Advancing Scene Understanding and AI-Generated Content through 3D Synergistic Multimodal Research

Linlian Jiang

jiangll21@mails.jlu.edu.cn
School of Artifical Intelligence, Jilin University

Self-assessment: Linlian Jiang is an exceptionally self-motivated individual, demonstrating outstanding observational skills and a remarkable aptitude for problem-solving. Her ability to summarize information effectively is unparalleled. Throughout her graduate studies, Linlian has diligently cultivated a vast knowledge base under the rigorous guidance of Prof. Rui Ma. Additionally, she has honed her programming skills through rigorous research training, becoming highly proficient in various programming languages. Linlian's disciplined and hardworking nature is evident in her meticulous planning and clear goals, as she strives to publish papers in prestigious computer conferences. Besides, She participated in the application and implementation of multiple national-level projects, and also participated in many top-level competitions and has achieved excellent results.

Keywords: Multimodal , 3D Generation, Scene Understanding, AIGC

1. Introduction

The world we live in is inherently a three-dimensional world. For example, the people, objects, and natural scenes around us are all three-dimensional. However, in order to represent this three-dimensional information in a computer, it needs to be digitized. In computer graphics, three-dimensional content can be primarily represented as meshes, point clouds, or voxels, among others. Three-dimensional content has a wide range of applications in computer graphics and related fields, such as 3D games and movie special effects. By utilizing 3D reconstruction techniques, it is also possible to create virtual cities, engage in 3D printing, and conduct system simulations, among other things.

Currently, there are significant challenges in the practical application of 3D content modeling. 3D object modeling primarily involves manual mesh editing, which requires a lot of detailed operations. On the other hand, the main method for 3D scene modeling is to gradually place individual 3D models into 3D modeling software and adjust their positions. This process requires certain 3D modeling skills and is not beginner-friendly. It can be time-consuming and labor-intensive when creating complex scenes. To address this issue, we aim to abstractly model complex three-dimensional data, extract and learn prior knowledge from existing datasets, and design intelligent, efficient, and controllable algorithms or neural networks to automatically reconstruct or generate intricate, realistic three-dimensional content. We employ data-driven and learning-based approaches to achieve three-dimensional content modeling. The ideal modeling involves a high degree of intelligence and controllability, where the system requires little to no manual modeling, is user-friendly for beginners, ensures the novelty and randomness of results, and meets specific modeling requirements of users.

Computer Vision (CV) and Natural Language Processing (NLP) have witnessed remarkable technological advancements in deep learning research. Over the years, significant progress has been made in single-modality models **[1][2]**. However, more recently, there has been a surge in research activities dedicated to the development of large-scale multimodal approaches **[3] [4] [5]**. These approaches aim to leverage the power of both CV and NLP to enhance the understanding and processing of complex data. By combining visual and textual information, researchers are pushing the boundaries of AI, enabling machines

to interpret and analyze multimodal data more effectively than ever before.

During my PhD studies, the research targets the following key objectives: (1) Synergistic Multimodal Scene Understanding, and (2) Controllable High-Quality 3D Content Generation.

2. Challenges

2.1. Multimodal representation and conflict

Typically, multimodal representation involves mapping individual input streams, such as images, videos, or sound, into high-level semantic representations—either linear or nonlinear. Multimodal representation takes advantage of the correlation among different sensory inputs by aggregating their spatial outputs. This requires adapting deep learning models to accurately represent both the source and target modality structures. For instance, representing a 2D image involves visual patterns, introducing challenges in integrating this data structure with non-visual concepts. In scenarios like autonomous driving with LiDAR and embedded sensors, weather conditions can impact visual perception. Additionally, the high dimensionality of the state space presents challenges in both structured and unstructured locations.

Combining visual and non-visual cues is challenging due to differences in data sparsity. Learning a joint embedding becomes crucial for constructing shared representation spaces, highlighting the necessity of multimodal fusion approaches. Multimodal data, available in diverse formats, poses challenges in extracting valuable information due to its variability. Despite the richness of multimodal information, its large dimension increases computational complexity and memory consumption during acquisition and processing. Synchronizing temporal data enhances correlation between representation levels. While feature-level fusion is more flexible than decision-level fusion for homogeneous data samples, existing dimensionality reduction algorithms (e.g., k-NN, PCA) compress input signals or extract low-dimensional patterns to facilitate analysis and processing, thereby mitigating computational costs.

