MSc Thesis Defense

Relation and Knowledge Aware Zero Shot Learning in 3D Object Recognition

Md Fokhrul Islam
Exam Roll: 120702
Reg. No.: 2016-914-692
Course Code: RME 5401

Supervisor:
Dr. Sejuti Rahman
Associate Professor
Dept. of RME, DU

Dept. of Robotics & Mechatronics Engineering
University of Dhaka

Agenda

- Introduction
  - Motivation
  - Problem Definition
- Related Work
- Novelty
  - Limitation of Existing Literature
  - Ours' Novelty
- Proposed Approach
  - Text-Image Alignment
  - Knowledge Graph Construction
  - Common Sense Embedding Learning
  - Contrastive GAN framework
  - Instance-level Contrastive Embedding
- Experiments & Results
  - Dataset
  - Overall Result (benchmark, backbone and prompt analysis)
  - Ablation (t-SNE plot, component and confusion matrix analysis)
- Conclusion

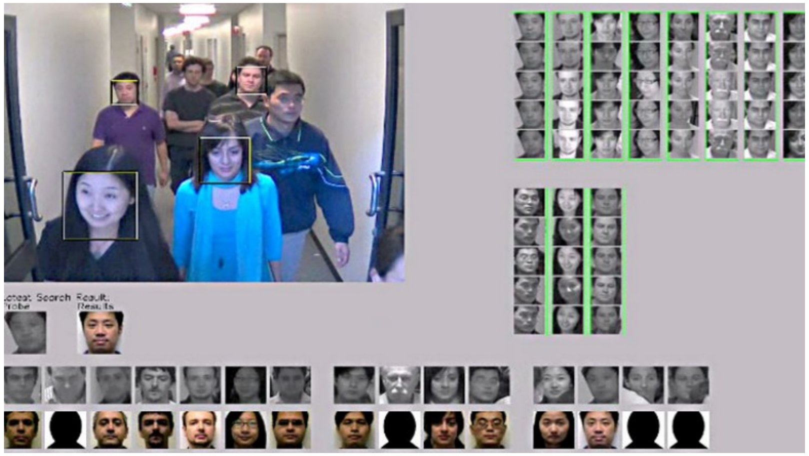
Motivation: 2D vs 3D

2D vision:

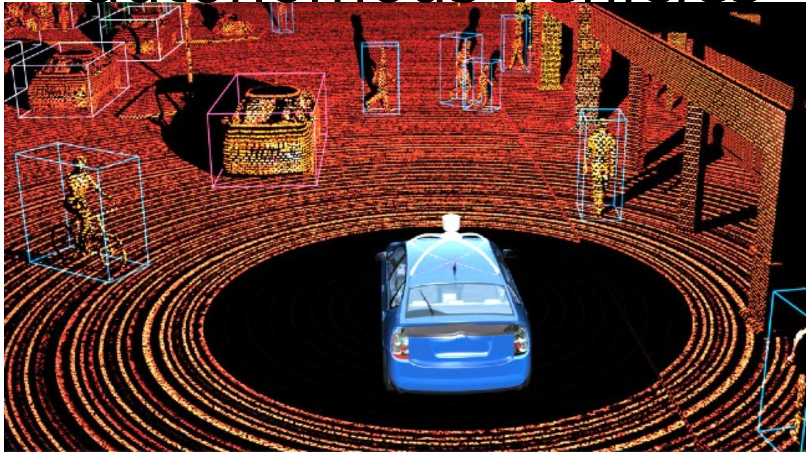
- Limitations: Lack of depth, viewpoint dependency
- harder for 2D vision to perform high level real-world tasks

3D vision:

- Advantages: Depth perception, robustness to pose variations
- Object recognition in robotics, augmented reality and autonomous vehicles



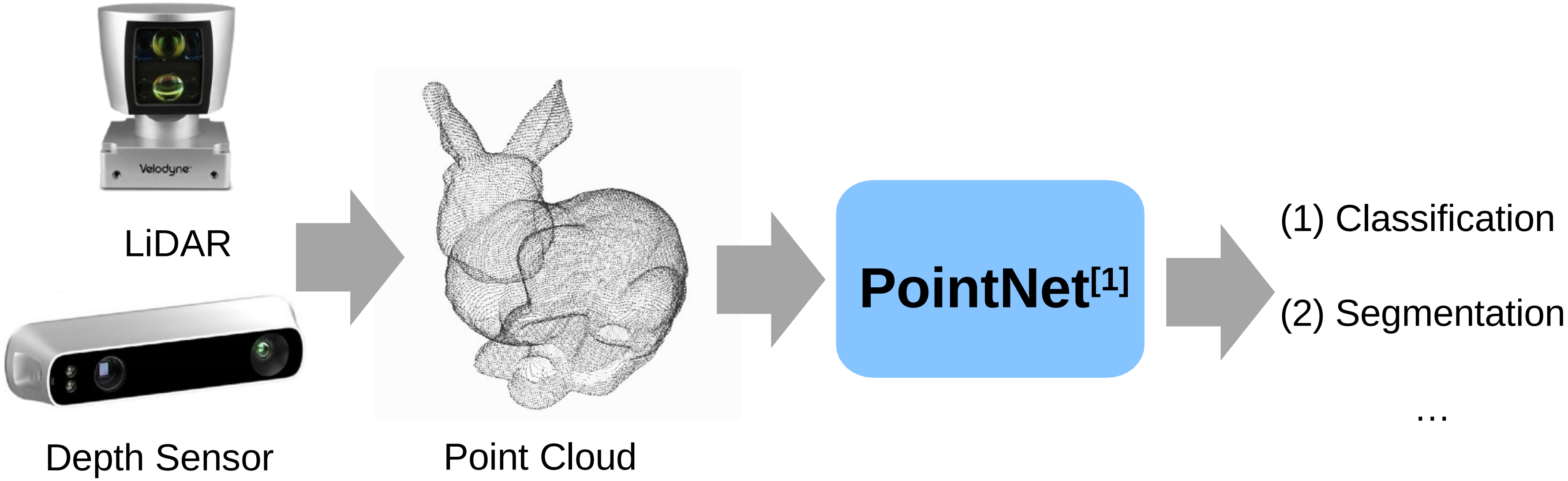
Facial recognition



Robot perception

3D Representation and Learning

- ✓ Point cloud is close to raw sensor data
- ✓ End-to-end learning for scattered, unordered point data
- ✓ Unified framework for various tasks

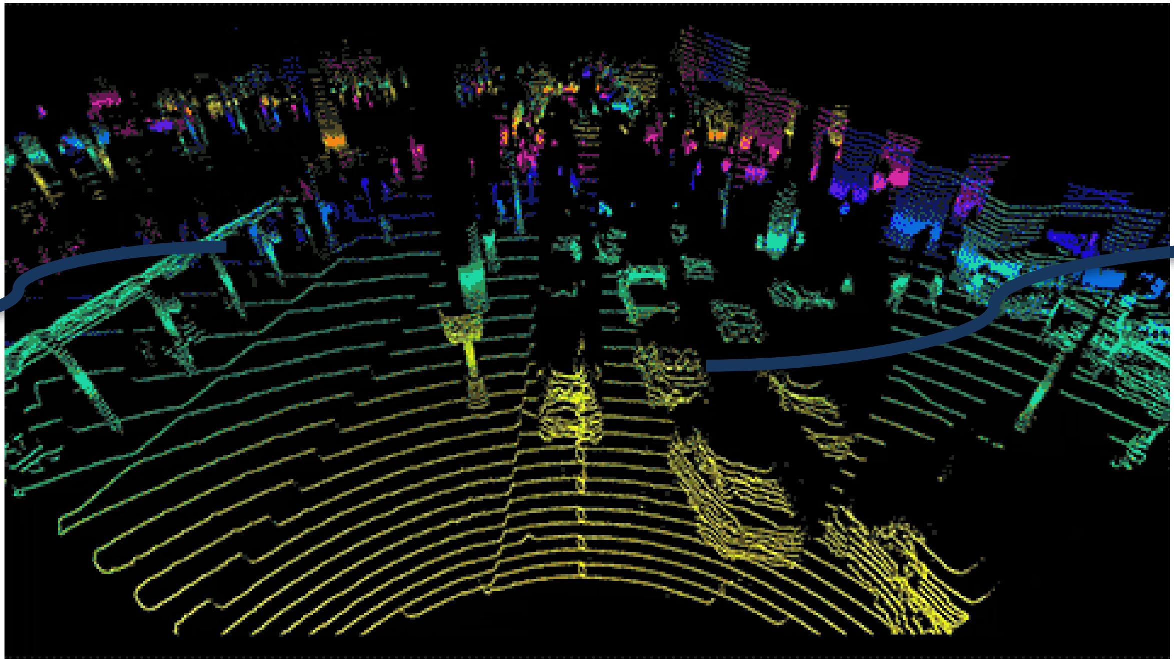


[1] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, “Pointnet: Deep learning on point sets for 3d classification and segmentation,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 652–660

Motivation: What Machine See in Real World?

