Learning in-the-wild 3D Vehicle Generation for Realistic Sensor Simulation

Ze Yang, Jingkang Wang, Jiageng Mao

Abstract—Generating realistic 3D objects and rendering them at arbitrary viewpoints is important to scale the content creation and sensor simulation for robotics training the testing. To achieve this goal, we need to learn a 3D generative model that is able to synthesize 3D objects with diverse and accurate geometry, robust and photo-realistic appearance, and can be rendered efficiently. The generated 3D contents thus can be immediately deployed to facilitate downstream applications. Existing works on 3D objects generation usually cannot generate high-fidelity geometry, or cannot generate photo-realistic renderings compared to real images. This introduces a large domain gap when composing those assets with real images. In this project, we plan to propose a 3D geometry-aware framework that learns 3D assets generation from real-world data. We model the shape and appearance of the 3D object using an implicit neural feature fields, and utilize the differentiable volume rendering and neural rendering to synthesize the 2D image.

Index Terms—Neural Scene Representation, Volume Rendering, Genarative Model

1 Introduction

Inspired by the tremendous progress in 2D image generation [1], [2], [3], 3D content generation has attracted more and more attention in recent years. Existing works demonstrated the high-quality generation in different representations including point cloud [4], [5], [6], [7], voxel grid [8], [9], [10], [11], [12], mesh [13], [14], [15] or implicit geometry [16], [17], [18], [19]. However, these works usually focus on the synthetic datasets where the observations are dense and the objects are created with simplified materials and lighting conditions. Those assumptions will not hold in the real world. Specifically, the observations are often sparse and noisy (e.g., noisy segmentation masks, imperfect calibration and localization, etc). Therefore, the quality for generated meshes is not sufficient for the real applications (See Figure N in the state-of-the-art work [19]) such as realistic sensor simulation for self-driving.

In this project, we will focus on the object-level in-the-wild 3D model generation. Built on top of the existing approaches (Pi-GAN [18], EG3D [16] and GET3D [19]), the ultimate goal is to generate a diverse set of 3D vehicles that contains realistic baked texture (more advanced material modeling is not considered) and can be rendered for actor insertion. Specifcially, we will explore two directions 1) volume rendering based approach (EG3D), 2) differentiable rendering based approach (GET3D). Since both works have released the code and partial pre-trained models, we plan to use the public codebase and finetune on the real data. The overall pipeline can be summarized as follows: Given some latent codes, the network will produce some implicit feature grids or SDFs; We render the images at random viewpoints given the implicit representations either by volume rendering the feature grids or using differentiable mesh extraction and rendering; Finally, we use a GAN to judge whether the rendered images are real or fake. The full pipeline is differentiable and end-to-end trainable.

After experimenting with these two approaches, we will analyse the performance and potentially propose some new techniques to improve the performance. Here are some initial thoughts: (1) Since the real world observations are quite sparse (several images with limited viewpoints), it is usually challenging to generate without sufficient data priors. We could add some template meshes (e.g., vehicle CAD models) as an initialization and let the network to predict the vertex offset, scale, etc. (2) We could first use the pre-trained models that already consumes the class specific priors and then plug some mapping layers for style transfer.

In summary, we would like to bridge the gap between synthetic and real 3D generation. The generated high-quality 3D vehicles can be potentially used to create an alternative asset bank compared to expensive/unrealistic 3D CAD models or inefficiently reconstructed assets that are widely used in the industry.

2 Related Work

Existing works on 3D generation from images can be divided based on the 3D representations they used and the supervisory signals. Occupancy networks [20] leverage implicit representations to learn 3D reconstruction from the functional space. PointFlow [21] learns to generate 3D point clouds from images with point-wise supervision. Texture3D [22] proposes to reconstruct 3D meshes and textures from images. Those methods generally rely on 3D supervisory signals. However, since 3D models are relatively expensive to obtain, these approaches are hard to generalize to real-world scenarios.

In addition to reconstruction leveraging 3D signals, there is also a category of works that generate novel views without 3D supervision. NeRF [23] is a pioneering work that proposes neural radiance field for novel view synthesis. Numerous papers [16], [17], [18], [24] have been trying to improve NeRF for 3D-aware novel view synthesis. However, these methods are restricted to view synthesis

of objects or simple indoor scenes, and cannot effectively handle complex driving scenarios.

3 Experiments and Plans

3.1 Datasets

We plan to conduct experiments on the real world PandaSet [25]. PandaSet is a dataset captured by the self-driving vehicle platform equipped with 6 cameras (front, front-left, left, front-right, right and back cameras) and two LiDARs (a 360° mechanical spinning LiDAR and a solid forward-facing LiDAR). The cameras and LiDARs are calibrated. PandaSet annotates instance-level 3D bounding boxes for common traffic participants in urban scenes, which can be used to extract camera images and LiDAR sweeps for diverse set of vehicles, motorcycles, *etc.* We are primarily interested in learning vehicle generation model, since vehicles are the most common actor in self-driving scenes.