2.2. 3D Vision-language Learning

Recently, there has been a growing interest in the realm of 3D vision-language (3D-VL) learning. In contrast to conventional scene understanding, 3D-VL tasks establish connections between the physical world and natural language, a vital aspect for achieving embodied intelligence [6]. Within this emerging field, Chen et al. [7] and Achlioptas et al. [8] concurrently introduced the ScanRefer and ReferIt3D datasets, aiming to benchmark natural language grounding to 3D object properties and relations. Beyond 3D visual grounding, Azuma et al. [9] developed a 3D question-answering dataset named ScanQA, challenging models to answer questions about objects and their relations within a 3D scene. More recently, Ma et al. [10] proposed a situated reasoning task called SQA3D, focusing on embodied scene understanding in 3D environments. Various models have been put forward for these benchmarks [6][7] [11] [12] [13]. Notably, 3D-SPS [11] and BUTD-DETR [12] adopt a cross-attention mechanism and language guidance to progressively discover the target object. 3DVG [11], MVT [14], and ViL3DRel [15] address 3D visual grounding by explicitly incorporating spatial relation information into their models. Despite achieving impressive results in bridging 3D vision and language, these approaches still heavily rely on task-specific knowledge in model design and employ sophisticated optimization techniques.

2.3. Scene Completion

Completion algorithms employed interpolation [16]or energy minimization [17] [18] [19] methods to address minor gaps in data. Initially, completion efforts were focused on filling in missing portions of object shapes, aiming to generate representations of objects free from occlusions. Various trends, such as leveraging symmetry [20] [21] [22], have been explored and are comprehensively discussed in [23]. Another prevalent approach involves utilizing pre-existing 3D models to optimally match sparse input data [24] [25]. In recent years, advancements in model-based techniques and the availability of expansive datasets have broadened the application scope. This expansion enables the inference of complete occluded segments not only in scanned objects

[26] [27] but also in entire scenes [28].

Conventional segmentation methodologies, as discussed in [29], were grounded in manually crafted features, statistical rules, and bottom-up processes, coupled with traditional classifiers. The landscape underwent a significant transformation with the advent of deep learning [30]. Initial forays into 3D deep techniques relied on multiviews processed by 2D CNNs [31], but they swiftly gave way to the utilization of 3D CNNs, which operate on voxels [32]. However, this shift came with challenges related to memory and computational limitations. Point-based networks [33] addressed these issues by working directly with points and gained popularity in segmentation tasks, although generative tasks remained a formidable challenge. Consequently, if the point space is pre-defined, SSC (Semantic Scene Completion) can be viewed as a task focused on segmenting points [34].

In contrast to Semantic Scene Completion (SSC), semantic instance completion specifically operates at the level of individual objects and does not extend to completing elements of the scene background such as walls and floors. Most existing approaches necessitate the detection of object instances, completing them either through CAD models [35] [36]or a dedicated completion head [37]. Noteworthy is the work presented in [38], which not only identifies objects but also the layout of the scene, generating a comprehensive lightweight CAD representation. These strategies typically rely on 3D object detection [39] [40] or instance segmentation. Additionally, recent methodologies address the challenge of completing multiple objects from single images, either by employing individual object detection and reconstruction [41]or by holistically understanding the scene structure for joint completion [42]

2.4. 3D Generation

Deep learning-based techniques for generating 3D shapes are gaining significant traction in both computer vision and architectural design. This involves a comparative analysis of the most recent methods in 3D object generation using deep generative models (DGMs), which encompass Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), 3D-aware images, and diffusion models. The prospect of attaining finely adjustable 3D shapes through text is captivating. Recent groundbreaking research has showcased the effectiveness of pre-trained text-to-image diffusion models in optimizing neural radiance fields, yielding notable outcomes in text-to-3D synthesis. Nonetheless, this approach is not without its challenges, and some researchers are actively working to address these issues.

