A Survey on 3D Deep Networks

Mei Han
meih3@illinois.edu

Wenqi Jia
wenqij5@illinois.edu

1. Introduction

3D deep network has made significant strides in computer vision in recent years, offering powerful tools for tasks such as object recognition, shape analysis, and scene segmentation. This brief survey provides an overview of key design decisions and techniques in 3D deep networks. We first discuss various 3D data representations, including multi-view, RGB-D, point cloud, voxel, mesh, and implicit representations, each optimized for specific tasks. We then explore network architecture designs, covering 3D Convolutional Neural Networks, Transformers, diffusion models, Graph Neural Networks, and hybrid architectures. We also examine various learning strategies, such as supervised, self-supervised, and transfer learning, to address the complexities of 3D data. Finally, we consider task-specific design considerations, including 3D object classification, segmentation, reconstruction, and detection. Through this concise overview, we aim to provide readers with a comprehensive perspective on the field of 3D deep learning, highlighting its current capabilities and future research directions.

2. Key Design Decisions and Techniques

2.1. 3D Data Representation

In recent years, 3D deep learning has been utilizing various representations tailored for tasks like object recognition, shape analysis, and segmentation.

Multi-view Representations represent 3D objects as a collection of 2D images captured from different viewpoints. Multi-view CNN (MVCNN) [39] achieves competitive performance on shape recognition tasks by fusing features across views. RotationNet [16] improves upon this by learning object rotation jointly with recognition, enhancing robustness to viewpoint variations.

RGB-D-based Representations, like 3D ShapeNets [46], use depth maps to capture geometric structures of 3D objects in conjunction with RGB images. Field Probing Neural Networks (FPNN) [20] introduces a field probing technique to efficiently process depth images for 3D object detection, further optimizing the use of 2.5D data.

Point Cloud-based Representations, led by PointNet [31], process raw point clouds directly, leveraging their unordered nature to simplify the learning of spatial features. PointNet++ [32] extends this by incorporating hierarchical structures, enabling local feature learning at multiple scales to deal with those non-uniform densities.

Voxel-based Representations represent 3D shapes as volumetric grids, making them compatible with 3D convolutional networks [26]. To address the high memory and computational cost of voxel grids, OctNet [34] introduces a hierarchical octree-based representation, reducing resource demands while maintaining performance for large-scale 3D data processing.

Mesh-based Representations treat 3D data as surfaces represented by vertices and edges, applying graph convolutions to the mesh structure for shape analysis tasks. FeaStNet [43] and MeshCNN [13] introduce edge-based convolutional operations to capture surface details more effectively, making them suitable for tasks like 3D mesh segmentation.

Implicit Representations model 3D shapes using Signed Distance Functions (SDFs), which represent surfaces continuously, enabling highly detailed shape representations. DeepSDF [29] and Occupancy Networks [27] generalize this concept by learning to predict whether a point is inside or outside a shape, offering a flexible and continuous representation of 3D objects.

2.2. Network Architecture Design

In 3D deep learning, various network architectures have been explored to handle the unique challenges posed by different types of 3D data. Each architecture offers specific strengths, depending on the representation and the task. In this section, we discuss the major architectural paradigms used in 3D deep learning, along with examples of relevant works.

3D Convolutional Neural Networks (CNNs) naturally extend 2D CNNs by applying convolutional operations in three dimensions and have been foundational in 3D deep learning, particularly for voxel-based and multi-view representations. VoxNet [26] employs 3D CNNs to recognize objects from voxelized input, demonstrating the effectiveness of convolutional architectures for spatial feature extraction. Similarly, OctNet [34] builds on voxel-based CNNs by introducing octree structures to reduce memory requirements

while preserving essential spatial resolution. In contrast, MVCNN [39] applies 2D CNNs to multi-view representations of 3D objects and aggregates information from different views, achieving strong performance in 3D shape recognition tasks. Another multi-view approach, RotationNet [16], learns object rotation along with recognition, utilizing CNNs to improve robustness to varying viewpoints.

