A Generalized Mesh Prediction Framework for Arbitrary Objects

Ziyi Zhou
Carnegie Mellon University

Chih-Chun Yang
Carnegie Mellon University

Cheng-Yen Hsieh
Carnegie Mellon University

Abstract

In this work, we address the challenge of open-category 3D mesh prediction using a deep learning approach. Our objective is to develop a highly accurate and efficient model capable of predicting the 3D mesh of objects belonging to a diverse set, with structures unseen during training. To achieve this, we adopt a CNN-based architecture with a mesh generation module that produces a 3D mesh from the CNN's output. We train our model on a toy dataset, manually selected from a large-scale 3D dataset, Objaverse, to simulate the open-category scenario. In addition to presenting qualitative and quantitative results, we conduct a thorough analysis of the correlation between 2D representations and 3D structures of objects, demonstrating the potential of achieving a generalized mesh prediction framework.

1 Introduction

Humans perceive objects in three dimensions, but we typically capture this information as two-dimensional images. Unfortunately, it's difficult to reconstruct a 3D object from a 2D image due to the loss of spatial information that occurs when transforming from 3D to 2D. Despite this challenge, deep neural networks have made impressive strides in reconstructing 3D shapes. These networks use various 3d shape representation to approach the problem, such as voxels, pointclouds, and meshes. However, the challenge of reconstructing 3D shapes from 2D images is complicated by the presence of occlusions, which require hallucinating the unseen part of the object.

The task of single-view 3D reconstruction has been a well-known problem within the computer vision community. However, previous works[6, 10, 13] have primarily focused on a group of known objects, without considering the fact that in the real world, individuals often encounter unknown objects. The generalization of 3D reconstruction from known to unknown objects remains an open issue, and there have been limited studies exploring this problem. One of the main difficulties in this task is domain shift, whereby the distribution of object shapes and appearances in unknown categories may differ significantly from those in known categories, leading to a discrepancy between the training and testing datasets. Moreover, there may be structural and semantic variations between known and unknown categories, making it challenging to transfer knowledge from one category to another. For instance, an object in the unknown category may have a different topology or a varying number of parts than those in the known categories.

In this work, we introduce a framework that addresses the challenge of single-view 3D reconstruction for open categories. Our proposed framework aims to accurately estimate the 3D mesh of known objects while maintaining strong generalization performance to unseen objects. The framework is composed of a convolutional neural network that estimates the coarse 3D voxel representation of the input image. The estimated 3D voxel is then cubified to serve as the prior for the subsequent mesh prediction. To further refine the mesh prediction, we incorporate a graph neural network that leverages local information from the coarse mesh. Our proposed framework paves the way for future research in the field of open-category 3D mesh prediction to analyze the latent representations learned by the

Preprint. Under review.

model, enabling a better understanding of the key factors contributing to the model's generalization ability towards unseen objects.

Now we highlight the contribution of this project as follows.

- We presents a framework that tackles the challenge of single-view 3D reconstruction for open categories. The proposed framework aims to estimate the 3D mesh of known objects while having strong generalization ability towards unseen objects.
- Incorporated in the framework is a graph neural network that utilizes local information to enhance the generalization ability and refine the mesh prediction.
- By examining the generalization performance of our 3D reconstruction framework from seen to unseen objects, we reveal the relationship between 2D images and 3D objects.

2 Related Work

2.1 Single-View Mesh Reconstruction

Single-view mesh reconstruction has gained significant attention in the computer vision community due to its potential applications in fields such as robotics, virtual reality, and augmented reality. However, it is a challenging task as it involves estimating the 3D geometry of an object from a single 2D image. Recent advances in deep learning and computer vision have led to the development of several powerful algorithms that can generate high-quality 3D mesh models from a single image.

One such method is Mesh R-CNN[6], which uses a two-stage CNN architecture for single-view mesh reconstruction. In the first stage, an object detector is used to detect the object in the input image and predict its 2D bounding box. In the second stage, a mesh generator generates a 3D mesh model of the object from the detected 2D bounding box. The method has shown to outperform state-of-the-art methods on several benchmark datasets. Pixel2Mesh[13] is another method that generates a coarse 3D mesh model from a single 2D image using a CNN. The coarse mesh is then refined using a graph CNN that updates the mesh vertices based on their local geometry and image features. This method has shown to generate high-quality 3D mesh models that are more accurate than other methods.