The conventional method of editing 3D models requires specialized tools and extensive training to manually shape, extrude, and re-texture given objects. This approach is both labor-intensive and resource-intensive. Recently, there has been a shift towards text-guided 3D model editing, as highlighted in this section. Instruct-NeRF2NeRF introduces an innovative text-instructable editing approach for NeRF scenes. It employs an iterative image-based diffusion model (InstructPix2Pix) to edit the input image while optimizing the underlying scene. This process results in an optimized 3D scene that adheres to the provided editing instructions. Experimental results demonstrate the method's capability to edit large-scale real-world scenes, achieving more realistic and targeted edits compared to previous approaches. Instruct 3D-to-3D introduces a novel 3D-to-3D transformation method utilizing a pre-trained image-to-image diffusion model. The method also proposes dynamic scaling and explicit conditioning on the input source 3D scene to enhance 3D consistency and controllability.

While Text-to-3D can produce remarkable outcomes, the inherent lack of constraints in text-to-image diffusion models often leads to guiding collapse. This limitation hampers their ability to precisely link object semantics with specific 3D structures, contributing to a longstanding issue of poor controllability.

3. Methodology

3.1. Multi-modal Fusion

The essence of multimodality lies in alignment, and there are various ways to achieve alignment, such as entity alignment, attention mechanisms, or initially aligning events and then performing fusion. Fusion methods include linking, grounding, structure, and others. Efficiently structuring how to align information from different modalities or determining which modalities work better

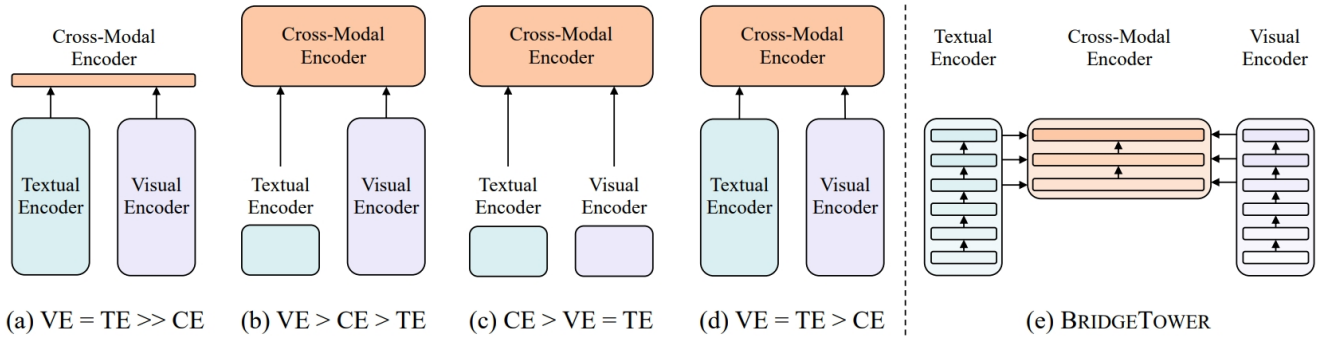
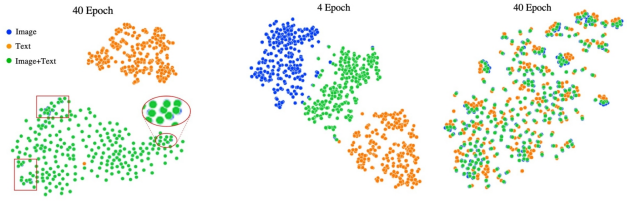


Figura 1: (a) to (d) represent the four categories of existing TWO-TOWER vision-language models, while (e) provides a concise depiction of the BRIDGETOWER architecture. In this context, VE, TE, and CE stand for Visual Encoder, Textual Encoder, and Cross-modal Encoder, respectively. The height of each rectangle corresponds to its relative computational cost. VE = TE signifies that the visual encoder and the textual encoder share a similar number of parameters or computational costs.

together to assist downstream tasks is a significant challenge.

As shown in Figure 1, BRIDGETOWER introduces multiple bridge layers designed to establish connections between the upper layers of unimodal encoders and each layer of the cross-modal encoder. This facilitates efficient bottom-up cross-modal alignment and fusion between visual and textual representations at different semantic levels from pre-trained unimodal encoders in the cross-modal encoder. Despite being pre-trained with only 4 million images, BRIDGETOWER attains state-of-the-art performance across various downstream vision-language tasks.