Two type of scenario may occur in this situation:
(a) Already known through supervision by trained on large amount of data
(b) Never seen this type of objects

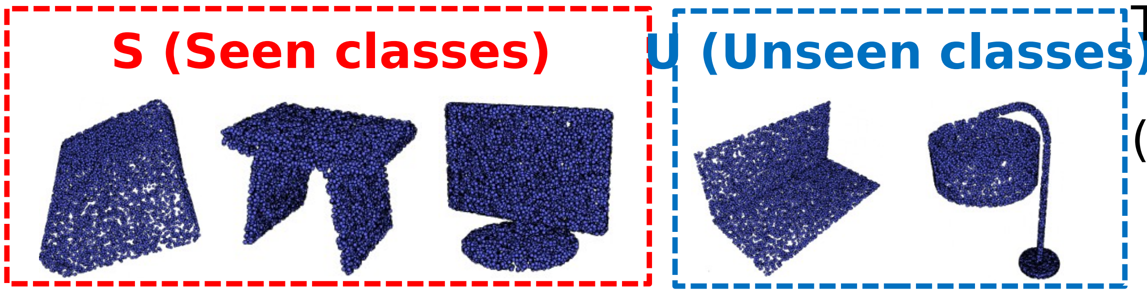
Appear first time



Already known

Courtesy: New York Times, What Self-Driving Cars See?

Problem Definition



Train on classes in S.

(a) For ZSL: Test on classes in U

(b) For GZSL: Test on classes in S and U

Need electricity	yes	no	yes	yes	yes
Placed on	table	floor	table	table	table
Made of	metal	wood	metal	metal	metal

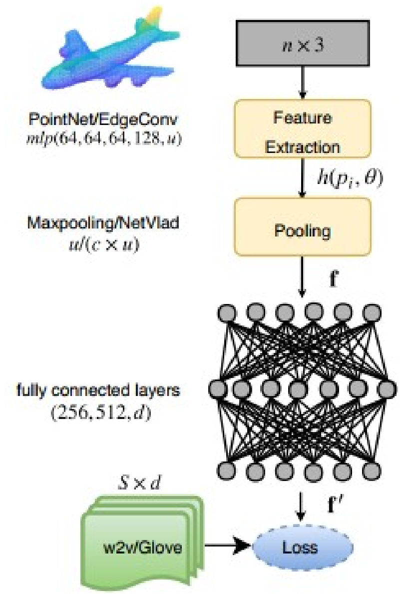
enerally, the learning is achieved using additional side-information called **sema embedding** or attributes vectors.

hese semantic embedding encode the intra class relationships between all the seen (S) and unseen (U) classes

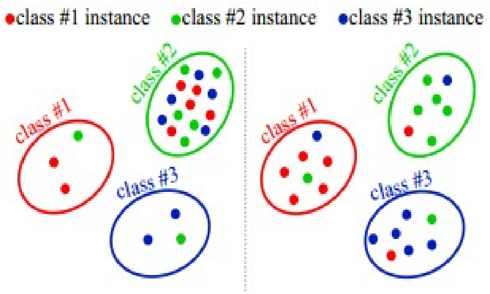
*Example images come from ModelNet40 dataset

Related Work

- Zero-shot Learning of 3D Point Cloud Objects[2]
  - First paper to introduced ZSL problem in 3D point cloud classification
  - Their architecture used semantic word vectors



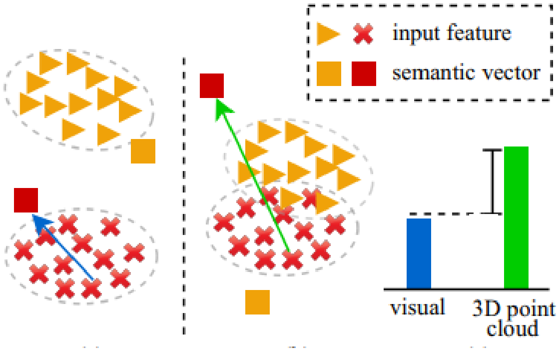
- Mitigating the Hubness Problem for Zero-Shot Learning of 3D Objects[3]
  - New loss proposed to mitigate the hubness problem of 3D-ZSL task(s)
  - More ZSL and GZSL benchmark result provided for 3D point cloud classification



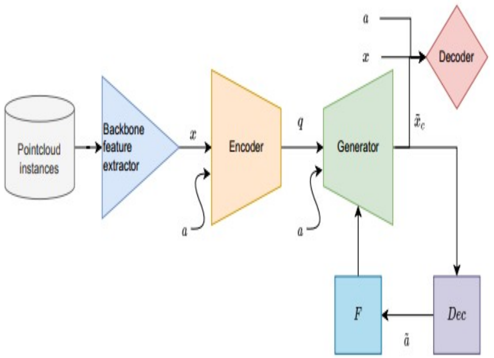
[2] A. Cheraghian, S. Rahman, and L. Petersson, "Zero-shot learning of 3d point cloud objects," in 2019 16th International Conference on Machine Vision Applications (MVA). IEEE, 2019, pp. 1–6. [3] A. Cheraghian, S. Rahman, D. Campbell, and L. Petersson, "Mitigating the hubness problem for zero-shot learning of 3d objects," arXiv preprint arXiv:1907.06371, 2019.

Related Work (2)

- Transductive Zero-Shot Learning for 3D Point Cloud Classification[4]
  - Reduced the bias and encouraged the projected semantic vectors to align with their true feature vector
  - Developed a novel triplet loss to minimize the average intra-class distance



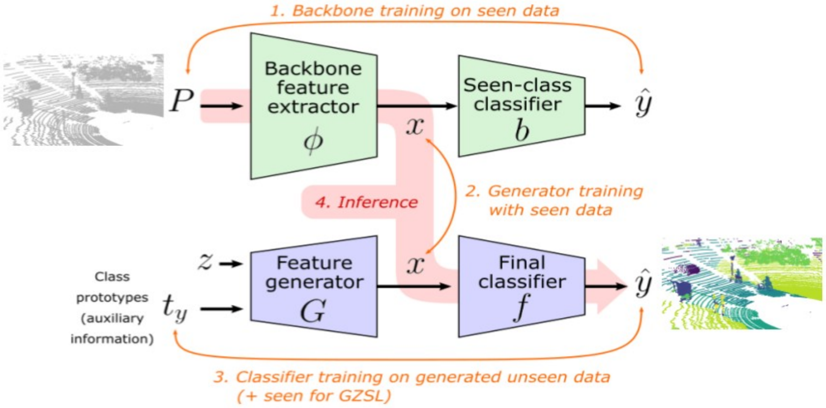
- Improving 3D object Recognition with Contextual Information and Meta Learning [5]
  - First method to apply GAN based approach on 3D ZSL
  - Used 2D image features as auxiliary info



[4] A. Cheraghian, S. Rahman, D. Campbell, and L. Petersson, “Transductive zero-shot learning for 3d point cloud classification,” in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 923–933. [5] Muhammad Tahmeed Abdullah. "Improving 3D object Recognition with Contextual Information and Meta Learning." Master's thesis in department of Robotics and Mechatronics Engineering,

Related Work (3)

- Generative Zero-Shot Learning for Semantic Segmentation of 3D Point Clouds[6]
  - Proposed a generative framework and provide 3 benchmarks.
  - Provided evidence that generation based method can perform better in GZSL
  - Failed to incorporate Image-text embedding and left it for future researcher



[6] Michele, Björn, et al. "Generative zero-shot learning for semantic segmentation of 3d point clouds." 2021 International Conference on 3D Vision (3DV). IEEE, 2021.