Transformers have become popular in the field of 3D deep learning due to their capacity to model long-range dependencies in irregular 3D data as point clouds. For example, Point Transformer [53] processes point clouds by leveraging self-attention to model interactions between points, leading to better performance in object classification and part segmentation tasks. Similarly, Point Cloud Transformer (PCT) [12], also uses a transformer-based architecture to capture local features in point clouds before applying global self-attention, achieving state-of-the-art performance in 3D point cloud analysis.

Diffusion models have been widely used in 3D learning tasks such as shape generation and completion. These models operate by introducing noise to the data, and the core objective is to progressively reverse this diffusion process, refining the noisy input step by step until a clean, denoised sample is generated. 3D Shape Diffusion [54] applies diffusion models to generate 3D objects from noise, offering a novel way to generate high-resolution shapes. DiffusionNet [36] using learned diffusion mechanisms to smooth and refine mesh and point cloud data, demonstrating significant improvements in reconstruction tasks.

Graph Neural Networks (GNNs) have been proven to be particularly effective for mesh-based and point cloud-based 3D data that can be represented as graphs, given the nature that the graph structure allows the network to capture both local and global features by propagating information along edges. Dynamic Graph CNN (DGCNN) [44] applies GNNs to dynamically compute graphs from point clouds, capturing relationships between neighboring points for tasks like segmentation and classification. Similarly, MeshCNN [13] introduces convolutional operations over mesh edges, which can effectively capture surface details for tasks such as 3D mesh segmentation.

Hybrid Architectures are emerging as an effective way to combine the strengths of different architectures to improve performance across various tasks. For example, 3D-R2N2 [5] integrates 3D CNNs with Recurrent Neural Networks (RNNs) to enable temporal feature learning to reconstruct 3D objects from a sequence of 2D images. Similarly, PointRCNN [37] combines the benefits of point cloud processing with region-based CNNs for effective 3D object detection.

2.3. Learning Strategy

Various learning strategies have been adopted to handle the challenges such as irregularity, sparsity, and high computational costs posed by complex 3D data.

Supervised Learning is the most common strategy in 3D deep learning, where models are trained on labeled 3D datasets. Methods such as PointNet and VoxNet are examples where deep networks are trained with labeled point clouds and voxel grids, respectively. Supervised learning is effective for tasks like 3D object classification, part segmentation, and scene understanding, however, large-scale labeled data might be expensive to obtain.

Self-Supervised Learning given the limited availability of labeled 3D data, self-supervised learning has been explored with unlabeled 3D data. Models learn useful features by solving auxiliary tasks, such as predicting the missing part of an object or reconstructing a 3D shape from partial inputs. For example, PointContrast [48] employs contrastive learning to enhance feature extraction from point clouds without explicit supervision.

Semi-Supervised Learning combines both labeled and unlabeled data to improve performance in situations where labeling is expensive or impractical. SPLATNet [40] is a semi-supervised model that learns from both labeled and unlabeled point clouds, using the latter to improve generalization on unseen data. The combination of small labeled datasets with large amounts of unlabeled 3D data enhances the network's robustness in real-world applications.

Few-Shot Learning aims to train models with minimal labeled data, leveraging prior knowledge to recognize new categories with few examples. GNN-based few-shot learning methods have been proposed for 3D shape recognition, where the network generalizes to unseen categories by learning relational features between 3D objects.

Multi-Task Learning trains a model to perform several related tasks simultaneously to leverage shared features across tasks. For example, PointNet++ can be extended for multi-task learning by adding branches for 3D classification and part segmentation in parallel.

2.4. Task-specific Considerations

When applying deep learning models to 3D data, it is crucial to account for the unique challenges and requirements of each specific task. The design of 3D networks is often influenced by the nature of the task, which may include object classification, segmentation, reconstruction, and detection.