(a) Naive single-tower ITC. (b) OneR at the beginning and the end of the training.

Figura 2: overview of the Spatial U-Net.We have designed a novel spatial attention mechanism that effectively generates 3D shape during the inference stage.

Contrastive learning, a variant of distance learning, seeks to acquire invariant features from two correlated representations. By treating an image and its caption as distinct perspectives of shared mutual information, the objective is to train a model that learns a cohesive vision-language representation space encompassing both modalities in a modality-agnostic manner. As shown in Figure 2, this approach uncovers distinctive attributes that set OneR apart from prior endeavors focusing on modality-specific representation spaces, showcasing capabilities like zero-shot object localization, text-guided visual reasoning, and multimodal retrieval.

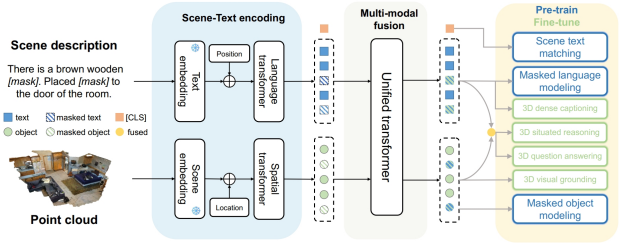


Figura 3: The architectural framework of our 3D-VisTA comprises modules for text encoding, scene encoding, and multi-modal fusion. 3D-VisTA undergoes pre-training through self-supervised learning objectives, encompassing masked language modeling, masked object modeling, and scene-text matching.

In 3D vision-language pertaining, in figure 3, 3D-VisTA straightforwardly employs self-attention layers for both single-modal modeling and multi-modal fusion, devoid of any intricate task-specific design. 3D-VisTA concatenates the text and 3D object tokens and inputs them into an L-layer Transformer for multi-modal fusion. Learnable type embeddings are incorporated into the tokens to distinguish between text and 3D ob-

jects. The output of the multi-modal fusion module is denoted as wcls, w1:M, o1:N for [CLS], text tokens, and 3D object tokens, respectively.

3.2. Generative Model

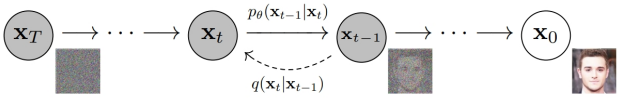


Figura 4: Diffusion Model

As shown in Figure 4, the diffusion probability model was initially inspired by non-equilibrium thermodynamics and proposed as a latent variable generative model. Such models involve two processes: first, a forward process that gradually disrupts the data distribution by adding noise at multiple scales, and then a reverse process that learns how to recover the data structure . From this perspective, the diffusion model can be viewed as a deeply hierarchical Variational Auto-Encoder (VAE), where the disruption and recovery processes correspond to the encoding and decoding processes in VAE. Consequently, much research has focused on learning the encoding and decoding processes, coupled with the design of variational lower bounds, to enhance model performance. Alternatively, the diffusion model's processes can be regarded as a discretization of a stochastic differential equation (SDE) , where the forward and reverse processes correspond to the forward SDE and reverse SDE, respectively. Analyzing the diffusion model through SDE enriches theoretical results, facilitates model improvements, and particularly shines in terms of sampling strategies.

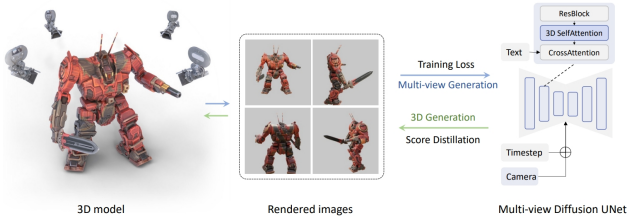


Figura 5: MVDream, Illustration of the multi-view diffusion model

The utilization of diffusion models for 3D generation is truly remarkable, As shown in Figure 5, MVDream is a multi-view diffusion model designed to generate consistent multi-view images based on a given text prompt. Leveraging insights from both 2D and 3D data, this model combines the generalizability of 2D diffusion models with the coherence found in 3D renderings. The multi-view prior employed in MVDream functions as a versatile 3D prior, agnostic to specific 3D representations. This prior proves beneficial for 3D generation using Score Distillation Sampling, significantly improving the consistency and stability of existing 2D-lifting methods. Moreover, MVDream can learn novel concepts from a limited set of 2D examples, akin to DreamBooth but specifically tailored for 3D generation.

