3D Visual Reasoning in Limited Label and Data Scarce Scenarios

Deepti Hegde

Research Statement

1. Introduction

Visual data represented in three dimensions is a rich source of information of the world around us, making it extremely useful in teaching machines to perform perception and reasoning tasks. There exist a variety of datasets curated for 3D visual reasoning, from LiDAR datasets for autonomous driving applications to rich 3D shape repositories. However, leveraging this type of data in deep learning research comes with its own set of challenges. The availability of such annotated 3D datasets is limited compared to 2D images. My research deals with training robust 3D perception models in scenarios where there is limited access to labelled data or even any data at all. Over the course of my doctoral graduate program, I have focused on three areas, of which I have gone into detail in the sections below. First, I addressed the issue of distribution shift in LiDAR datasets when training 3D object detection networks [1–3] (Section 2). Second, I worked on developing a self supervised learning framework to enable data-efficient fine-tuning of 3D object detection networks [4] (Section 3). Third, I explored the role of language to perform zero-shot recognition, 3D shape retrieval, and referring detection [5] (Section 4).

2. Addressing distribution shift in LiDAR datasets

LiDAR point clouds collected from different environments vary widely between one another in terms of point cloud density, scene properties and object dimensions due to different modes of capture, locations and weather conditions. LiDAR scenes from different datasets have large differences that are easily visible. A 3D object detection network trained on a particular source distribution drops in performance when evaluated on a different target distribution.

Unsupervised domain adaptation has been broadly successful in addressing this problem for 3D object detection networks [6–10], but depends on annotated source-domain data during adaptation, limiting its applicability in scenarios where it is unavailable due to privacy concerns or memory constraints. We propose two works that address this issue by adapting 3D object detection networks to a target distribution in a "source-free" setting, in which we do not access any source distribution samples.

Source-free unsupervised domain adaptation: In [1], we propose AttProto, an unsupervised, source-free domain adaptation framework for 3D object detection that addresses the issue of incorrect, over-confident pseudo-labels during self-training by training a network to learn noise-free class prototypes. In [2], we propose an uncertainty-aware mean teacher framework which implicitly filters incorrect pseudo-labels during training. Leveraging model uncertainty allows the mean teacher network to perform implicit filtering by down-weighing losses corresponding to uncertain pseudo-labels. We demonstrate our domain adaptation method on an adverse weather dataset created by augmenting lidar scenes from KITTI with rain, snow, and fog and show that it out-performs current domain adaptation frameworks.

Domain generalization: Although a very relevant and useful problem setting, we point out some issues with the domain adaptation formulation. This formulation can be unrealistic, as the target domain characteristics are often unknown and can change dynamically. Additionally, the adapted model is suitable only for the target domain it is trained for, and must be re-trained for every new target distribution. In contrast, we formulate and propose a method to address the domain **generalization** (DG) problem, which is a more practical and challenging setting for the 3D object detection task [11, 12]. In the DG setting, no information about the target domain(s) is available during training. The guiding principle of DG is that training a model able to generalize over diverse source domains can help generalize to unseen target distributions [13]. Additionally, we explore the role of multi-modal data for training robust detectors. We suggest that including image information helps not only the baseline performance, but also in training networks robust to distribution shifts. Images provide dense color and texture information, while LiDAR point clouds provide sparse but accurate depth measurements. In [3], we propose CLIX^{3D}, a multimodal fusion and supervised contrastive learning framework for 3D object detection that performs alignment of object features from same-class samples of different domains while pushing

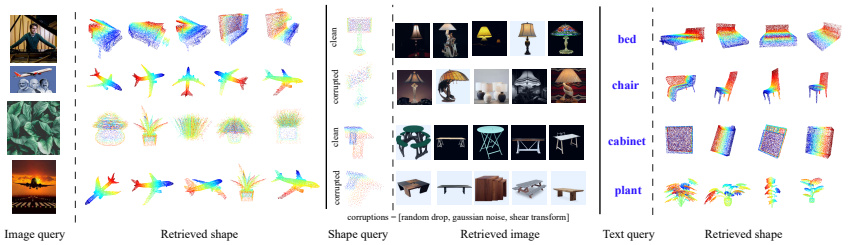

Fig. 1: Retrieval results for CG3D.

the features from different classes apart. We show that **E-SSL^{3D}** yields state-of-the-art domain generalization performance under multiple dataset shifts.