3D Object Classification In object classification, the primary challenge is to effectively capture global shape information from 3D data. Methods like PointNet and VoxNet focus on processing point clouds and voxel grids, respectively, to learn features that differentiate between object categories. The representation of 3D data significantly impacts performance; for example, multi-view approaches like

MVCNN aggregate information from 2D projections of 3D objects to improve recognition accuracy. In this task, the choice of representation—whether it be point clouds, voxels, or multi-view images—plays a central role in the model's ability to capture the necessary geometric details.

3D Segmentation 3D segmentation involves predicting a label for each point or voxel in the 3D object or scene, making it essential for models to capture both local and global geometric features. Networks such as PointNet++ and DGCNN are designed to capture hierarchical structures and neighborhood relationships, allowing them to perform fine-grained segmentation tasks. In mesh-based segmentation, methods like MeshCNN operate on the mesh structure to extract edge-level details. The ability to capture fine local features, while maintaining context from the entire object or scene, is crucial for accurate segmentation.

3D Shape Reconstruction Shape reconstruction tasks, such as those tackled by voxel-based networks or implicit surface models like Occupancy Networks, require the ability to generate complete 3D shapes from partial or noisy input. Implicit representations, such as Signed Distance Functions (SDF) used in DeepSDF, have become popular for these tasks due to their continuous and resolution-independent nature. The challenge here lies in generating high-quality, smooth reconstructions that are faithful to the input, whether from point clouds, partial views, or other modalities.

3D Object Detection Object detection in 3D spaces presents challenges due to the sparsity and irregularity of 3D data, particularly in real-world environments such as LiDAR scans. Networks like VoteNet propose voting-based mechanisms to detect 3D bounding boxes around objects in point clouds. Another approach, PointRCNN, builds on the region proposal network (RPN) architecture from 2D object detection but adapts it for point clouds. For detection tasks, the ability to accurately localize objects and estimate their orientation in 3D space is paramount, and models must effectively handle sparse and unstructured data.

3D Motion and Scene Flow Estimation In dynamic environments, estimating motion or scene flow from 3D data is essential for applications like autonomous driving and robotics. Approaches such as FlowNet3D extend point cloud processing to estimate scene flow, capturing point-wise motion in 3D. These tasks require networks to understand temporal evolution and local motion patterns, and they often combine sequential data from multiple frames to predict movement in 3D space. The ability to model temporal dependencies is a key consideration in this domain, leading to the use of architectures like recurrent neural networks (RNNs) or temporal convolutions.

3. Evaluation Method

3.1. Benchmarks and Datasets

To evaluate the performance of 3D deep networks across various tasks including object classification, segmentation, reconstruction, and detection, several key benchmarks and datasets have emerged. These datasets provide rich and diverse sources of 3D data, including point clouds, meshes, voxels, and multi-view images, allowing researchers to test their models on both real-world and synthetic 3D objects.

3.1.1 3D Object Shape

ModelNet [46] is one of the most widely used benchmarks for 3D object classification and retrieval tasks. It consists of two main subsets: **ModelNet10** (with 10 object categories) and **ModelNet40** (with 40 object categories). The dataset contains over 12,000 3D CAD models, providing a standard for evaluating 3D shape recognition methods. Many early deep learning approaches, such as PointNet and VoxNet, were first evaluated on ModelNet. **ShapeNet** [4] is a large-scale repository of 3D object models, containing several categories. The **ShapeNetCore** subset contains over 50,000 models from 55 object categories, making it a highly popular benchmark for 3D object classification, segmentation, and retrieval tasks. The **ShapeNet Parts** subset is specifically designed for part segmentation tasks, with point-wise annotations indicating different parts of objects in 16 categories. **PartNet** [28] focuses on fine-grained 3D part segmentation and is built on ShapeNet models. It contains more than 26,000 3D models from 24 object categories, with detailed part-level annotations for each object. This dataset enables fine-grained part segmentation tasks and detailed shape analysis. **Pix3D** [41] is a dataset designed for 3D shape reconstruction from single images. It contains over 10,000 real-world images of furniture along with corresponding 3D models, making it a valuable resource for image-to-3D shape reconstruction tasks.