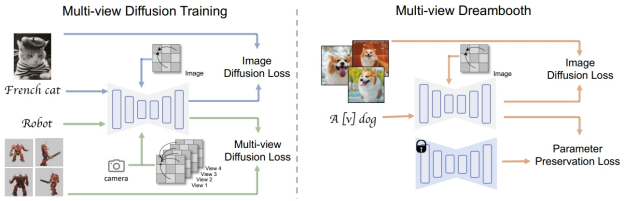


Figura 6: The training pipeline of MVDream

As shown in Figure 6, MVDream initial strategy involves adjusting the attention layers to account for cross-view dependencies while retaining the rest of the network as a 2D model designed for operation within a single image. However, the direct application of temporal attention proves insufficient in capturing multi-view consistency, and despite fine-tuning the model on a 3D rendered dataset, content drift remains an issue. Consequently, the article adopts a 3D attention approach. Specifically, the original 2D self-attention layer is transformed into a 3D version by establishing connections.

4. Conclusiones and Goals

This research direction aims to explore the synergy between different modalities to enhance scene understanding. By integrating information from multiple sources, such as visual and textual inputs, the objective is to develop comprehensive models for scene understanding that surpass the limitations of unimodal approaches. The investigation into Artificial Intelligence-Generated Content will involve exploring the generation of content using advanced AI techniques. This encompasses creating algorithms and models capable of producing diverse and high-quality content across various domains, contributing to the broader field

of AI-driven content creation.

These research directions collectively aim to contribute to the advancements in 3D reconstruction methodologies, multimodal scene understanding, and the generation of content through Artificial Intelligence. The interdisciplinary nature of these topics will provide a holistic perspective on cutting-edge developments in computer vision and artificial intelligence.

During my doctoral studies, I plan to diligently acquire in-depth knowledge in my field, actively engage in discussions with my advisor, and strive to publish at least two high-quality research papers.

5. Timeframe for the Study

After discussions with Prof. Zichun Zhong, my time plan during my PhD is as follows:

Year 1: Synergistic Multimodal Scene Understanding. In the first year, our focus will be on advancing synergistic multimodal scene understanding. Drawing inspiration from successful models like CLIP , CrossPoint, and Y2Seq2Seq, our objective is to bridge the modality gap in the feature space of cross-modal data. Through the exploration of contrastive learning techniques and effective cross-modal alignment methods, we aim to develop a comprehensive model that outperforms existing single-modal methods. Expected outputs for this year include algorithmic advancements, code implementations, and preliminary results demonstrating improved cross-modal scene understanding. And submit advanced research findings to CVPR.

Year 2: Controllable High-Quality 3D Content Generation. In the second year, our research will shift towards controllable high-quality 3D content generation. Building on the foundation laid in Year 1, we will delve into the generation of diverse and high-quality content using advanced AI techniques. This will involve the creation of algorithms and models capable of producing 3D content across various domains. Our goal is to surpass the capabilities of existing models like ShapeGPT, and initial demonstrations showcasing precise generation and modification at the part level. Additionally, we aspire to achieve highly controllable content generation results while optimizing computational costs. We plan to submit one to two key research findings to international conferences such as SIGGRAPH, AAAI, or ICCV.

Year 3: Scene Reconstruction. The third year will focus on scene reconstruction. Integrate features from different data sources into a unified model. Scene construction: Based on the extracted features, establish scene reconstruction models, including site models, building models, vegetation models, etc. Model training: Use machine learning, deep learning and other technologies to train models to improve the accuracy and efficiency of scene reconstruction. Model optimization: Perform parameter adjustments, structural optimization, etc. on the model to further improve the reconstruction effect.

Year 4: Integration and Comprehensive Model Development. The fourth year will focus on synthesizing the results of the preceding years to formulate a comprehensive model. This model will be capable of synergistic multimodal scene understanding and controllable high-quality 3D content generation. We aim to integrate industrial standards to align our research with practical applications, achieving genuinely impactful results.

