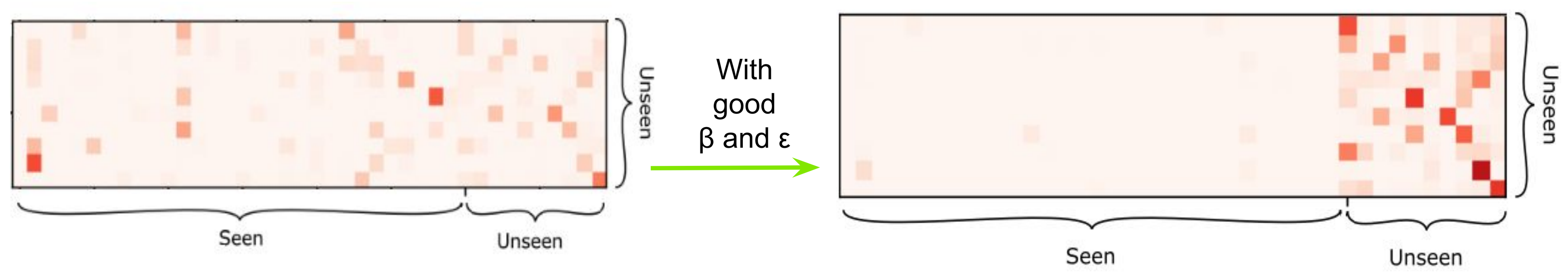
[1] Bucher et al. Generating visual representations for zero-shot classification. ICCV, 2017.
[2] Bucher et al. Zero-shot semantic segmentation. NeurIPS, 2019.
[3] Cheraghian et al. Zero-shot learning of 3D point cloud objects. MVA, 2019.
[4] Cheraghian et al. Mitigating the hubness problem for zero-shot learning of 3D objects. 2019.
[5] Cheraghian et al. Transductive zero-shot learning for 3D point cloud classification. WACV, 2020.
[6] Chao et al. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. ECCV, 2016.
[7] Frome et al. DeViSE: A deep visual-semantic embedding model. NIPS, 2013.
[8] Behley et al. SemanticKITTI: A dataset for Semantic Scene Understanding of LiDAR Sequences. ICCV, 2019.
[9] Armeni et al. 3D semantic parsing of large-scale indoor spaces. CVPR, 2016.
[10] Dai et al. ScanNet: Richly-annotated 3d reconstructions of indoor scenes. CVPR, 2017.

Reducing bias towards seen classes

- In GZSL a strong bias toward seen classes can be observed [6].

- Bias reduction techniques:
  - *Class-dependent weighting*: Loss for unseen classes is weighted with a factor β >1 in classifier training.
  - *Calibrated Stacking* [6]: At test time a small value ε is subtracted from the seen-class score (after softmax).
  - β and ε are estimated by cross-validation.

- GZSL is the naturally setting for semantic segmentation.



Results

- **Classification:**
  - ZSL and GZSL on ModelNet40 (10 unseen, 30 seen classes).
  - Different auxiliary information: W2V and GloVe

Method	Generative	Full supervision Acc.	ZSL W2V Acc.	ZSL GloVe Acc.	GZSL Bias reduct.	GZSL W2V Acc. $\mathcal{S}$	GZSL W2V Acc. $\mathcal{U}$	GZSL W2V HM	GZSL GloVe Acc. $\mathcal{S}$	GZSL GloVe Acc. $\mathcal{U}$	GZSL GloVe HM
PointNet		89.2									
f-CLSWGAN* [5]	✓		20.7	-		76.3	3.7	7.0	-	-	-
CADA-VAE* [5]	✓		23.0	-		84.7	1.3	2.6	-	-	-
ZSLPC [3]			28.0	20.9		40.1	22.5	28.8	49.2	18.2	26.6
MHPC [4]			**33.9**	28.7	✓	**53.8**	26.2	35.2	**53.8**	25.7	**34.8**
3DGenZ (ours)	✓		28.6	**29.3**	✓	48.8	**29.3**	**36.6**	44.7	**28.4**	34.7

*: adaptation of 2D methods to 3D point clouds, implemented in [5]

- **Semantic segmentation**
  - GZSL on S3DIS, ScanNet and SemanticKITTI.
  - 2 Baseline methods + additional bias reduction.

Upper bounds

Baselines

	Training set Backbone	Training set Classifier	S3DIS mIoU $\mathcal{S}$	S3DIS mIoU $\mathcal{U}$	S3DIS HmIoU	ScanNet mIoU $\mathcal{S}$	ScanNet mIoU $\mathcal{U}$	ScanNet HmIoU	SemanticKITTI mIoU $\mathcal{S}$	SemanticKITTI mIoU $\mathcal{U}$	SemanticKITTI HmIoU
Supervised methods with different levels of supervision											
Full supervision	$\mathcal{S}\cup\mathcal{U}$	$\mathcal{S}\cup\mathcal{U}$	74.0	50.0	59.6	43.3	51.9	47.2	59.4	50.3	54.5
ZSL backbone	$\mathcal{S}$	$\mathcal{S}\cup\mathcal{U}$	60.9	21.5	31.8	41.5	39.2	40.3	52.9	13.2	21.2
ZSL-trivial	$\mathcal{S}$	$\mathcal{S}$	70.2	0.0	0.0	39.2	0.0	0.0	55.8	0.0	0.0
Generalized zero-shot-learning methods											
ZSLPC-Seg* [3]†	$\mathcal{S}$	$\mathcal{U}$	65.5	0.0	0.0	28.2	0.0	0.0	49.1	0.0	0.0
DeViSe-3DSeg* [7]†	$\mathcal{S}$	$\mathcal{U}$	70.2	0.0	0.0	20.0	0.0	0.0	49.7	0.0	0.0
ZSLPC-Seg [3]†	$\mathcal{S}$	$\mathcal{U}$	5.2	1.3	2.1	16.4	4.2	6.7	26.4	10.2	14.7
DeViSe-3DSeg [7]†	$\mathcal{S}$	$\mathcal{U}$	3.6	1.4	2.0	12.8	3.0	4.8	42.9	4.2	7.5
3DGenZ (ours)	$\mathcal{S}$	$\mathcal{S}\cup\mathcal{U}$	**53.1**	**7.3**	**12.9**	**32.8**	**7.7**	**12.5**	**41.4**	**10.8**	**17.1**

†Our adaption of the method. *Direct, unrepaired (failing) adaptation.