3. Representation learning for 3D object detection

Self-supervised learning (SSL) enables the learning of generic visual representations of unlabelled data by completing tasks designed based on human intuition about what information can be inferred from its inherent properties, without the need for explicit supervision. The availability of large amounts of unlabelled LiDAR data thus makes SSL pre-training methods a natural choice for improving performance of perception tasks when a limited amount of annotated data is available. Using pre-training loss functions that encourage *equivariance* of features under certain transformations provides a strong self-supervision signal while also retaining information of geometric relationships between transformed feature representations. This can enable improved performance in downstream tasks that are equivariant to such transformations. In [4], we propose ESSL^{3D}, a spatio-temporal equivariant learning framework by considering both spatial and temporal augmentations jointly. Our experiments show that the best performance arises with a pre-training approach that encourages equivariance to translation, scaling, and flip, rotation and scene flow.

4. Leveraging Language for 3D Visual Reasoning

Comprehending the semantics and characteristics 3D shapes is crucial for addressing a wide range of issues in downstream tasks, and can enable open-world scene understanding, removing the need for annotated samples. Vision-language models like CLIP [17] have been widely adopted for various tasks due to their impressive zero-shot capabilities. However, CLIP is not suitable for extracting 3D geometric features as it was trained on only images and text by natural language supervision. In [5], we propose the framework CLIP Goes 3D (CG3D) where a 3D encoder is learned to exhibit zero-shot capabilities. CG3D is trained using triplets of point clouds, corresponding rendered 2D images, and texts using natural language supervision. To align the features in a multimodal embedding space, we utilize contrastive loss on 3D features obtained from the 3D encoder, as well as visual and text features extracted from CLIP. In order to mitigate distribution shift, we employ prompt tuning and introduce trainable parameters in the input space to shift CLIP towards the 3D pre-training dataset utilized in CG3D. We demonstrate capabilities in zero-shot, open scene understanding, and retrieval tasks. Some retrieval results can be seen in Figure 1. Further, it also serves as strong starting weights for fine-tuning in downstream 3D recognition tasks.

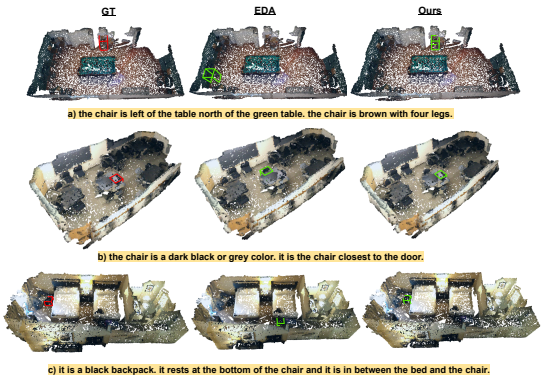

Fig. 2: A qualitative comparison of visual grounding results on 3 scenes from ScanNet [14] for the ScanRefer [15] benchmark. The referring sentence expression is below each scene. The first column shows the ground truth bounding box in red. The second column shows the box detected by EDA [16] in green. The third column shows the box detected by our method in green.

Connecting vision and language through referential grounding has emerged as an important task in 3D scene understanding. The objective of this task is to localize objects in a scene based on a descriptive language query that describes their attributes and relative position. This requires learning the semantic relationships between the point cloud and text modalities. We thus address the more challenging task of referring detection and propose 3D-SPL, a novel visual grounding framework that performs optimal cross-modal alignment and improves object matching by incorporating key insights on language modulation. We propose that maintaining a fixed language representation during alignment and modulating the language representation during matching is optimal. This is because solely aligning the vision features to a fixed language representation allows the transformer to focus on learning visual information relevant to the referring expression during alignment. Our model demonstrates significant improvements in grounding performance and achieves state-of-the-art results on two widely used 3D visual grounding benchmarks. We show qualitative results in Figure 2, demonstrating superior localization performance.