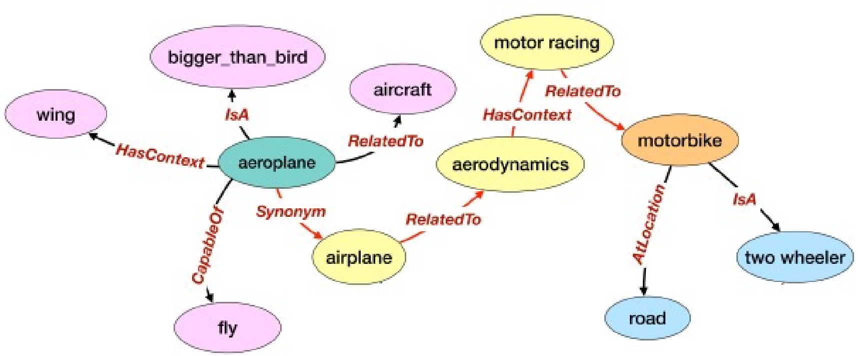
Limitations in Existing Literatures

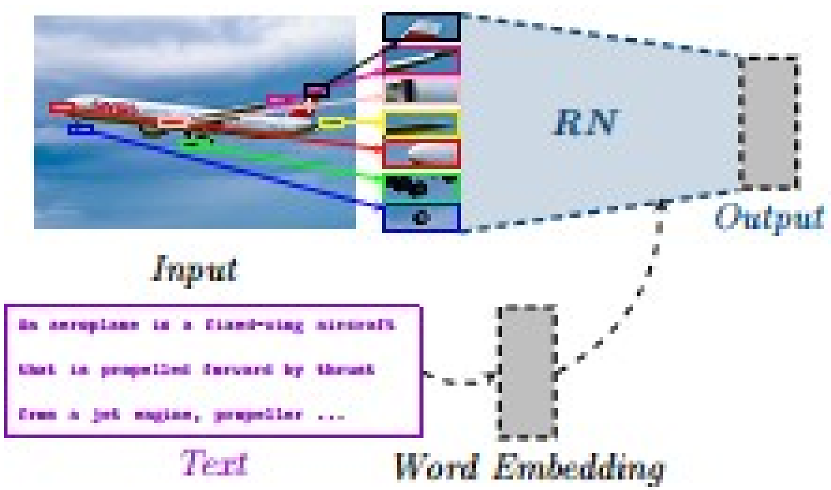
- Limitations in existing works:
  - Cannot connect different class in semantic space [No relation]
  - Using only 2D (one/some) image can not be viable because of image variations [need to learn class attribute]
  - Generative unseen features are more bias toward seen class [no distinct feature representation]

Novelty

- Knowledge aware:
  - Connect all concepts through knowledge graph.
  - They relate each other through learned embedding

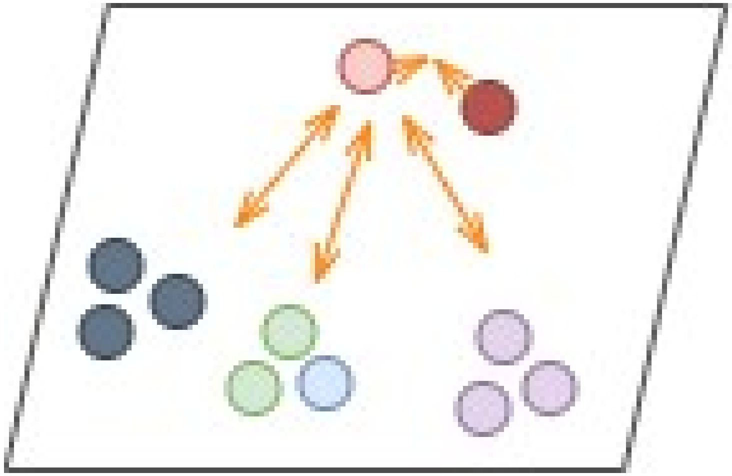
- Relational aware:
  - Extend the implicit information by incorporating 2D image feature with text
  - Both modalities make relation based on leaned feature



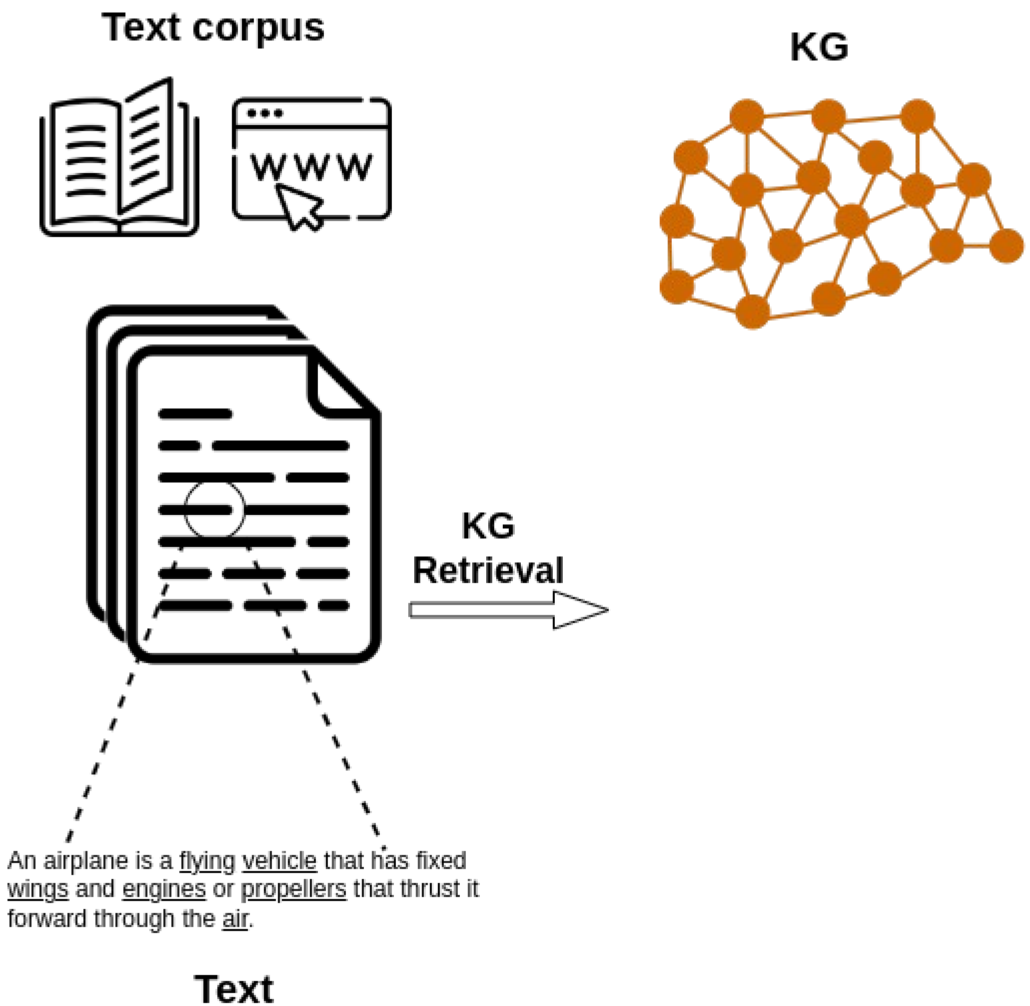


Novelty(2)

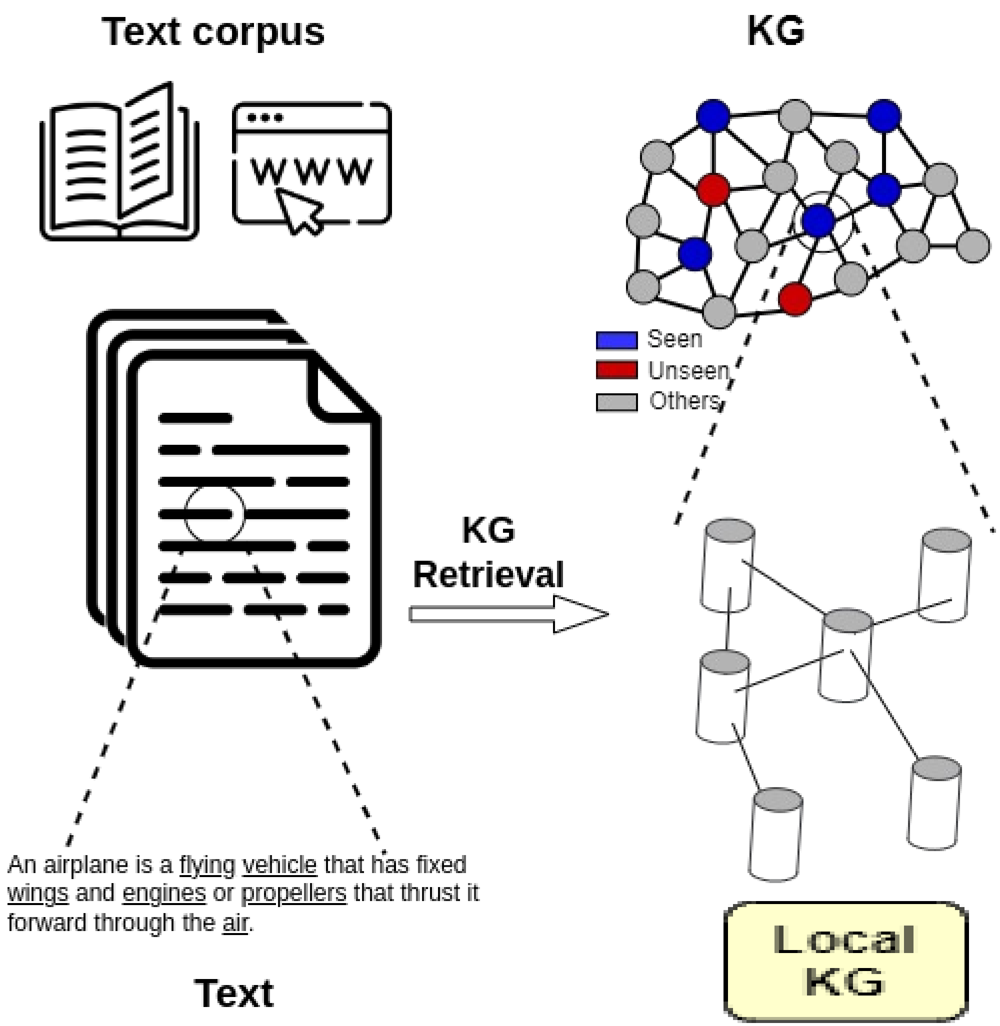
- More distinct synthetic feature leaning:
  - Learn a more accurate 3D visual feature representation using contrastive semi supervised leaning
  - This will also solve bias problem occur in seen-unseen classes