Referencias

[1] T. Brown *et al.*, Language models are few-shot learners, *Advances in neural information processing systems*, *33*, 1877–1901, **2020**.

[2] C. Wang *et al.*, Neural codec language models are zero-shot text to speech synthesizers, *arXiv preprint arXiv:2301.02111*, **2023**.

[3] N. Fei *et al.*, Towards artificial general intelligence via a multimodal foundation model, *Nature Communications*, *13*, no. 1, 3094, **2022**.

[4] A. Koubaa, Gpt-4 vs. gpt-3.5: A concise showdown, **2023**.

[5] A. Ramesh *et al.*, Zero-shot text-to-image generation, in *International Conference on Machine Learning*, PMLR, **2021**, 8821–8831.

[6] J. Duan, S. Yu, H. L. Tan, H. Zhu, and C. Tan, A survey of embodied ai: From simulators to research tasks, *IEEE Transactions*

on Emerging Topics in Computational Intelligence, *6*, no. 2, 230–244, **2022**.

[7] A. Abdelreheem, U. Upadhyay, I. Skorokhodov, R. Al Yahya, J. Chen, and M. Elhoseiny, 3dreftransformer: Fine-grained object identification in real-world scenes using natural language, in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, **2022**, 3941–3950.

[8] A. Abdelreheem, K. Olszewski, H.-Y. Lee, P. Wonka, and P. Achlioptas, Scanents3d: Exploiting phrase-to-3d-object correspondences for improved visio-linguistic models in 3d scenes, *arXiv preprint arXiv:2212.06250*, **2022**.

[9] D. Azuma, T. Miyanishi, S. Kurita, and M. Kawanabe, Scanqa: 3d question answering for spatial scene understanding, in *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, **2022**, 19129–19139.

[10] X. Ma *et al.*, Sqa3d: Situated question answering in 3d scenes, *arXiv preprint arXiv:2210.07474*, **2022**.

[11] J. Luo *et al.*, 3d-sps: Single-stage 3d visual grounding via referred point progressive selection, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, **2022**, 16454–16463.

[12] A. Jain, N. Gkanatsios, I. Mediratta, and K. Fragkiadaki, Bottom up top down detection transformers for language grounding in images and point clouds, in *European Conference on Computer Vision*, Springer, **2022**, 417–433.

[13] D. Cai, L. Zhao, J. Zhang, L. Sheng, and D. Xu, 3djcg: A unified framework for joint dense captioning and visual grounding on 3d point clouds, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, **2022**, 16464–16473.

[14] S. Huang, Y. Chen, J. Jia, and L. Wang, Multi-view transformer for 3d visual grounding, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, **2022**, 15524–15533.

[15] S. Chen, P.-L. Guhur, M. Tapaswi, C. Schmid, and I. Laptev, Language conditioned spatial relation reasoning for 3d object grounding, *Advances in Neural Information Processing Systems*, *35*, 20522–20535, **2022**.

[16] J. Davis, S. R. Marschner, M. Garr, and M. Levoy, Filling holes in complex surfaces using volumetric diffusion, in *Proceedings. First international symposium on 3d data processing visualization and transmission*, IEEE, **2002**, 428–441.

[17] M. Kazhdan, M. Bolitho, and H. Hoppe, Poisson surface reconstruction, in *Proceedings of the fourth Eurographics symposium on Geometry processing*, *7*, **2006**, 0.

[18] A. Nealen, T. Igarashi, O. Sorkine, and M. Alexa, Laplacian mesh optimization, in *Proceedings of the 4th international conference on Computer graphics and interactive techniques in Australasia and Southeast Asia*, **2006**, 381–389.

[19] O. Sorkine and D. Cohen-Or, Least-squares meshes, in *Proceedings Shape Modeling Applications, 2004.*, IEEE, **2004**, 191–199.

[20] R. de Charette and S. Manitsaris, 3d reconstruction of deformable revolving object under heavy hand interaction, *arXiv preprint arXiv:1908.01523*, **2019**.

[21] M. Pauly, N. J. Mitra, J. Wallner, H. Pottmann, and L. J. Guibas, Discovering structural regularity in 3d geometry, in *ACM SIGGRAPH 2008 papers*, **2008**, 1–11.