3.2 Metrics

To evaluate the quality of our generations, we compute both geometry and appearance metrics on the generated shapes. For geometry, we adopt the aggregated point cloud as ground-truth shapes, and use the Chamfer Distance to compute the Minimum Marching Distance between the generated shapes and the ground-truth shapes. For appearance, we compute the FID metric between the observed camera images and our rendered camera images for the vehicles.

3.3 Future Plans

3.3.1 Prepare the datasets (Nov 17th - Nov 20th):

We will use the annotated 3D vehicle bounding boxes in PandaSet to obtain the LiDAR point cloud and cropped vehicle images for each vehicle in the dataset. In order to obtain the segmentation mask, we will use the off-the-shelf algorithm to segment the vehicles from the backgrounds.

3.3.2 Implement the model (Nov 20th - Nov 27th):

We will build on top of the official repository for EG3D and GET3D. We choose StyleGAN as our initial architecture for generator. The feature map generated by the StyleGAN can be used to derive the implicit representation of the generated assets, from which we can render the camera image, sihouette and LiDAR point clouds. We train a discriminators to classifier whether the inputs are real or fake. The generator is trained to maximize the likelihood of the discriminator output.

3.3.3 Analyze and improve the model (Nov 27th - Dec 4th):

After implementing the model, we will conduct experiments and tune the hyper-parameters / architecture, and analyse the algorithm to find the optimum designs.

3.3.4 Write report and prepare poster (Dec 4th - Dec 7th):

We will summarize the results, write the reports and prepare the poster for presentation.

References

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[2] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.

[3] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8780–8794, 2021.

[4] L. Caccia, H. Van Hoof, A. Courville, and J. Pineau, "Deep generative modeling of lidar data," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 5034–5040.

[5] G. Yang, X. Huang, Z. Hao, M.-Y. Liu, S. Belongie, and B. Hariharan, "Pointflow: 3d point cloud generation with continuous normalizing flows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4541–4550.

[6] P. Achlioptas, O. Diamanti, I. Mitliagkas, and L. Guibas, "Learning representations and generative models for 3d point clouds," in *International conference on machine learning*. PMLR, 2018, pp. 40–49.

[7] L. Zhou, Y. Du, and J. Wu, "3d shape generation and completion through point-voxel diffusion," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5826–5835.

[8] J. Wu, C. Zhang, T. Xue, B. Freeman, and J. Tenenbaum, "Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling," *Advances in neural information processing systems*, vol. 29, 2016.

[9] S. Lunz, Y. Li, A. Fitzgibbon, and N. Kushman, "Inverse graphics gan: Learning to generate 3d shapes from unstructured 2d data," *arXiv preprint arXiv:2002.12674*, 2020.

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

[11] R. Girdhar, D. F. Fouhey, M. Rodriguez, and A. Gupta, "Learning a predictable and generative vector representation for objects," in *European Conference on Computer Vision*. Springer, 2016, pp. 484–499.

[12] P. Henzler, N. J. Mitra, and T. Ritschel, "Escaping plato's cave: 3d shape from adversarial rendering," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9984–9993.

[13] T. Groueix, M. Fisher, V. G. Kim, B. C. Russell, and M. Aubry, "A papier-mâché approach to learning 3d surface generation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 216–224.

[14] G. Gkioxari, J. Malik, and J. Johnson, "Mesh r-cnn," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9785–9795.

[15] T. Shen, J. Gao, K. Yin, M.-Y. Liu, and S. Fidler, "Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis," *Advances in Neural Information Processing Systems*, vol. 34, pp. 6087–6101, 2021.

[16] E. R. Chan, C. Z. Lin, M. A. Chan, K. Nagano, B. Pan, S. De Mello, O. Gallo, L. J. Guibas, J. Tremblay, S. Khamis *et al.*, "Efficient geometry-aware 3d generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 123–16 133.

[17] K. Schwarz, Y. Liao, M. Niemeyer, and A. Geiger, "Graf: Generative radiance fields for 3d-aware image synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 20 154–20 166, 2020.

[18] E. R. Chan, M. Monteiro, P. Kellnhofer, J. Wu, and G. Wetzstein, "pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 5799–5809.

[19] J. Gao, T. Shen, Z. Wang, W. Chen, K. Yin, D. Li, O. Litany, Z. Gojcic, and S. Fidler, "Get3d: A generative model of high quality 3d textured shapes learned from images," *arXiv preprint arXiv:2209.11163*, 2022.

[20] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, "Occupancy networks: Learning 3d reconstruction in function space," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4460–4470.

[21] G. Yang, X. Huang, Z. Hao, M.-Y. Liu, S. Belongie, and B. Hariharan, "Pointflow: 3d point cloud generation with continuous normalizing flows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4541–4550.

[22] D. Pavllo, J. Kohler, T. Hofmann, and A. Lucchi, "Learning generative models of textured 3d meshes from real-world images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 879–13 889.

[23] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[24] J. Gu, L. Liu, P. Wang, and C. Theobalt, "Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis," *arXiv preprint arXiv:2110.08985*, 2021.

[25] P. Xiao, Z. Shao, S. Hao, Z. Zhang, X. Chai, J. Jiao, Z. Li, J. Wu, K. Sun, K. Jiang *et al.*, "Pandaset: Advanced sensor suite dataset for autonomous driving," in *ITSC*, 2021.