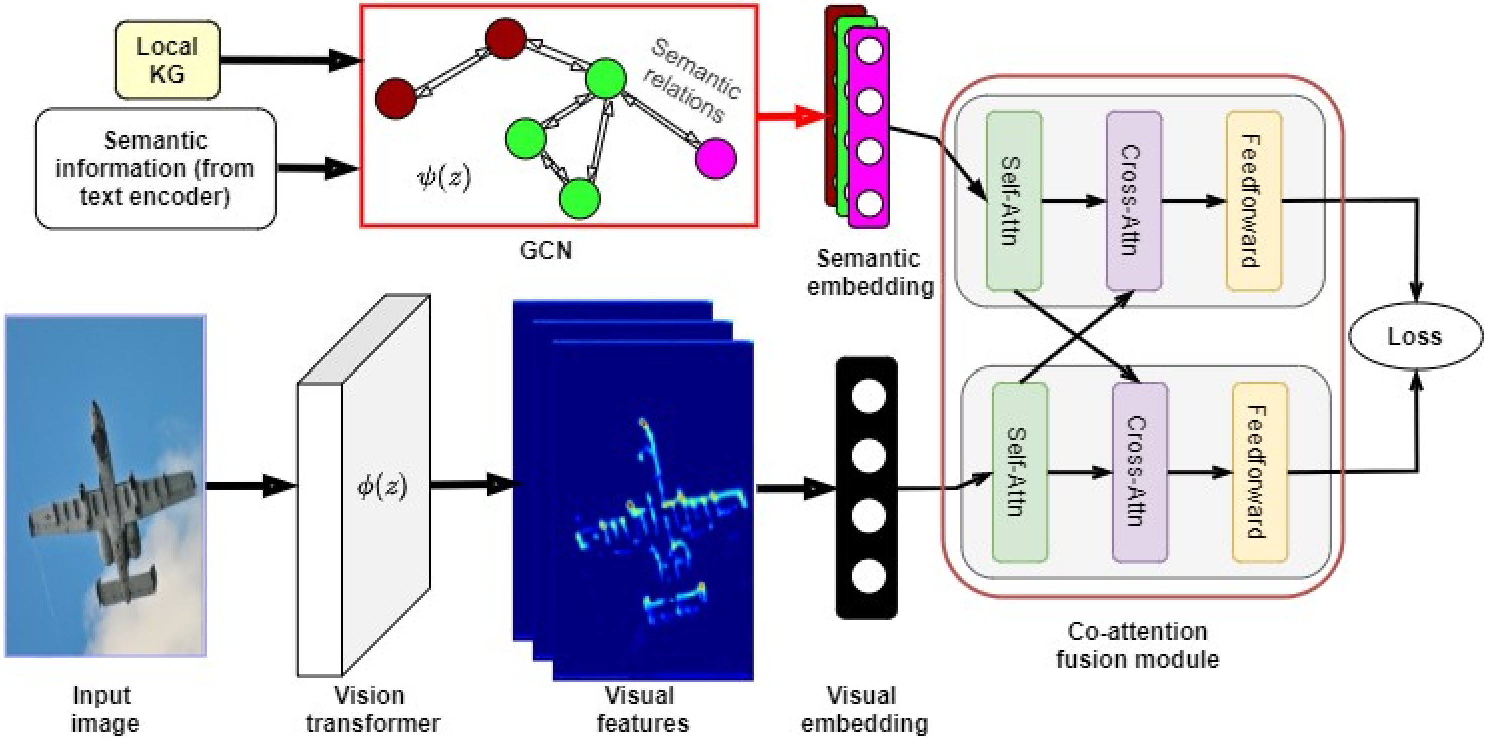
Knowledge Graph Construction



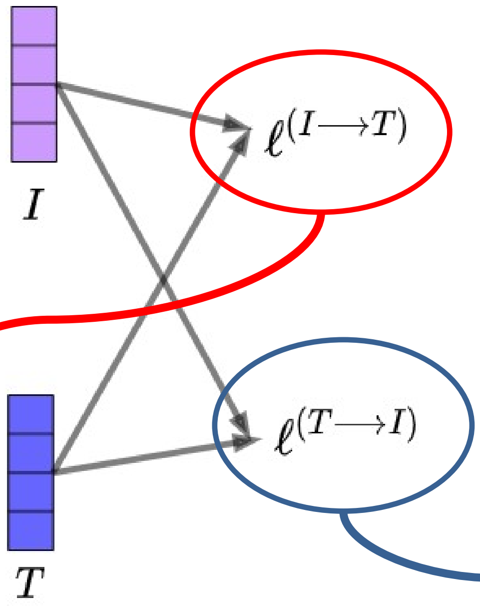
Knowledge Graph Construction



Common Sense Embedding Learning



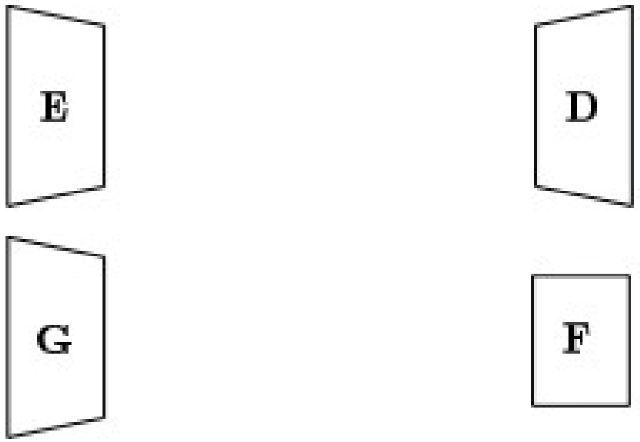
Text-Image Alignment Loss



$$\ell^{(I \rightarrow T)} = -log \frac{exp(s(I_i, T_i)/\tau)}{\sum_{k=1}^{N} exp(s(I_i, T_k)/\tau)} \qquad \ell^{(T \rightarrow I)} = -log \frac{exp(s(T_i, I_i)/\tau)}{\sum_{k=1}^{N} exp(s(T_i, I_k)/\tau)}$$

$$\mathcal{L}_{mm} = \frac{1}{N} \sum_{n=1}^{N} \lambda \ell_n^{(I \rightarrow T)} + (1 - \lambda) \ell_n^{(T \rightarrow I)}$$