Deep Marching Cubes[11]is a method that predicts the occupancy probabilities of the 3D voxels that represent the object using a deep learning approach. These occupancy probabilities are then converted into a 3D mesh using the marching cubes algorithm. The method has shown to generate high-quality 3D mesh models with fine details. Another method is 3D-R2N2[2], which generates a voxel grid of the object from a single image using a CNN. The voxel grid is then converted into a 3D mesh using a modified marching cubes algorithm. The method has shown to generate high-quality 3D mesh models that are comparable to state-of-the-art methods. Finally, AtlasNet[7] is a novel method for 3D mesh generation from point clouds that can be used for single-view mesh reconstruction. The method uses a deep neural network to map point clouds onto a 2D surface parameterization of a sphere, which is then converted into a 3D mesh using the inverse of the parameterization. The method has shown to generate high-quality 3D mesh models that are more accurate than other methods.

In conclusion, single-view mesh reconstruction is a challenging task that has gained significant attention from the research community. The recent advances in deep learning and computer vision have led to the development of several powerful algorithms that can generate high-quality 3D mesh models from a single image. These methods have shown promising results on existing datasets, but there is much improvement needed for the open category object reconstruction.

2.2 Open-World Prediction

The open-world prediction problem has been extensively studied in the field of machine learning and artificial intelligence. In 2011, Scheirer et al. introduced the open-set recognition problem, where they extended the traditional classification problem to include unknown classes. They proposed several methods for open-set recognition, including the Nearest Open Set and the Open-Set Support Vector Machine [12].

In recent years, deep learning-based approaches have become increasingly popular for open-world prediction. In 2018, Bendale and Boult [1]proposed a technique called Open Set Domain Adaptation, which uses adversarial learning to adapt a classifier to unknown classes. In the same year, Ge et al.

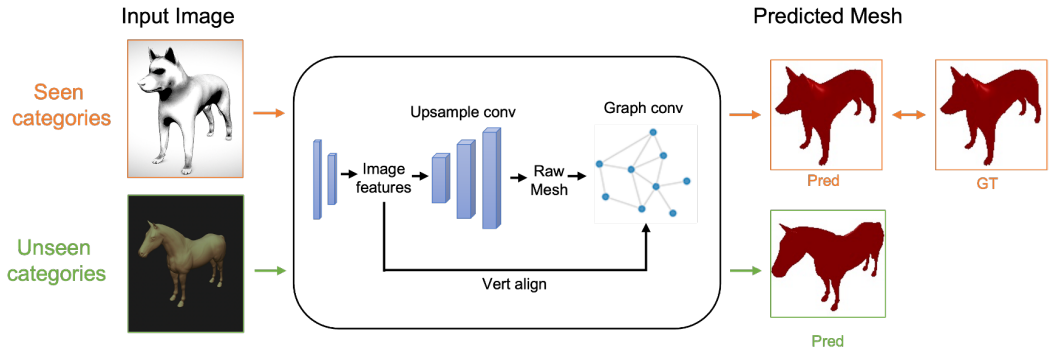

Figure 1: **3D open-category mesh prediction framework.** Given single-view 2D images, the method reconstructs 3D meshes for objects from seen and unseen categories. During training, 3D ground truth meshes are provided for objects from seen categories. Graph convolutional layers are employed to better capture local 3D structures, enabling better generalization across diverse categories.

proposed an approach called Open-Set Domain Generalization, which extends Open Set Domain Adaptation to the domain generalization setting [5].

Overall, the above works represent a diverse range of techniques for open-world prediction, including traditional machine learning methods as well as deep learning-based approaches. The development of new algorithms and frameworks in this area is expected to continue in the future, as the open-world prediction problem remains an important challenge in the field of machine learning.

3 Methods

3.1 Task Overview - Open-Category Mesh Prediction

The aim of this study is to develop a system that takes a single image as input and outputs a 3D triangle mesh for objects from both seen categories and unseen categories. To simulate open-category scenarios in the real world, the model could be trained using both seen and unseen category images, while relying solely on 3D ground truth meshes from seen categories. This approach enables the model to effectively reconstruct 3D meshes for novel objects and achieve superior performance in open-category settings. Even though images from unseen categories are provided in the designed open-category setting, we did not employ them in our training objectives, as our primary focus is to conduct a comprehensive analysis of the relationship between representations and 3D structures.

3.2 Single-view Image Mesh Predictor

Similar to [6], our framework first generates a raw mesh prediction from extracted 2D image representations and refine the raw mesh before generating the final mesh prediction. In contrast to [6], our framework refrains from using classification heads, which may limit the ability of the feature extractor and mesh prediction head to acquire information beyond known categories. Instead, our approach prioritizes the capture of structural information and thus ensures that the framework is able to capture 3D structures of novel objects.