3.1.2 Real-World Indoor Scenes

ScanNet [7] is a richly annotated dataset of 3D meshes and point clouds derived from real-world indoor scenes. It contains over 1,500 scenes captured by RGB-D sensors, with dense 3D reconstructions and semantic labels for object detection and segmentation tasks. ScanNet is widely used in semantic segmentation and 3D scene understanding benchmarks due to its comprehensive labeling of real-world environments. **S3DIS (Stanford Large-Scale 3D Indoor Spaces)** [2] is a benchmark dataset focusing on 3D semantic segmentation of indoor spaces. The dataset consists of several areas from the Stanford campus, with dense point clouds and detailed annotations of objects including furniture, walls, and floors. It is commonly used in

tasks like room layout prediction, object segmentation, and scene classification. **SUN RGB-D** [38] is another large-scale dataset for 3D object detection and scene understanding in indoor environments. It contains over 10,000 RGB-D images annotated with 3D bounding boxes for a wide variety of object categories. SUN RGB-D is a common benchmark for depth-based 3D object detection models.

3.1.3 Real-World Outdoor Scenes

KITTI [10] dataset is widely used for autonomous driving research and contains annotated 3D point clouds, RGB images, and stereo data. It provides benchmarks for tasks like 3D object detection, scene flow estimation, and visual odometry, making it highly relevant for 3D learning models applied to outdoor, real-world scenarios. **NuScenes** [3] is a large-scale autonomous driving dataset that includes 3D point clouds, camera images, and radar data. It provides annotations for 3D object detection, tracking, and segmentation, making it an essential benchmark for 3D deep networks in the field of autonomous driving.

3.2. Metrics

Evaluating the performance of 3D deep learning models requires a variety of metrics tailored to different tasks like classification, segmentation, reconstruction, and detection.

Accuracy is a widely used metric for 3D object classification and segmentation tasks. It represents the percentage of correctly classified or labeled points, objects, or parts within a dataset. For multi-class tasks, accuracy is calculated as the ratio of correctly predicted classes to the total number of classes.

Precision and Recall are commonly used in 3D object detection and segmentation tasks. Precision measures the proportion of correctly predicted positive samples among all predicted positives, while recall measures the proportion of correctly predicted positive samples among all actual positives. These metrics are especially important when dealing with imbalanced datasets or rare object classes.

F1 Score is the harmonic mean of precision and recall, providing a single metric that balances the trade-off between the two. It is particularly useful for evaluating 3D object detection models where both false positives and false negatives are critical to performance.

Average Precision (AP) is a standard metric for 3D object detection, particularly in benchmarks like KITTI [10] and NuScenes [3]. It measures the precision-recall curve, calculating the area under the curve. Higher AP values reflect better detection performance across different levels of confidence.

Mean Average Precision (mAP) is an extension of AP, typically used when evaluating multi-class object detection tasks. It calculates the average precision across all object categories, providing a single metric for comparing detection models' performance across different datasets.

Intersection over Union (IoU) is a critical metric for 3D segmentation and object detection tasks. It measures the overlap between the predicted segmentation or bounding box and the ground truth. Specifically, IoU is calculated as the area of the intersection divided by the area of the union of the predicted and ground truth regions. Higher IoU values indicate better segmentation or detection performance.

Mean IoU (mIoU) IoU is a more general form of IoU, used in segmentation tasks involving multiple classes. It calculates the IoU for each class and then computes the mean value, providing a more holistic view of the segmentation performance across different categories.

Volumetric IoU (vIoU) is a specialized metric for evaluating voxel-based 3D object detection or reconstruction. Similar to standard IoU, it measures the overlap between the predicted and ground truth volumes. It is particularly important for tasks involving voxel grids, where accurate prediction of 3D occupancy is crucial.