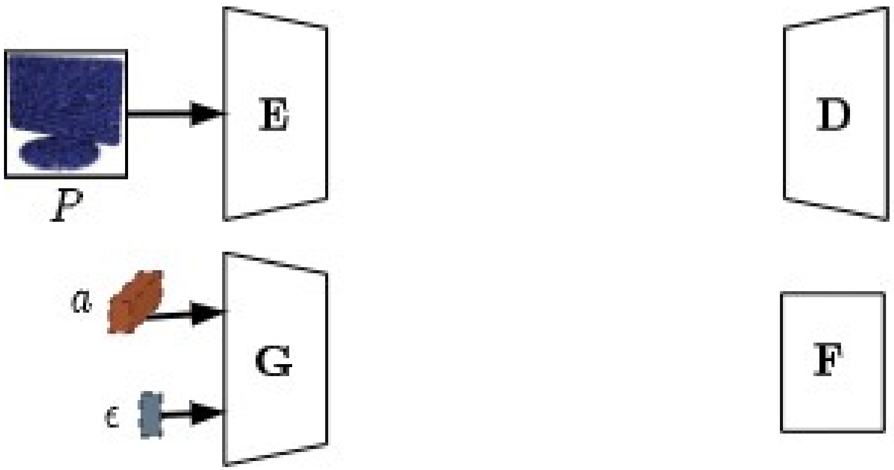
Contrastive GAN Framework



E: Point cloud encoder
G: Generator network
D: Discriminator network
F: Classifier network

Contrastive GAN Framework

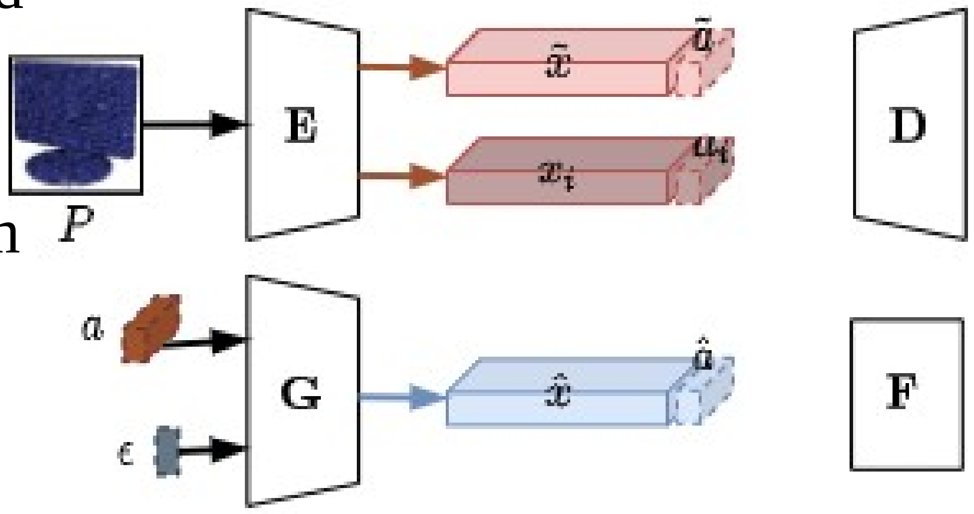
- Provide input 3D point cloud to encoder
- Generator net. take both attribute of class and random noise



E: Point cloud encoder
G: Generator network
D: Discriminator network
F: Classifier network
P: Point cloud sample
a: Attribute vector
ε: noise

Contrastive GAN Framework(2)

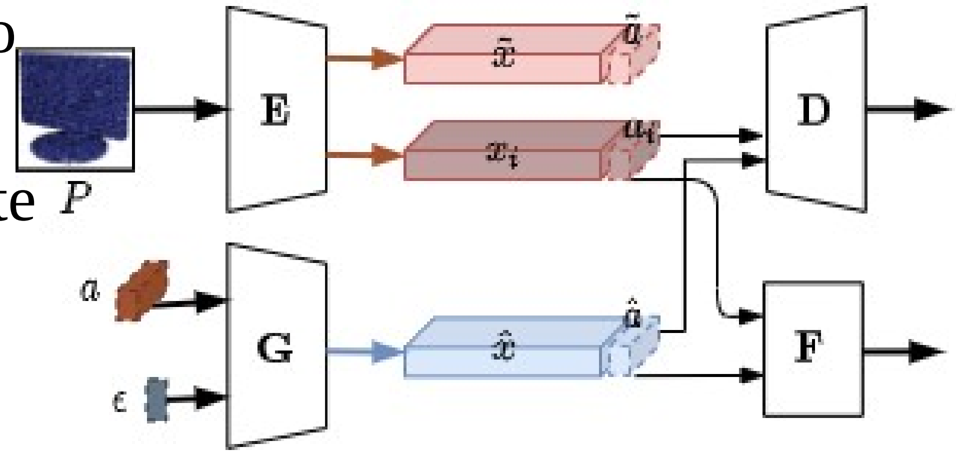
- Provide input 3D point cloud to encoder
- Generator net take both attribute of class and random noise
- Encoder convert each **P** into respective embedding
- Generator create embedding based on given input



E: Point cloud encoder
G: Generator network
D: Discriminator network
F: Classifier network
P: Point cloud sample
a: Attribute vector
ε: noise
x: point cloud embedding

Contrastive GAN Framework(3)

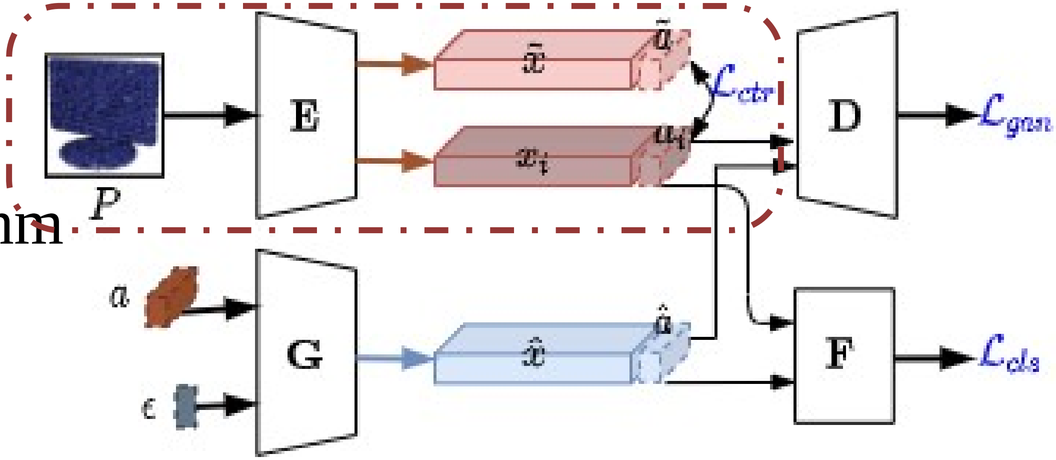
- Provide input 3D point cloud to encoder
- Generator net take both attribute of class and random noise
- Encoder convert each **P** into respective embedding
- Generator create embedding based on given input
- Discriminator classify the given input is real or fake



E: Point cloud encoder
G: Generator network
D: Discriminator network
F: Classifier network
P: Point cloud sample
a: Attribute vector
ε: noise
x: point cloud embedding

Contrastive GAN Framework(4)

- Then calculate three loss function and optimize through backprop algorithm



$$\mathcal{L}_{ctr} = -log \frac{exp(sim(h_i, h^+)/\tau_e)}{exp(sim(h_i, h^+)/\tau_e) + \sum_{k=1}^{K} exp(sim(h_i, h^-)/\tau_e)}$$

$$\mathcal{L}_{gan} = \mathbb{E}_{p(x,a)}[logD(x,a)] + E_{p_G(\hat{x},a)}[log(1 - D(\hat{x},a))]$$

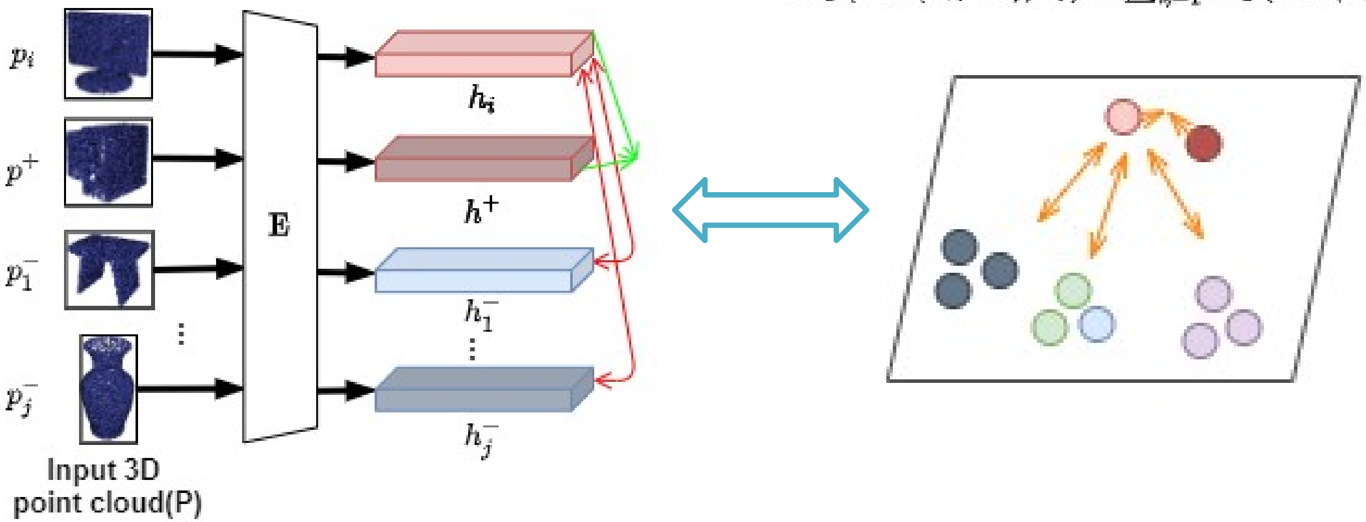
$$\mathcal{L}_{cls} = -\mathbb{E}_{\tilde{x} \sim p_{\tilde{x}}}[logP(y|\tilde{x};\theta)]$$