3.2.1 Raw Mesh Generation

We use CNN-based architecture (e.g., Resnet-50 [8]) to extract 2D image representations of shape (C, H, W). The feature maps are then upsampled by using several Transpose 2D convolutional layers. In this project, we used five upsampled layers with a sigmoid layer to produce (G, G, G) voxel occupancy probabilities $\hat{\mathbf{V}}$, where G denotes the voxel grid size and is selected as 48 in our implementation. The predicted voxels are transformed into meshes and cubified to generate an initial coarse approximation of the 3D structures. These raw meshes serve as a solid foundation that can be further refined to achieve more detailed and precise 3D object shapes. To ensure the model generates

a good coarse approximation of the object, we employ binary cross entropy loss $BCE(.,.)$ between the predicted voxels and the ground truth voxels, which could be obtained by voxelizing the ground truth meshes, as follows:

$$L_v = \frac{1}{N}\sum_{i=1}^{N}\sum_{g=1}^{G^3} BCE(\hat{\mathbf{V}}_{i,g}, \mathbf{V}_{i,g}) \tag{1}$$

where L_v denotes voxel loss, $\mathbf{V}_{i,g}$ denotes the g-th grid in the i-th ground truth voxel, and N denotes the batch size.

3.2.2 Mesh Refinement

To capture fine-grained details of the object, raw meshes are further refined with a mesh refinement branch. The mesh refinement branch employs a series of refinement stages that adjust the vertex positions. Each refinement stage [13] consists of three operations: vertex alignment, which extracts image features for the vertices; graph convolution, which conveys information along the edges of the mesh and is important to capture 3D local structures; and vertex refinement, which updates the positions of the vertices.

Specifically, ***vertex alignment*** is used to obtain image representations for each vertex. Given a image feature map, we can use the bilinearly interpolated image feature [9] to obtain the feature vector for each vertex based on its position.

The ***graph convolutional layers*** transfer information between adjacent vertices. The vertex features f_i are transformed into updated features $f_i' = ReLU(W_0 f_i + \sum_{j \in N(i)} W_1 f_j)$, where $N(i)$ is the set of all neighboring vertices of the i-th vertex, and W_0 and W_1 are weight matrices that can be trained. The graph convolution operation inherently captures 3D local areas by analyzing the relationship between nearby vertices.

vertex refinement updates the vertex positions v_i to $v_i' = v_i + \tanh(R(f_i))$, where R consists of several graph convolutional layers that transforms vertex features into offset feature that could be further transformed into vertex offsets with tanh operation. In practice, we concatenate vertex positions and vertex features to include positional information during the refinement process. Lastly, to ensure the refined meshes are precise, we use chamfer distance [4] loss L_{cham} and edge loss [6] L_{edge}, which serves as a shape regularizer. Some work also use normal distance loss to penalize mismatched normal vectors between two meshes. Yet, we empirically found that using normal distance loss degrades the visual quality of predicted 3D meshes.

3.2.3 Training Objective Summarization

In summary, the training objective consists of the voxel loss L_v, chamfer distance loss L_{cham} and the edge loss L_{edge}. The training objectives could thus be formulated as:

$$L = L_v + \lambda_{cham} L_{cham} + \lambda_{edge} L_{edge}, \tag{2}$$

where λ_{cham} and λ_{edge} are hyperparameters used to balance each loss term, and are selected as 10.0 and 5.0 respectively. With these designed objectives, our method achieves promising results in terms of both quantitative and qualitative results, as shown in the following section.

4 Experiments

4.1 Dataset: Objaverse

Objaverse [3] is a large-scale dataset of 3D objects created by the Allen Institute for AI (AI2), a nonprofit research organization that focuses on advancing artificial intelligence research. The dataset consists of over 62,000 3D models, each represented as a mesh with texture, and covers a wide range of categories, including animals, vehicles, furniture, and more. Each object is represented as a 3D mesh with texture and is accompanied by its metadata, including its category, subcategory, and semantic attributes.