5. Conclusion

In summary, my doctoral research has focused on addressing critical challenges in training robust 3D perception models using limited or unlabeled data. Specifically, I tackled the issue of distribution shift in LiDAR datasets during the training of 3D object detection networks. Additionally, I developed a self-supervised learning framework for data-efficient fine-tuning and explored the role of language in zero-shot recognition, 3D shape retrieval, and referring detection. These contributions pave the way for more effective and adaptable 3D perception systems, even in scenarios with scarce data availability.

References

1. D. Hegde and V. M. Patel, "Attentive prototypes for source-free unsupervised domain adaptive 3d object detection," *arXiv preprint arXiv:2111.15656*, 2021.
2. D. Hegde, V. Kilic, V. Sindagi, A. B. Cooper, M. Foster, and V. M. Patel, "Source-free unsupervised domain adaptation for 3d object detection in adverse weather," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6973–6980, IEEE, 2023.
3. D. Hegde, S. Lohit, K.-C. Peng, M. J. Jones, and V. M. Patel, "Multimodal 3d object detection on unseen domains," *arXiv preprint arXiv:2404.11764*, 2024.
4. D. Hegde, S. Lohit, K.-C. Peng, M. J. Jones, and V. M. Patel, "Equivariant spatio-temporal self-supervision for lidar object detection," *arXiv preprint arXiv:2404.11737*, 2024.
5. D. Hegde, J. M. J. Valanarasu, and V. Patel, "Clip goes 3d: Leveraging prompt tuning for language grounded 3d recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2028–2038, 2023.
6. Q. Xu, Y. Zhou, W. Wang, C. R. Qi, and D. Anguelov, "Grid-GCN for fast and scalable point cloud learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.
7. J. Yang, S. Shi, Z. Wang, H. Li, and X. Qi, "ST3D: Self-training for unsupervised domain adaptation on 3D object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.
8. C. Saltori, S. Lathuili'ere, N. Sebe, E. Ricci, and F. Galasso, "SF-UDA^{3D}: Source-free unsupervised domain adaptation for LiDAR-based 3D object detection," *2020 International Conference on 3D Vision (3DV)*, pp. 771–780, 2020.
9. B. Caine, R. Roelofs, V. Vasudevan, J. Ngiam, Y. Chai, Z. Chen, and J. Shlens, "Pseudo-labeling for scalable 3D object detection," *ArXiv*, vol. abs/2103.02093, 2021.
10. Z. Luo, Z. Cai, C. Zhou, G.-D. Zhang, H. Zhao, S. Yi, S. Lu, H. Li, S. Zhang, and Z. Liu, "Unsupervised domain adaptive 3D detection with multi-level consistency," *ArXiv*, vol. abs/2107.11355, 2021.
11. G. Blanchard, G. Lee, and C. Scott, "Generalizing from several related classification tasks to a new unlabeled sample," in *Advances in Neural Information Processing Systems* (J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, eds.), vol. 24, Curran Associates, Inc., 2011.
12. K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.
13. S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, "Analysis of representations for domain adaptation," *Advances in neural information processing systems*, vol. 19, 2006.
14. A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5828–5839, 2017.
15. D. Z. Chen, A. X. Chang, and M. Nießner, "Scanrefer: 3d object localization in rgb-d scans using natural language," in *European conference on computer vision*, pp. 202–221, Springer, 2020.
16. Y. Wu, X. Cheng, R. Zhang, Z. Cheng, and J. Zhang, "Eda: Explicit text-decoupling and dense alignment for 3d visual and language learning," *arXiv preprint arXiv:2209.14941*, 2022.
17. A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning (ICML)*, 2021.