[22] I. Sipiran, R. Gregor, and T. Schreck, Approximate symmetry detection in partial 3d meshes, in *Computer Graphics Forum*, Wiley Online Library, *33*, **2014**, 131–140.

[23] N. J. Mitra, M. Pauly, M. Wand, and D. Ceylan, Symmetry in 3d geometry: Extraction and applications, in *Computer Graphics Forum*, Wiley Online Library, *32*, **2013**, 1–23.

[24] D. Li, T. Shao, H. Wu, and K. Zhou, Shape completion from a single rgbd image, *IEEE transactions on visualization and computer graphics*, *23*, no. 7, 1809–1822, **2016**.

[25] Y. Li, A. Dai, L. Guibas, and M. Nießner, Database-assisted object retrieval for real-time 3d reconstruction, in *Computer graph-*

ics forum, Wiley Online Library, *34*, **2015**, 435–446.

[26] A. Dai, C. Ruizhongtai Qi, and M. Nießner, Shape completion using 3d-encoder-predictor cnns and shape synthesis, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, **2017**, 5868–5877.

[27] X. Han, Z. Li, H. Huang, E. Kalogerakis, and Y. Yu, High-resolution shape completion using deep neural networks for global structure and local geometry inference, in *Proceedings of the IEEE international conference on computer vision*, **2017**, 85–93.

[28] E. J. Smith and D. Meger, Improved adversarial systems for 3d object generation and reconstruction, in *Conference on Robot Learning*, PMLR, **2017**, 87–96.

[29] A. Nguyen and B. Le, 3d point cloud segmentation: A survey, in *2013 6th IEEE conference on robotics, automation and mechatronics (RAM)*, IEEE, **2013**, 225–230.

[30] Y. Xie, J. Tian, and X. X. Zhu, Linking points with labels in 3d: A review of point cloud semantic segmentation, *IEEE Geoscience and remote sensing magazine*, *8*, no. 4, 38–59, **2020**.

[31] A. Boulch, J. Guerry, B. Le Saux, and N. Audebert, Snapnet: 3d point cloud semantic labeling with 2d deep segmentation networks, *Computers & Graphics*, *71*, 189–198, **2018**.

[32] A. Dai and M. Nießner, 3dmv: Joint 3d-multi-view prediction for 3d semantic scene segmentation, in *Proceedings of the European Conference on Computer Vision (ECCV)*, **2018**, 452–468.

[33] L. Landrieu and M. Simonovsky, Large-scale point cloud semantic segmentation with superpoint graphs, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, **2018**, 4558–4567.

[34] M. Zhong and G. Zeng, Semantic point completion network for 3d semantic scene completion, in *ECAI 2020*, IOS Press, **2020**, 2824–2831.

[35] A. Avetisyan, M. Dahnert, A. Dai, M. Savva, A. X. Chang, and M. Nießner, Scan2cad: Learning cad model alignment in rgb-d scans, in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, **2019**, 2614–2623.

[36] L. Nan, K. Xie, and A. Sharf, A search-classify approach for cluttered indoor scene understanding, *ACM Transactions on Graphics (TOG)*, *31*, no. 6, 1–10, **2012**.

[37] J. Hou, A. Dai, and M. Nießner, Revealnet: Seeing behind objects in rgb-d scans, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, **2020**, 2098–2107.

[38] A. Avetisyan, T. Khanova, C. Choy, D. Dash, A. Dai, and M. Nießner, Scenecad: Predicting object alignments and layouts in rgb-d scans, in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, Springer, **2020**, 596–612.

[39] S. Shi, Z. Wang, J. Shi, X. Wang, and H. Li, From points to parts: 3d object detection from point cloud with part-aware and part-aggregation network, *IEEE transactions on pattern analysis and machine intelligence*, *43*, no. 8, 2647–2664, **2020**.

[40] S. Song and J. Xiao, Deep sliding shapes for amodal 3d object detection in rgb-d images, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, **2016**, 808–816.

[41] G. Gkioxari, J. Malik, and J. Johnson, Mesh r-cnn, in *Proceedings of the IEEE/CVF international conference on computer vision*, **2019**, 9785–9795.

[42] Y. Nie, X. Han, S. Guo, Y. Zheng, J. Chang, and J. J. Zhang, Total3dunderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, **2020**, 55–64.