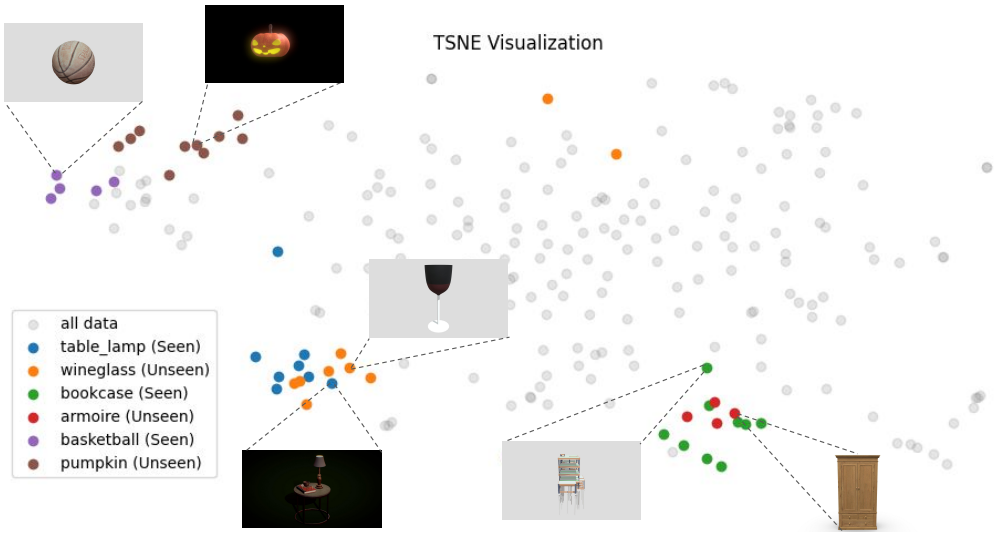

Figure 2: **TSNE visualization.**

Categories	$F1^{0.1}$	$F1^{0.3}$	$F1^{0.5}$	Chamfer Distance
Seen	85.9427	99.5634	99.9941	0.0102
Unseen	79.1717	99.0235	99.9849	0.0148

Table 1: **Quantitative result.**

4.1.1 Toy dataset construction

We split the Objaverse dataset by randomly selecting 15 seen categories and 15 unseen categories. Objects in seen and unseen categories are further divided into training and validation set. During the training phase, the model was trained on 2D images and 3D annotations of objects in the training set of seen categories, as well as training set 2D images of objects in the unseen categories. For evaluation, the model was tested on 2D images of validation set objects in both the seen and unseen categories.

4.2 Quantitative Results

Evaluation Metrics: Our framework is evaluated using metrics similar to those used in recent works on single-view 3D object reconstruction [6]. The Chamfer distance and $F1^{\tau}$ at different distance thresholds τ are calculated. The $F1^{\tau}$ is calculated as the harmonic mean of precision and recall at τ, where precision is the fraction of predicted points that are within τ distance from a ground-truth point, and recall is the fraction of ground-truth points that are within τ distance from a predicted point. Better performance is indicated by lower Chamfer distance values, and higher values of $F1^{\tau}$.

Results: In Table 1, we present the quantitative results for both seen and unseen objects. Our framework demonstrates promising performance on seen objects. For unseen objects, our framework exhibits only a minor decline in performance, indicating its strong generalization ability.

4.3 Qualitative Results

4.3.1 Correlation between 2D representations and 3D structure

Now we explore the relationship between 2D representations and 3D structures by analyzing the learned representations from the ResNet encoder in our framework using TSNE visualization. The results, as shown in Figure 2, reveal a clear domain shift between the seen and unseen objects.

