Supplementary Material for CRM

Zhengyi Wang[1,2], Yikai Wang[1,2], Yifei Chen[1], Chendong Xiang[1,2], Shuo Chen[1], Dajiang Yu[1], Chongxuan Li[3], Hang Su[1] and Jun Zhu[⋆1,2]

[1] Dept. of Comp. Sci. & Tech., BNRist Center, Tsinghua-Bosch Joint ML Center,
Tsinghua University;
[2] ShengShu, Beijing, China;
[3] Gaoling School of Artificial Intelligence, Renmin University of China, Beijing Key
Laboratory of Big Data Management and Analysis Methods, Beijing, China.

A Limitations

Although our method can generate a high-fidelity textured mesh in 10 seconds, there are still some limitations. As also a limitation in ImageDream [12], if the input image has a large elevation or different FoV, the results are sometimes not satisfactory. Also, it is very hard to ensure that the multi-view diffusion model always generates fully consistent results, and inconsistent images may make the 3D results degrade. What's more, the Flexicubes grid size is only 80 owing to the limited computing resource, which cannot represent very detailed geometry.

Additionally, since we use rendered images surrounding the objects to train the model (similar to previous works like LRM), it is hard to reconstruct surfaces inside the objects. One possible solution is to directly regress on the sdf/occupancy values of the 3D meshes, which may provide supervision inside the objects. We leave this as future work.

B Why do we use CCM as geometry representation?

In this work, we utilize the Canonical Coordinate Map (CCM) for geometric representation. Given that the pose is fixed to six orthographic views, each CCM has only one valid channel, while the other two channels maintain a constant value. This indicates that the CCM can equivalently be transformed into a depth map. We actually don't need the object to be canonically oriented. The choice of CCM over the depth map is motivated by its ability to represent identical parts of an object in the same color, which we hypothesize may aid the multi-view diffusion model in learning correspondences between object parts.

We do not assume canonical orientation for objects in the training set. During the training of the multi-view image diffusion model, we randomly select one of the six orthographic images as the input and use the diffusion model to predict the remaining five images. This approach enhances the model's generalization ability regarding the input pose.

[⋆] The Corresponding author.

C Additional Experiments

C.1 Additional Ablation Study

Flexicubes v.s. DMTet We conduct an ablation study on the use of geometry representation of our CRM. We compare between Flexicubes [11] and DMTet [10], two gradient-based mesh optimization methods. We trained a reconstruction model using DMTet geometry, with results shown in Fig. 12. Overall, both Flexicubes and DMTet are capable of reconstructing good geometry. However, the surface reconstructed with DMTet exhibits some non-smooth artifacts. Therefore, we finally selected Flexicubes for our geometry representation.

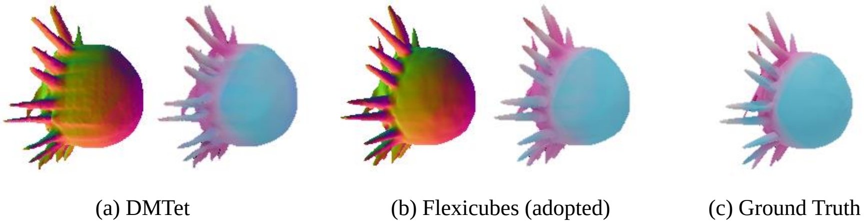

Fig. 12: Ablation study on different choices of geometry representation for CRM (zoom in for better details).

C.2 Failure Cases

In Fig. 13, we present some failure cases of the textured meshes generated by our method. As observed in the figure, our results occasionally exhibit a darker back view, which we suspect may result from the lighting conditions in the training dataset. Additionally, when the input image features a large elevation, the outcomes are sometimes unsatisfactory, which also reflects a limitation inherited from ImageDream [12]. Also, sometimes the generated meshes exhibit slight non-smooth geometry, which may result from the inconsistency in the multi-view images.

C.3 Results of Multi-view Diffusion Models

Here we provide additional results for the multi-view image diffusion models along with the results of CCM diffusion in Fig. 15. We demonstrate the diversity of multi-view image diffusion model in Fig. 16.