Image based prototypes

- Alternative auxiliary information based on image representations.

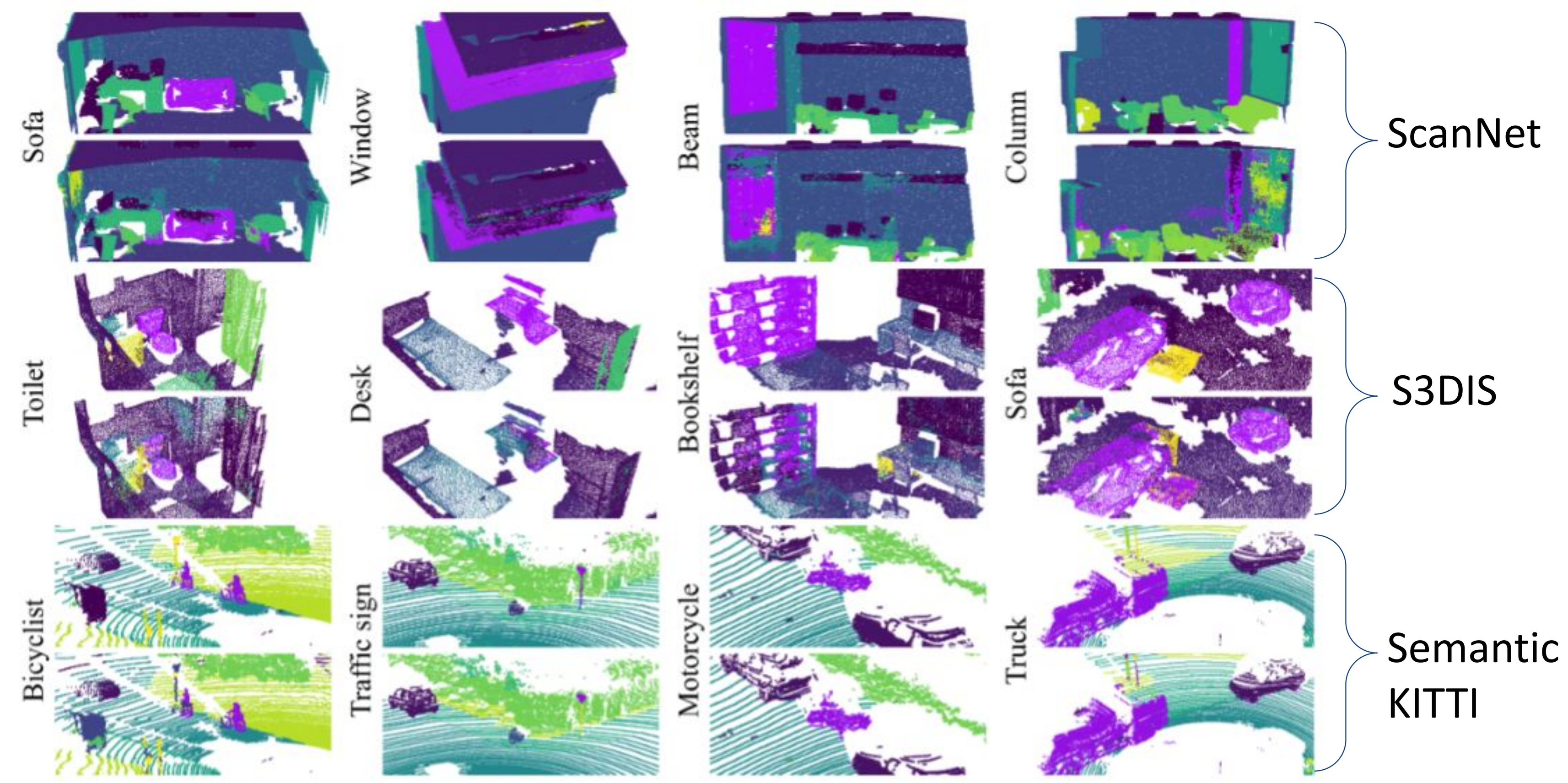
- Images capturing the appearance of objects, perhaps a better link to the object geometry.

- Description of an object class with a small set of images → generate a visual representation by averaging features extracted using a pre-trained CNN (ResNet).

Representation	*Classif.* ModelNet40 ZSL	*Classif.* ModelNet40 GZSL	*Segmentation* HmIoU ScanNet	*Segmentation* HmIoU KITTI
W2V+GloVe (self-sup.)	36.8	**41.3**	12.5	**17.1**
ResNet-18 (sup.)	**43.6**	40.0	13.9	3.6
ResNet-50 (self-sup.)	37.0	36.5	**15.5**	5.3

Only image based

Visualizations



Top: Ground Truth, Bottom: Ours
■ Unseen class

Summary

- Zero shot learning for 3D point clouds with generative approach:
  - reaches state of the art in classification with additional bias reduction.
  - outperforms natural baselines in GZSL for semantic segmentation.

- We proposed to use image-based auxiliary information.