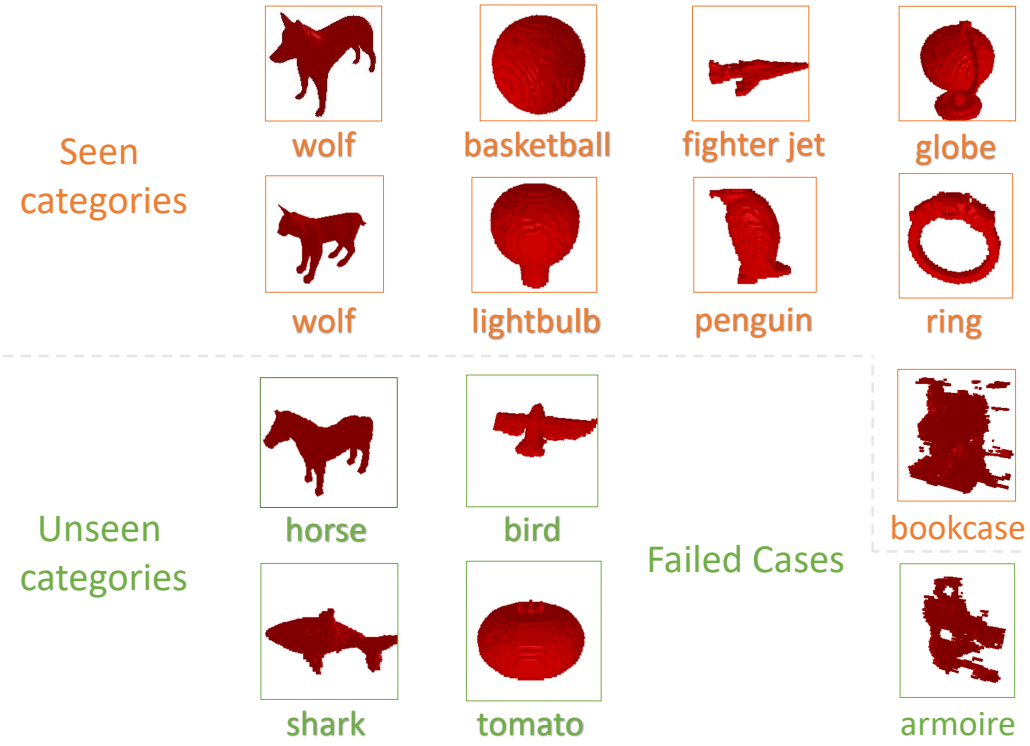

Figure 3: **3D mesh visualization on validation set.**

Moreover, we observe that objects with similar 3D structures tend to cluster together in the feature space. For example, objects such as basketball and pumpkin, table lamp and wineglass, and bookcase and armoire share similar 3D shapes, and thus their 2D representations are closer to each other compared to other object categories. These findings demonstrate the potential of leveraging the learned representations to achieve better generalization in 3D reconstruction tasks.

4.3.2 3D Mesh Visualization

The effectiveness of our proposed method is demonstrated through qualitative results of 3D mesh predictions, as illustrated in Figure 3. Our method successfully reconstructs fine details for most objects from seen categories, such as the legs of animals, and accurately predicts complex topologies for objects with holes, such as the ring. For objects from unseen categories, our method produces reasonable meshes for those with similar shapes to objects from seen categories. Based on this observation and our hypothesis, as detailed in Section 4.3.1, we propose that the quality of mesh prediction for an unseen category may be positively correlated with the mesh prediction results for similar seen categories. We provide evidence to support this claim through our failure to predict reasonable structures for bookcase (seen category) and armoire (unseen category), which have similar shapes and are close in 2D representation space shown in Figure 2. We further suggest that it may be possible to enhance the quality of mesh prediction for unseen categories by leveraging the 3D labels of seen categories that are close in 2D representation space, which will be left as our future work.

5 Conclusion

In this work, we aim to open up promising avenues for future research in open-category 3D mesh prediction. We present a CNN-based approach with a mesh generation branch for open-category 3D mesh prediction. Our framework is evaluated on a toy dataset manually constructed from the Objaverse 3D dataset to simulate the open-category setting. The use of graph convolutional layers in the mesh generation branch improves the capture of 3D local structures, as demonstrated by our quantitative and qualitative results. Furthermore, our in-depth TSNE analysis highlights the

correlation between 2D representations and 3D structures, suggesting the potential for improved 3D mesh predictions for unseen categories by incorporating 2D representations as priors. In future work, we aim to leverage 2D representations of objects from unseen categories to enhance 3D mesh prediction performance. We are also excited to extend our approach to toy datasets with more diverse categories, with the ultimate goal of achieving a generalized mesh prediction framework.

References

[1] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1563–1572, 2018.

[2] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 628–644, 2016.

[3] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. 2022.

[4] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 605–613, 2017.

[5] Weijian Ge, Yu Yang, Yinghao Yu, and Thomas Huang. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9107–9115. IEEE, 2018.

[6] Georgia Gkioxari, Jitendra Malik, and Justin Johnson. Mesh r-cnn. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9785–9795, 2019.

[7] Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan Russell, and Mathieu Aubry. Atlasnet: A papier-mache approach to learning 3d surface generation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 216–224, 2018.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[9] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28, 2015.

[10] Angjoo Kanazawa, Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 371–386, 2018.

[11] Jonas Lorenz and Suvrit Sra. Deep marching cubes: Learning explicit surface representations. In *International Conference on Learning Representations*, 2019.

[12] Walter J Scheirer, Anderson Rocha, Ajaya Sapkota, and Terrance E Boult. Towards open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1757–1772, 2011.

[13] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 52–67, 2018.