Optimize: $\max_{D} \min_{G,F,H} \mathcal{L}_{gan} + \alpha \mathcal{L}_{ctr} + \beta \mathcal{L}_{cls}$

E: Point cloud encoder
G: Generator network
D: Discriminator network
F: Classifier network
P: Point cloud sample
a: Attribute vector
ε: noise
x: point cloud embedding

Instance-level Contrastive Embedding

$$\mathcal{L}_{ctr} = -log\frac{exp(sim(h_i, h^+)/\tau_e)}{exp(sim(h_i, h^+)/\tau_e) + \sum_{k=1}^{K} exp(sim(h_i, h^-)/\tau_e)}$$



- Make similar class instance closer to one another
- Repulsing dissimilar classes sample
- Learn it though contrastive loss

Datasets: Real-world

1. RGB-D[7]

- Total class: 51
- Seen/ Unseen: 41/10
- Train/Valid/Test: 8032/2135/1607
- Newly benchmarked real world dataset

2. ScanObjectNN[8]

- Total class: 37
- Seen/ Unseen: 26/11
- Train/Valid/Test: 2902/819/495
- Real world dataset

[7] K. Lai, L. Bo, X. Ren, and D. Fox, “A large-scale hierarchical multi-view rgb-d object dataset,” in 2011 IEEE international conference on robotics and automation. IEEE, 2011, pp. 1817–1824. [8] M. A. Uy, Q.-H. Pham, B.-S. Hua, T. Nguyen, and S.-K. Yeung, “Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data,” in Proceedings of the IEEE/CVF ICCV, 2019.

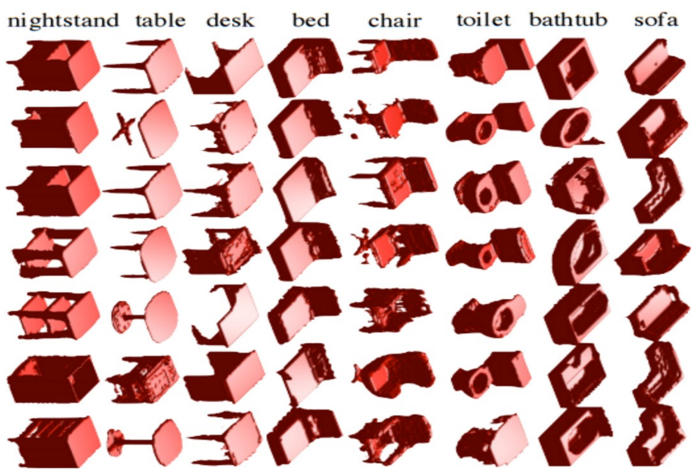
Datasets: Synthetic

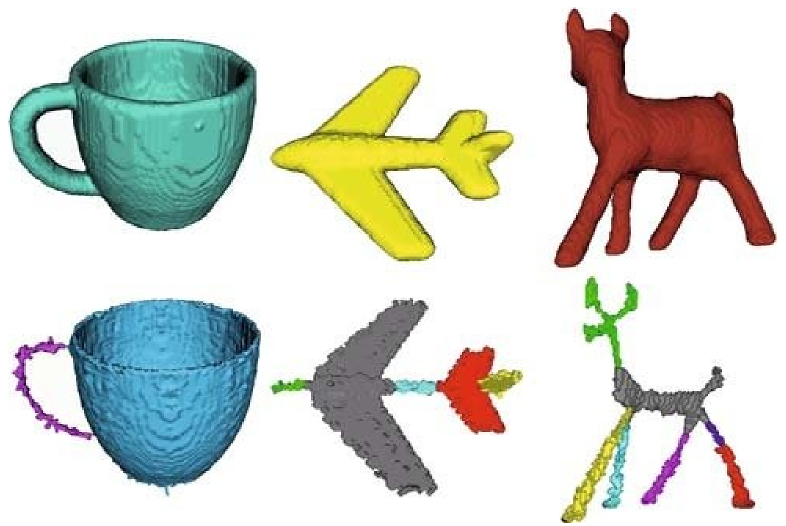
3. ModelNet40[9]

- Total class: 40
- Seen/ Unseen: 30/10
- Train/Valid/Test: 5852/1560/908
- Widely used and synthetic dataset

4. McGill[10]

- Total class: 44
- Seen/ Unseen: 30/14
- Train/Valid/Test: 5852/1560/115
- Synthetic dataset





[9] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, “3d shapenets: A deep representation for volumetric shapes,” in Proceedings of the IEEE CVPR, 2015. [10] K. Siddiqi, J. Zhang, D. Macrini, A. Shokoufandeh, S. Bouix, and S. Dickinson, “Retrieving articulated 3-d models using medial surfaces,” Machine vision and applications, vol. 19, pp. 261–275, 2008.

Overall Results

Method	ModelNet40				McGill				ScanObjectNN			
	ZSL	GZSL			ZSL	GZSL			ZSL	GZSL		
	Acc	Acc_s	Acc_u	HM	Acc	Acc_s	Acc_u	HM	Acc	Acc_s	Acc_u	HM
(2D & End-to-End Network.)												
AREN[11]	19.5	**78.5**	0.0	0.0	13.1	77.4	0.0	0.0	14.5	**79.6**	0.0	0.0
LFGAA[12]	16.2	76.8	0.7	1.4	12.8	73.4	2.3	4.4	11.7	78.4	1.7	3.3
RGEM[13]	18.9	73.8	6.8	12.4	14.1	76.2	4.6	8.6	15.5	77.4	6.2	11.4
APN[14]	18.6	77.3	5.8	10.8	15.0	76.2	6.8	12.5	13.6	76.2	4.2	8.0
GOGE[15]	19.9	75.4	8.3	15.0	12.4	**77.5**	5.8	10.8	14.8	76.4	6.8	12.5
LGE[16]	14.7	77.6	2.3	4.5	11.3	76.8	2.8	5.4	12.7	76.4	1.2	2.4
(2D & Generative Network.)												
f-CLSWGAN[17]	12.8	69.3	10.5	18.3	14.2	73.2	9.1	16.2	15.7	74.8	11.2	19.5
CADA-VAE[18]	15.2	82.6	2.6	5.0	6.8	83.6	8.7	15.8	17.1	78.8	**13.8**	**23.5**
GDAN[19]	18.4	76.4	8.3	15.0	12.8	73.6	9.4	**16.7**	16.5	75.4	9.2	16.4
TF-VAEGAN[20]	23.8	64.4	10.1	18.4	15.5	72.3	**9.4**	16.6	18.2	59.8	13.6	22.1
FG-DFC[21]	20.1	68.5	11.4	19.5	14.6	73.2	8.6	15.4	12.8	71.6	10.7	18.6
CE-GZSL[22]	16.4	70.3	9.8	17.2	14.8	76.2	7.6	13.8	14.2	68.1	11.8	20.1
(3D)												
ZSLPC [2]	28.0	40.1	22.5	28.8	**16.1**	-	-	-	-	-	-	-
G-ZSLPC [3]	**33.9**	53.8	**26.2**	**35.2**	12.5	-	-	-	-	-	-	-
ZSLPC&B [4]	32.5	89.4	6.8	12.7	15.7	71.1	8.7	15.5	24.1	89.4	5.7	10.7
3DGenZ [6]	36.8	47.8	36.5	41.3	9.4	49.6	8.6	14.5	-	-	-	-
MPCN [23]	31.3	70.6	12.6	21.4	17.0	77.6	9.8	17.4	**20.5**	73.4	16.2	26.5
Ours	46.9	69.6	38.9	49.9	19.7	65.0	10.4	17.9	26.6	83.6	20.9	33.5