Chamfer Distance is commonly used for 3D shape reconstruction and point cloud completion tasks. It measures the average distance between each point in the predicted shape and the nearest point in the ground truth shape, and vice versa. This metric assesses how well the predicted shape matches the ground truth in terms of point correspondence.

Earth Mover's Distance (EMD) is another metric used for evaluating shape reconstruction and point cloud generation. It measures the minimum cost required to transform one point cloud into another by solving a transportation problem. EMD is particularly effective for assessing the similarity between two sets of points, with lower values indicating better alignment between the predicted and ground truth shapes.

4. Current Capabilities and Gaps

4.1. Representation and Scene Complexity

Diversity in Data Representation poses a significant challenge in 3D deep learning [1]. 3D data can be represented through various formats, each with distinct advantages and limitations. Meshes provide a compact representation but struggle with complex topologies. Voxels offer a regular structure conducive to 3D convolutions, yet they suffer from high memory consumption, particularly at high resolutions. Point clouds, easily acquired from sensors like LiDAR, lack explicit topological information. Implicit functions, such as signed distance fields, offer continuous representation of 3D surfaces but are computationally intensive. Multi-view images leverage existing 2D vision techniques but necessitate additional processing for 3D structure inference. This lack of a unified representation hinders model generalization across diverse datasets and tasks [46].

Scale and Sparsity Disparities across different scenes present another crucial challenge for 3D deep networks [24, 32]. Indoor environments typically feature densely packed, detailed objects, whereas outdoor scenes are characterized by sparse, large-scale structures with inconsistent point densities. These stark contrasts complicate the development of models capable of effectively managing both fine-grained indoor representations and expansive outdoor scenes

4.2. Challenges in Data Acquisition and Processing

Real-world 3D data is frequently affected by occlusions, motion blur, and illumination variations, compromising data quality.

Motion Blur in dynamic scenes featuring camera motion or rapidly moving objects introduces inter-frame inconsistencies [6, 17, 23, 30, 50]. This phenomenon impedes accurate feature extraction, compromises depth estimation, and results in distorted 3D reconstructions.

Illumination Variation due to diverse lighting conditions, including shadows, specular reflections, and global illumination changes, significantly affects the perception and reconstruction of 3D scenes [21, 52]. These variations complicate surface texture recognition and compromise the accuracy of 3D shape inference. The interplay between illumination and geometry introduces ambiguities in shape estimation, particularly challenging for photometric stereo and multi-view reconstruction techniques.

Occlusions are prevalent in complex scenes, presenting a fundamental challenge in 3D vision [44]. Partial or complete object occlusions result in incomplete 3D reconstructions, erroneous object segmentation, and misidentification, particularly impacting instance-level 3D understanding [14]. The absence of information in occluded regions necessitates sophisticated inference and completion algorithms to reconstruct the full 3D structure.

Sensor Noise and Artifacts from various 3D sensing technologies, such as LiDAR and depth cameras, introduce specific noise patterns and imperfections [42]. These issues propagate through the processing pipeline, affecting the accuracy of 3D representations and subsequent analyses [33]. The characteristics of sensor noise vary across different modalities, requiring tailored denoising and artifact removal techniques for each sensor type.

5. New Research Ideas

5.1. Physics-Guided Deep Learning

Physics-guided deep learning integrates physical principles into neural networks, enhancing 3D vision accuracy and interpretability [45]. This approach bridges data-driven methods and physical understanding, enabling more robust and realistic 3D modeling.

Physically Consistent 3D Reconstruction involves incorporating laws of optics, material properties, and behavior into neural rendering pipelines and 3D deep learning models. This approach leads to more accurate and physically plausible 3D reconstructions [52], improving tasks such as deformable object reconstruction and interaction prediction [22]. Models constrained by principles of light transport ensure that reconstructed surfaces interact with light in physically correct ways, enhancing the realism of generated 3D environments and leading to better understanding of complex 3D scenes with diverse material properties.