C.4 User Study

We conduct a user study to demonstrate the effectiveness of CRM in a subjective perspective. We evaluate on 115 diverse input images, including 34 real-captured photos of daily-life object taken by us, 20 real-captured photos from

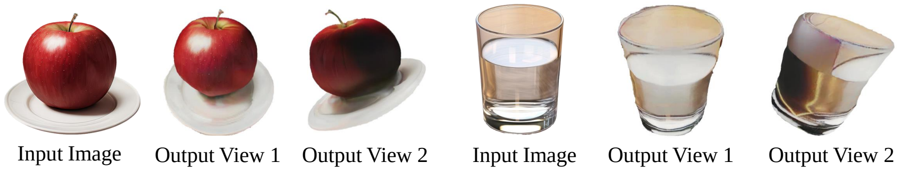

Fig. 13: Failure cases of our method include instances where the generated mesh displays a darker back view, possibly due to the lighting conditions in the training dataset. Additionally, results may exhibit distortion when the input image is at a large elevation.

Table 5: Results of user study. The percentage of user preference (↑) is reported in the table. We report the results on all 115 diverse images, 34 real-captured photos take by us and 20 real-captured images from MVImgNet, respectively.

Name	All	Captured	MVImgNet
Prefer OpenLRM	40.5	38.2	39.2
Prefer CRM (ours)	**59.5**	**61.8**	**60.8**

MVImgNet [14], 20 from GSO [1] scanned objects, 10 from OmniObject3D [13] scanned objects and 31 images of daily objects from the web. We compare with the most competitive baseline, OpenLRM [5]. We hire 55 volunteers from a crowd-source platform to compare the generated meshes based on texture and geometry quality, yielding 6325 pairwise comparisons. In Table 5, our method outperforms baseline, which demonstrate that our method is effective in a subjective perspective and is robust on diverse input images.

C.5 More Generation Results

We provide more results on real-captured photos in Fig. 14 to show the effectiveness of our method for real-world input images.

D Open-sourced Assets

Here, we list the URL and citations of open-sourced assets that we use in our work in Table 6.

References

1. Downs, L., Francis, A., Koenig, N., Kinman, B., Hickman, R., Reymann, K., McHugh, T.B., Vanhoucke, V.: Google scanned objects: A high-quality dataset of 3d scanned household items. In: 2022 International Conference on Robotics and Automation (ICRA). pp. 2553–2560. IEEE (2022) 3

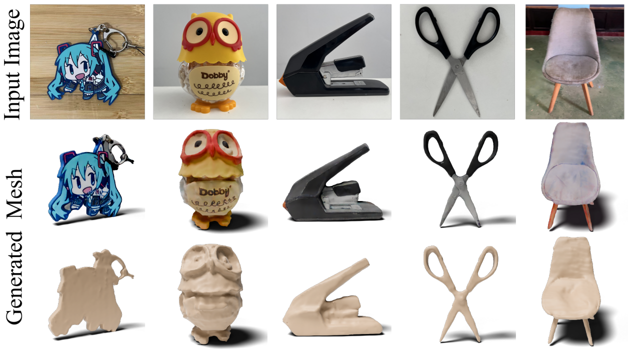

Fig. 14: More results of CRM on real-captured photos.

Table 6: URL and citations of open-sourced assets that we use in our work.

URL	Citation
https://github.com/nv-tlabs/GET3D	[3]
https://github.com/threestudio-project/threestudio	[4]
https://github.com/Stability-AI/stablediffusion	[9]
https://github.com/NVIDIAGameWorks/kaolin	[2]
https://github.com/huggingface/diffusers	[8]
https://github.com/bytedance/ImageDream	[12]
https://github.com/NVlabs/nvdiffrast	[6]
https://github.com/NVlabs/nvdiffrec	[7]

2. Fuji Tsang, C., Shugrina, M., Lafleche, J.F., Takikawa, T., Wang, J., Loop, C., Chen, W., Jatavallabhula, K.M., Smith, E., Rozantsev, A., Perel, O., Shen, T., Gao, J., Fidler, S., State, G., Gorski, J., Xiang, T., Li, J., Li, M., Lebaredian, R.: Kaolin: A pytorch library for accelerating 3d deep learning research. https://github.com/NVIDIAGameWorks/kaolin (2022) 4
3. Gao, J., Shen, T., Wang, Z., Chen, W., Yin, K., Li, D., Litany, O., Gojcic, Z., Fidler, S.: Get3d: A generative model of high quality 3d textured shapes learned from images. Advances In Neural Information Processing Systems **35**, 31841–31854 (2022) 4
4. Guo, Y.C., Liu, Y.T., Shao, R., Laforte, C., Voleti, V., Luo, G., Chen, C.H., Zou, Z.X., Wang, C., Cao, Y.P., Zhang, S.H.: threestudio: A unified framework for 3d content generation. https://github.com/threestudio-project/threestudio (2023) 4
5. He, Z., Wang, T.: Openlrm: Open-source large reconstruction models. https://github.com/3DTopia/OpenLRM (2023) 3
6. Laine, S., Hellsten, J., Karras, T., Seol, Y., Lehtinen, J., Aila, T.: Modular primitives for high-performance differentiable rendering. ACM Transactions on Graphics **39**(6) (2020) 4
7. Munkberg, J., Hasselgren, J., Shen, T., Gao, J., Chen, W., Evans, A., Müller, T., Fidler, S.: Extracting Triangular 3D Models, Materials, and Lighting From Images.