Table 1: Overall result

[2] A. Cheraghian, S. Rahman, and L. Petersson, “Zero-shot learning of 3d point cloud objects,” in 2019 16th International Conference on Machine Vision Applications (MVA). IEEE, 2019, pp. 1–6. [3] A. Cheraghian, S. Rahman, D. Campbell, and L. Petersson, “Mitigating the hubness problem for zero-shot learning of 3d objects,” in British Machine Vision Conference (BMVC’19), 2019 [4] A. Cheraghian, S. Rahman, D. Campbell, and L. Petersson, “Transductive zero-shot learning for 3d point cloud classification,” in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 923–933 [6] Michele, Björn, et al. "Generative zero-shot learning for semantic segmentation of 3d point clouds." 2021 International Conference

Overall Results (2)

	Acc	Acc_s	Acc_u	HM
Tahmeed *et. al.*[5]	45.22	**85.76**	25.39	39.18
Ours	**52.56**	81.91	**28.73**	**42.54**

Table 2: Overall result on RGB-D dataset

Prompts	Acc	Acc_s	Acc_u	HM
"a photo of a [CLASS]."	42.57	64.10	34.62	44.96
"[class] description."	42.26	65.40	32.24	43.19
"a Point cloud of a [CLASS]."	46.67	64.90	37.50	47.53
"[CLASS]."	**46.95**	**69.60**	**38.86**	**49.88**

Table 3: Result on different text prompt

[5] Muhammad Tahmeed Abdullah. "Improving 3D object Recognition with Contextual Information and Meta Learning." Master's thesis in department of Robotics and Mechatronics Engineering, University of Dhaka (2022).

Ablation Studies: Component Analysis

Only text embedding

Final model

Concatenate of image and text embedding

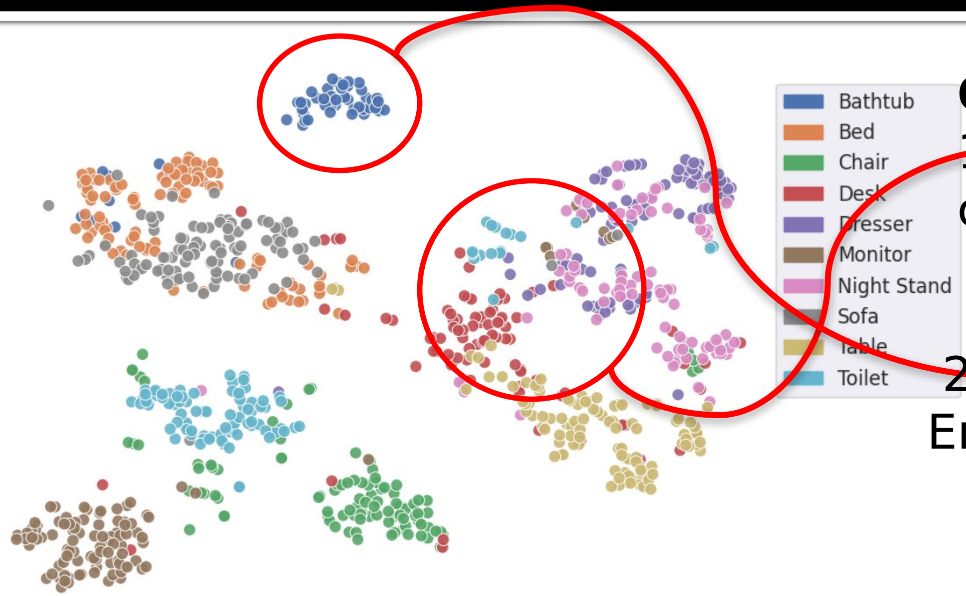
Attribute Feature	SSL modeule	Knowledge transfer	Acc	Acc_s	Acc_u	HM
T			34.89	63.83	23.13	33.92
I $\oplus$ **T**			36.80	67.17	27. 03	36.47
Learning **I** & **T**			40.18	66.47	29. 76	41.11
Learning **I** & **T**	✓		42.57	64.10	34.62	44.96
Learning **I** & **T**	✓	✓	**46.95**	**69.60**	**38.86**	**49.88**

Table 4: Comparing result with different components of model. Here ⊕, **I** and **T** represents concatenate, image embedding and text embedding respectively.

Learning embedding from both modalities

Incorporating point-wise contrastive learning in intermediate feature map

Ablation Studies: t-SNE Analysis



Observations:

1. Several unseen classes share common embedding space(they closer to each other).

2. Some class has distinct 3D visual Embedding(which is expected).

(a) Baseline

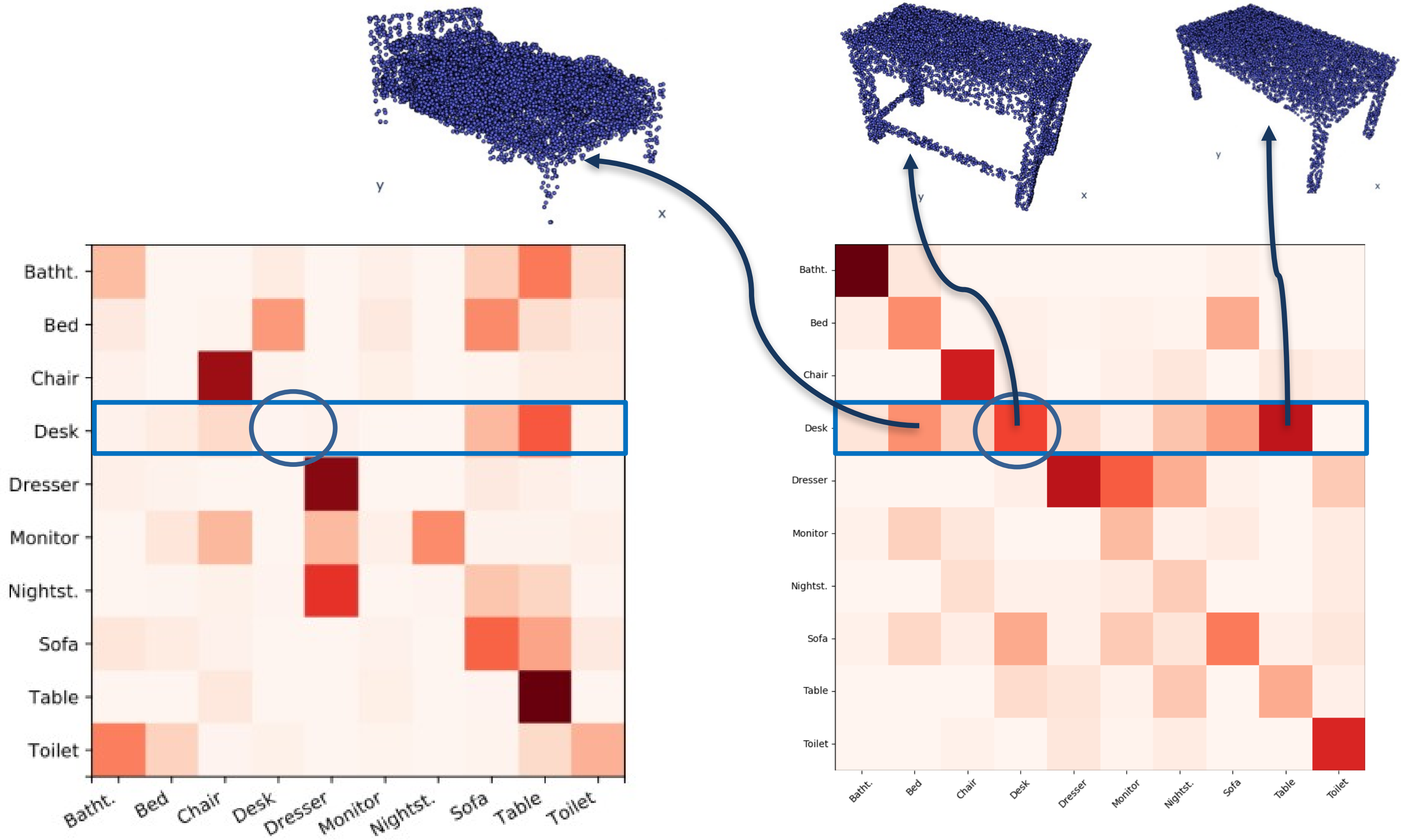
Solutions:

1. Providing an inter and intra-class self-supervised contrastive loss can further improve the embedding

2. Which make instances of similar class closer to each other and repulse the instances of other classes

(b) Ours

Results Analysis: ZSL Confusion Matrix



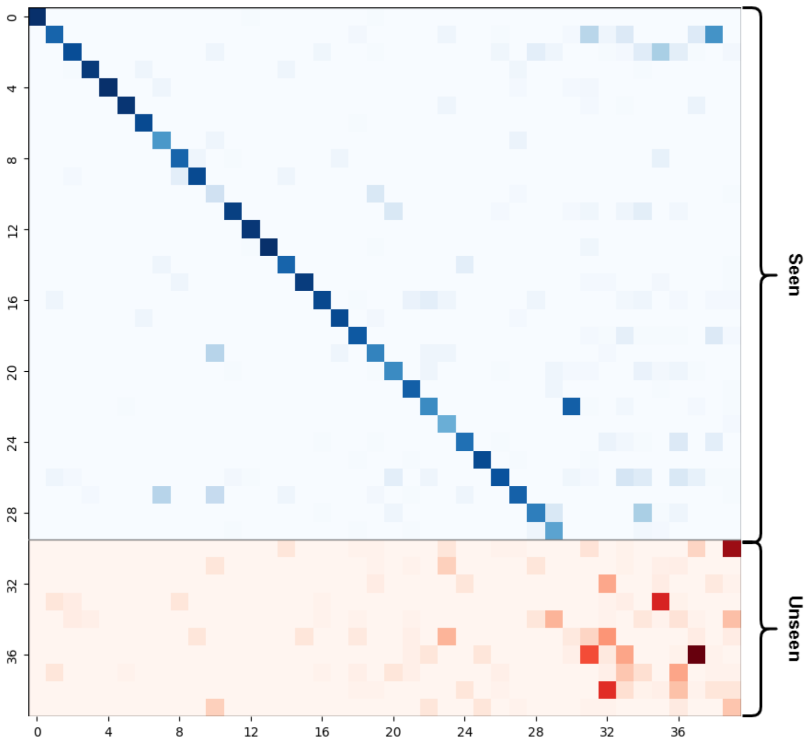
(a) 3DGenZ[6]

(b) Ours

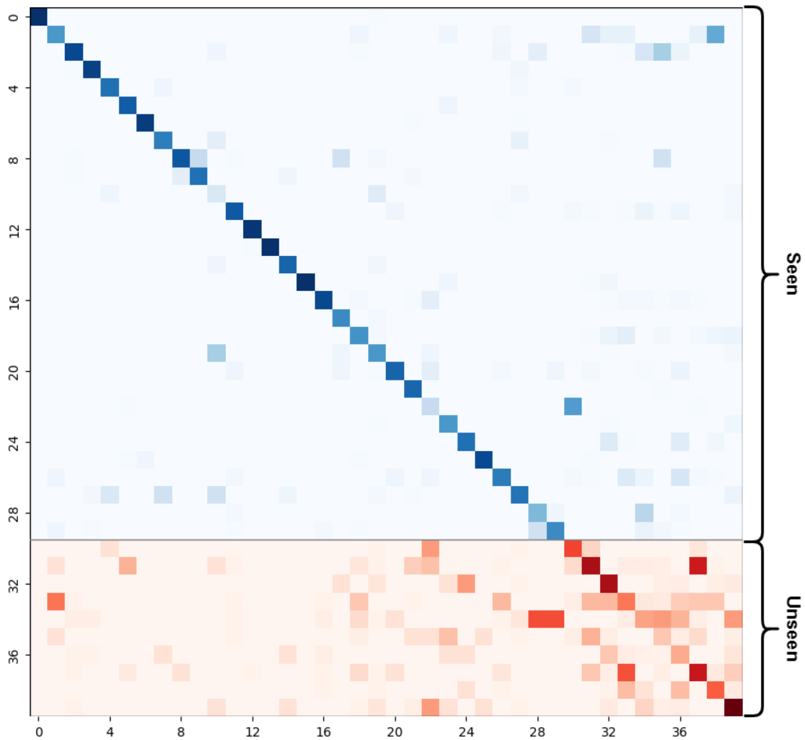
Figure : ZSL confusion matrix

[6] B. Michele, A. Boulch, G. Puy, M. Bucher, and R. Marlet, “Generative zero-shot learning for semantic segmentation of 3d point clouds,” in 2021 International Conference on 3D Vision (3DV). IEEE, 2021, pp. 992–

Results Analysis: GZSL Confusion Matrix



(a) Baseline



(b) Ours

Figure : GZSL confusion matrix

Conclusion

- We use the commonsense knowledge graph with GCN to produce commonsense embedding
- We propose co-attention based multi-model learning to distil info from both semantic and 2D visual data
- Also design an contrastive based generative framework to distil unseen augmented features
- Experimental findings showed that our method outperformed existing approach on a 3D visual dataset
- Further research should be done on other task like segmentation, detection in Zero shot setup

Reference

[1] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, “Pointnet: Deep learning on point sets for 3d classification and segmentation,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 652–660

[2] A. Cheraghian, S. Rahman, and L. Petersson, “Zero-shot learning of 3d point cloud objects,” in 2019 16th International Conference on Machine Vision Applications (MVA). IEEE, 2019, pp. 1–6.

[3] A. Cheraghian, S. Rahman, D. Campbell, and L. Petersson, “Mitigating the hubness problem for zero-shot learning of 3d objects,” arXiv preprint arXiv:1907.06371, 2019

[4] A. Cheraghian, S. Rahman, D. Campbell, and L. Petersson, “Transductive zero-shot learning for 3d point cloud classification,” in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 923–933.

[5] Muhammad Tahmeed Abdullah. "Improving 3D object Recognition with Contextual Information and Meta Learning." Master's thesis in department of Robotics and Mechatronics Engineering, University of Dhaka (2022).

[6] Michele, Björn, et al. "Generative zero-shot learning for semantic segmentation of 3d point clouds." 2021 International Conference on 3D Vision (3DV). IEEE, 2021.

[7] K. Lai, L. Bo, X. Ren, and D. Fox, “A large-scale hierarchical multi-view rgb-d object dataset,” in 2011 IEEE international conference on robotics and automation. IEEE, 2011, pp. 1817–1824.

[8] M. A. Uy, Q.-H. Pham, B.-S. Hua, T. Nguyen, and S.-K. Yeung, “Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data,” in Proceedings of the IEEE/CVF ICCV, 2019.

[9] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, “3d shapenets: A deep representation for volumetric shapes,” in Proceedings of the IEEE CVPR, 2015.

[10] K. Siddiqi, J. Zhang, D. Macrini, A. Shokoufandeh, S. Bouix, and S. Dickinson, “Retrieving articulated 3-d models using medial surfaces,” Machine vision and applications, vol. 19, pp.

Reference(2)

[11] G.-S. Xie, L. Liu, X. Jin, F. Zhu, Z. Zhang, J. Qin, Y. Yao, and L. Shao, “Attentive region embedding network for zero-shot learning,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp.
9384–9393
[12] Y. Liu, J. Guo, D. Cai, and X. He, “Attribute attention for semantic disambiguation in zero-shot learning,” in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 6698–6707.
[13] G.-S. Xie, L. Liu, F. Zhu, F. Zhao, Z. Zhang, Y. Yao, J. Qin, and L. Shao, “Region graph embedding network for zero-shot learning,” in Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16. Springer, 2020, pp. 562–580.
[14] W. Xu, Y. Xian, J. Wang, B. Schiele, and Z. Akata, “Attribute prototype network for zero-shot learning,” Advances in Neural Information Processing Systems, vol. 33, pp. 21 969–21 980, 2020.
[15] Y. Liu, L. Zhou, X. Bai, Y. Huang, L. Gu, J. Zhou, and T. Harada, “Goal-oriented gaze estimation for zero-shot learning,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 3794–3803.
[16] M. F. Naeem, Y. Xian, F. Tombari, and Z. Akata, “Learning graph embeddings for compositional zero-shot learning,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 953–962.
[17] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata, “Feature generating networks for zero-shot learning,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 5542–5551
[18] E. Schonfeld, S. Ebrahimi, S. Sinha, T. Darrell, and Z. Akata, “Generalized zeroand few-shot learning via aligned variational autoencoders,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 8247–8255.

Reference(3)

[21] D. Huynh and E. Elhamifar, “Compositional zero-shot learning via fine-grained dense feature composition,” Advances in Neural Information Processing Systems, vol. 33, pp. 19 849–19 860, 2020.
[22] Z. Han, Z. Fu, S. Chen, and J. Yang, “Contrastive embedding for generalized zero-shot learning,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2371–238
[23] Y. Cao, Y. Su, G. Lin, and Q. Wu, “Mutex parts collaborative network for 3d point cloud zero-shot classification,” in Available at SSRN 4332138, 2023.

Thank You

Questions?