Physics-Informed Point Cloud Processing leverages physical constraints to improve the robustness of object detection, scene understanding [47], and trajectory prediction [35]. By enforcing physical priors such as gravity, object stability, and principles from classical mechanics, these models generate more realistic and consistent 3D scene interpretations and motion predictions. This approach is particularly crucial for applications like autonomous driving and robotics, where accurate prediction of object motion significantly improves safety and performance, especially in complex environments where traditional methods might struggle.

Energy-Conserving Neural Networks represent an innovative direction where network architectures are designed to inherently conserve energy or other physical quantities. This approach results in more stable and physically consistent 3D predictions [11], which is particularly relevant for long-term predictions in dynamic 3D scenes, enhancing the reliability of simulation outcomes. These networks provide a foundation for more accurate and physically grounded 3D vision systems, potentially revolutionizing fields ranging from environmental modeling to virtual reality applications.

5.2. Interactable 3D Generation

Articulated Object Modeling is gradually occupying an important position in 3D vision [8, 15].Modeling and tracking articulated objects with moving parts involves detecting and reconstructing individual parts of objects and tracking their poses across a sequence of observations. Traditional approaches often relied on point cloud data or multi-view observations. These methods range from analyzing sequences of articulated motion to processing single point clouds. However, the complexity of articulated objects has pushed researchers to explore more advanced techniques. RSRD [17] combines Segment Anything (SAM) [19] with GARField [18] for 3D part understanding. PhysPart [25] introduces a diffusion-based part generation model with physical constraints. Interactive3D [9] presents a two-stage framework enhancing controllability and quality in 3D object generation.

Iteractable 3D Scenes PhyScene [49] responds to the

growing demand in Embodied Artificial Intelligence (EAI) for physically plausible and interactive 3D scenes. It uses a conditional diffusion model with physics- and interactivity-based guidance mechanisms to generate scenes with realistic layouts, articulated objects, and rich physical interactivity tailored for embodied agents. ClimateNeRF [21] demonstrates the fusion of physical simulations with Neural Radiance Fields (NeRF) to visualize dynamic weather effects in 3D scenes. This allows for realistic representations of climate change impacts, controlled by physically meaningful variables. For Human-Environment Interaction, recent studies [51] have incorporated physics-aware constraints to model interactions between realistic human bodies and the environment, reducing collisions and enhancing realism through accurate body-scene contact.

6. Conclusion

In this survey, we have explored the key advancements in 3D deep learning technologies, focusing on Data Representation, Network Architectures, Learning Strategies, and Task-specific Considerations. These technologies have enabled significant progress in various 3D vision tasks, improving the ability to perceive, understand, and interact with complex 3D environments. Recent innovations like implicit representations and physics-guided deep learning have significantly enhanced 3D reconstruction and analysis capabilities. Looking ahead, the integration of interactable 3D generation with 3D deep learning presents an exciting opportunity to further enhance 3D vision systems. By combining advanced 3D modeling techniques with physics-based constraints, we could achieve more realistic and interactive 3D representations, process complex scenes, and even collaborate more effectively in dynamic and unstructured environments.

References

[1] Eman Ahmed, Alexandre Saint, Abd El Rahman Shabayek, Kseniya Cherenkova, Rig Das, Gleb Gusev, Djamila Aouada, and Bjorn Ottersten. A survey on deep learning advances on different 3d data representations. *arXiv preprint arXiv:1808.01462*, 2018. 4

[2] Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1534–1543, 2016. 3

[3] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 4

[4] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015. 3

[5] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 628–644. Springer, 2016. 2

[6] Yaodong Cui, Ren Chen, Wenbo Chu, Long Chen, Daxin Tian, Ying Li, and Dongpu Cao. Deep learning for image and point cloud fusion in autonomous driving: A review. *IEEE Transactions on Intelligent Transportation Systems*, 23(2): 722–739, 2021. 5