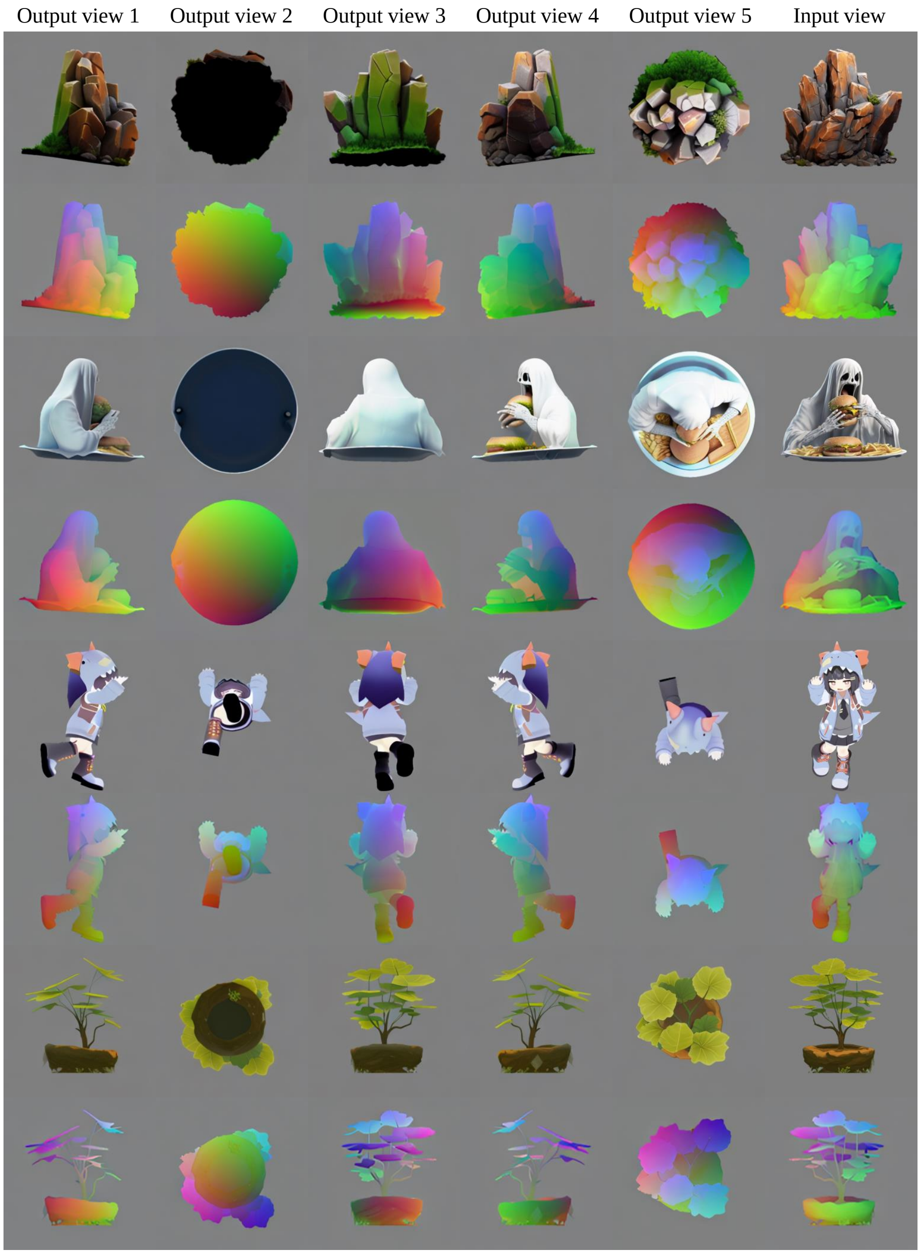

Fig. 15: More results from our multi-view image diffusion model along with the CCM diffusion model. Given an input image, the multi-view image diffusion model generates the other 5 images and CCM diffusion model generates the 6 CCMs. Our model can generalize to different types of input images.