[7] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 3

[8] Congyue Deng, Jiahui Lei, William B Shen, Kostas Daniilidis, and Leonidas J Guibas. Banana: Banach fixed-point network for pointcloud segmentation with inter-part equivariance. *Advances in Neural Information Processing Systems*, 36, 2024. 5

[9] Shaocong Dong, Lihe Ding, Zhanpeng Huang, Zibin Wang, Tianfan Xue, and Dan Xu. Interactive3d: Create what you want by interactive 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4999–5008, 2024. 5

[10] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 4

[11] Samuel Greydanus, Misko Dzamba, and Jason Yosinski. Hamiltonian neural networks. *Advances in neural information processing systems*, 32, 2019. 5

[12] Meng-Hao Guo, Jun-Xiong Cai, Zheng-Ning Liu, Tai-Jiang Mu, Ralph R Martin, and Shi-Min Hu. Pct: Point cloud transformer. *Computational Visual Media*, 7:187–199, 2021. 2

[13] Rana Hanocka, Amir Hertz, Noa Fish, Raja Giryes, Shachar Fleishman, and Daniel Cohen-Or. Meshcnn: a network with an edge. *ACM Transactions on Graphics (ToG)*, 38(4):1–12, 2019. 1, 2

[14] Ji Hou, Angela Dai, and Matthias Nießner. Revealnet: Seeing behind objects in rgb-d scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2098–2107, 2020. 5

[15] Zhenyu Jiang, Cheng-Chun Hsu, and Yuke Zhu. Ditto: Building digital twins of articulated objects from interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5616–5626, 2022. 5

[16] Asako Kanezaki, Yasuyuki Matsushita, and Yoshifumi Nishida. Rotationnet: Joint object categorization and pose estimation using multiviews from unsupervised viewpoints.

In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5010–5019, 2018. 1, 2

[17] Justin Kerr, Chung Min Kim, Mingxuan Wu, Brent Yi, Qianqian Wang, Angjoo Kanazawa, and Ken Goldberg. Robot see robot do: Part-centric feature fields for visual imitation of articulated objects. In *8th Annual Conference on Robot Learning*. 5

[18] Chung Min Kim, Mingxuan Wu, Justin Kerr, Ken Goldberg, Matthew Tancik, and Angjoo Kanazawa. Garfield: Group anything with radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21530–21539, 2024. 5

[19] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 5

[20] Yangyan Li, Soeren Pirk, Hao Su, Charles R Qi, and Leonidas J Guibas. Fpnn: Field probing neural networks for 3d data. *Advances in neural information processing systems*, 29, 2016. 1

[21] Yuan Li, Zhi-Hao Lin, David Forsyth, Jia-Bin Huang, and Shenlong Wang. Climatenerf: Extreme weather synthesis in neural radiance field. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3227–3238, 2023. 5, 6

[22] Hubert Lin, Melinos Averkiou, Evangelos Kalogerakis, Balazs Kovacs, Siddhant Ranade, Vladimir Kim, Siddhartha Chaudhuri, and Kavita Bala. Learning material-aware local descriptors for 3d shapes. In *2018 International Conference on 3D Vision (3DV)*, pages 150–159. IEEE, 2018. 5

[23] Zhi-Hao Lin, Wei-Chiu Ma, Hao-Yu Hsu, Yu-Chiang Frank Wang, and Shenlong Wang. Neurmips: Neural mixture of planar experts for view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15702–15712, 2022. 5

[24] Ze Liu, Han Hu, Yue Cao, Zheng Zhang, and Xin Tong. A closer look at local aggregation operators in point cloud analysis. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*, pages 326–342. Springer, 2020. 5

[25] Rundong Luo, Haoran Geng, Congyue Deng, Puhao Li, Zan Wang, Baoxiong Jia, Leonidas Guibas, and Siyuang Huang. Physpart: Physically plausible part completion for interactable objects. *arXiv preprint arXiv:2408.13724*, 2024. 5

[26] Daniel Maturana and Sebastian Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 922–928. IEEE, 2015. 1

[27] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4460–4470, 2019. 1

[28] Kaichun Mo, Shilin Zhu, Angel X Chang, Li Yi, Subarna Tripathi, Leonidas J Guibas, and Hao Su. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 909–918, 2019. 3

[29] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 1

[30] Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild. *arXiv preprint arXiv:2409.12259*, 2024. 5

[31] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas 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*, pages 652–660, 2017. 1

[32] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 1, 5

[33] Jiaxiong Qiu, Zhaopeng Cui, Yinda Zhang, Xingdi Zhang, Shuaicheng Liu, Bing Zeng, and Marc Pollefeys. Deeplidar: Deep surface normal guided depth prediction for outdoor scene from sparse lidar data and single color image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3313–3322, 2019. 5

[34] Gernot Riegler, Ali Osman Ulusoy, and Andreas Geiger. Octnet: Learning deep 3d representations at high resolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3577–3586, 2017. 1

[35] Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter Battaglia. Learning to simulate complex physics with graph networks. In *International conference on machine learning*, pages 8459–8468. PMLR, 2020. 5

[36] Nicholas Sharp, Souhaib Attaiki, Keenan Crane, and Maks Ovsjanikov. Diffusionnet: Discretization agnostic learning on surfaces. *ACM Transactions on Graphics (TOG)*, 41(3): 1–16, 2022. 2

[37] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 770–779, 2019. 2

[38] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015. 4

[39] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 945–953, 2015. 1, 2

[40] Hang Su, Varun Jampani, Deqing Sun, Subhransu Maji, Evangelos Kalogerakis, Ming-Hsuan Yang, and Jan Kautz. Splatnet: Sparse lattice networks for point cloud processing.

In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2530–2539, 2018. 2

[41] Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B Tenenbaum, and William T Freeman. Pix3d: Dataset and methods for single-image 3d shape modeling. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[42] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *2017 international conference on 3D Vision (3DV)*, pages 11–20. IEEE, 2017. 5

[43] Nitika Verma, Edmond Boyer, and Jakob Verbeek. Feastnet: Feature-steered graph convolutions for 3d shape analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2598–2606, 2018. 1

[44] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (tog)*, 38(5):1–12, 2019. 2, 5

[45] Jared Willard, Xiaowei Jia, Shaoming Xu, Michael Steinbach, and Vipin Kumar. Integrating physics-based modeling with machine learning: A survey. *arXiv preprint arXiv:2003.04919*, 1(1):1–34, 2020. 5

[46] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015. 1, 3, 4

[47] Kevin Xie, Tingwu Wang, Umar Iqbal, Yunrong Guo, Sanja Fidler, and Florian Shkurti. Physics-based human motion estimation and synthesis from videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11532–11541, 2021. 5

[48] Saining Xie, Jiatao Gu, Demi Guo, Charles R Qi, Leonidas Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 574–591. Springer, 2020. 2

[49] Yandan Yang, Baoxiong Jia, Peiyuan Zhi, and Siyuan Huang. Physcene: Physically interactable 3d scene synthesis for embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16262–16272, 2024. 5

[50] Hongguang Zhang, Yuchao Dai, Hongdong Li, and Piotr Koniusz. Deep stacked hierarchical multi-patch network for image deblurring. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5978–5986, 2019. 5

[51] Siwei Zhang, Yan Zhang, Qianli Ma, Michael J Black, and Siyu Tang. Place: Proximity learning of articulation and contact in 3d environments. In *2020 International Conference on 3D Vision (3DV)*, pages 642–651. IEEE, 2020. 6

[52] Xiuming Zhang, Pratul P Srinivasan, Boyang Deng, Paul Debevec, William T Freeman, and Jonathan T Barron. Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. *ACM Transactions on Graphics (ToG)*, 40(6):1–18, 2021. 5

[53] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16259–16268, 2021. 2

[54] Linqi Zhou, Yilun Du, and Jiajun Wu. 3d shape generation and completion through point-voxel diffusion. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5826–5835, 2021. 2