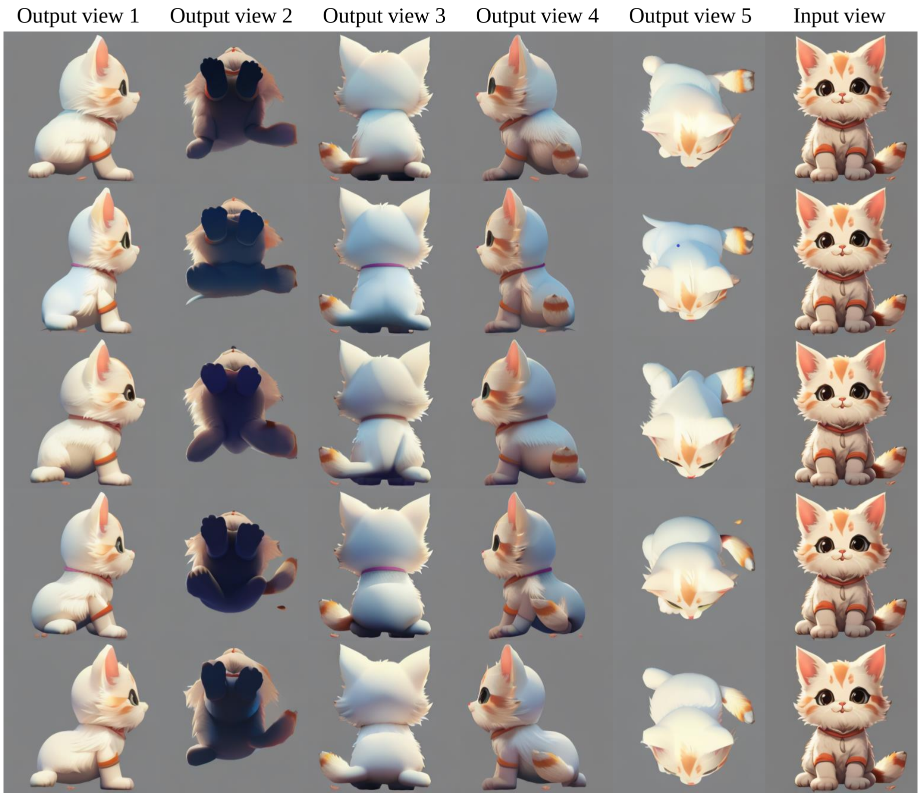

Fig. 16: Diversity of our multi-view image diffusion model.

In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8280–8290 (June 2022) 4

8. von Platen, P., Patil, S., Lozhkov, A., Cuenca, P., Lambert, N., Rasul, K., Davaadorj, M., Wolf, T.: Diffusers: State-of-the-art diffusion models. `https://github.com/huggingface/diffusers` (2022) 4
9. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) 4
10. Shen, T., Gao, J., Yin, K., Liu, M.Y., Fidler, S.: Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. Advances in Neural Information Processing Systems **34**, 6087–6101 (2021) 2
11. Shen, T., Munkberg, J., Hasselgren, J., Yin, K., Wang, Z., Chen, W., Gojcic, Z., Fidler, S., Sharp, N., Gao, J.: Flexible isosurface extraction for gradient-based mesh optimization. ACM Transactions on Graphics (TOG) **42**(4), 1–16 (2023) 2
12. Wang, P., Shi, Y.: Imagedream: Image-prompt multi-view diffusion for 3d generation. arXiv preprint arXiv:2312.02201 (2023) 1, 2, 4
13. Wu, T., Zhang, J., Fu, X., Wang, Y., Ren, J., Pan, L., Wu, W., Yang, L., Wang, J., Qian, C., et al.: Omniobject3d: Large-vocabulary 3d object dataset for realis-

tic perception, reconstruction and generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 803–814 (2023) 3

14. Yu, X., Xu, M., Zhang, Y., Liu, H., Ye, C., Wu, Y., Yan, Z., Zhu, C., Xiong, Z., Liang, T., et al.: Mvimgnet: A large-scale dataset of multi-view images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9150–9161 (2023) 3