Annotated Bibliography **Ben Berry & Cole Determan**

Chen, Z., Tang, J., Dong, Y., Cao, Z., Hong, F., Lan, Y., Wang, T., Xie, H., Wu, T., Saito, S., Pan, L., Lin, D., & Liu, Z. (2025). 3DTopia-XL: Scaling High-quality 3D Asset Generation via Primitive Diffusion. In Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://openaccess.thecvf.com/content/CVPR2025/papers/Chen_3DTopia-XL_Scaling_High-quality_3D_Asset_Generation_via_Primitive_Diffusion_CVPR_2025_paper.pdf

In this paper, the authors describe a new 3D generation model called 3DTopia-XL. 3DTopia-XL is made from PrimX, a 3D representation of shape, albedo (which is essentially the color of a surface before lighting), and the material of a textured mesh (a 3D object) in a N x D tensor (where N represents the number of building blocks or primitives and D represents the dimensionality or feature information like 3d position, scale, properties, etc.). This model generates PBR (Physically Based Rendering) assets so that game developers and 3D designers can use them in software like Blender or Unity in a GLB format (a format that stores all necessary details like textures, geometry, etc. in one file). PrimX is also very efficient, allowing rapid tensorization (taking lower-order data like a vector matrix into a higher-order data type). The model builds a generative framework on top of PrimX using a Diffusion Transformer architecture (a specific neural network architecture for denoising in a diffusion model). The paper found that this model can generate 3D assets with high-resolution geometry and textures with realistic materials. And, because of the Diffusion Transformer architecture, as well as how PrimX stores its data, the model is very efficient and scalable. This paper cites Denoising Diffusion Probabilistic Models because it established the framework and efficient training method that all modern diffusion models (including 3DTopia-XL) are built on.

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, 33, 6840–6851. https://proceedings.neurips.cc/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf

The authors introduce diffusion probabilistic models in this paper. These models are used to generate high quality images. Diffusion works by starting with an image, and gradually adding more and more random noise to it over many time steps. When this process is complete, the image will resemble something like TV static. Then, the model learns to predict what the image looked like one time step before, until it gets to the original image. Once the model is trained, it can be given pure noise and create new realistic images. They found that this model could match or outperform other image generation methods at the time.

Liang, J., Ma, X., Han, X., Lu, J., & Zhou, B. (2024). LucidDreamer: Towards High-Fidelity Text-to-3D Generation via Interval Score Matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 20076–20085). https://openaccess.thecvf.com/content/CVPR2024/papers/Liang_LucidDreamer_Towards_High-Fidelity_Text-to-3D_Generation_via_Interval_Score_Matching_CVPR_2024_paper.pdf

The authors introduce LucidDreamer, a framework for generating high quality 3D assets from text by improving prior methods utilizing pretrained 2D generation models to supervise 3D asset generation. Firstly, they explain the problems with commonly used Score Distillation Sampling (SDS), which is a method of accomplishing this. SDS works by rendering a 3D model from a random camera angle to get an image. Then, the 2D model generates a score (a gradient to make the image closer to what it should look like) which is used by SDS to adjust its parameters and match the 2D models expectations. The author's main concern with SDS is that it produces inconsistent and low quality, commonly over-smoothed, results. So, they present Interval Score Matching (ISM). This method uses interval based score matching between 2 steps to allow more stable and higher quality results. They find that this method produces better visual quality, finer textures, and faster convergence compared to SDS.

Lin, J., Yang, X., Chen, M., Xu, Y., Yan, D., Wu, L., Xu, X., Xu, L., Zhang, S., & Chen, Y.-C. (2025). Kiss3DGen: Repurposing Image Diffusion Models for 3D Asset Generation. In Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://openaccess.thecvf.com/content/CVPR2025/papers/Lin_Kiss3DGen_Repurposing_Image_Diffusion_Models_for_3D_Asset_Generation_CVPR_2025_paper.pdf

The authors present Kiss3DGen, which repurposes a pretrained 2D image diffusion model into a 3D asset generator. Instead of starting from scratch on some large 3D dataset, the 2D model is tuned to create a “3D Bundle Image” This consists of multi-view images, which are a series of 2D images of a 3D object from different viewpoints to capture the texture of the object, along with normal maps which represent the geometry of the object. These are then put together to construct the 3D model. Doing so allows this problem to effectively be converted from a 3D generation task to a 2D task, allowing the many strategies of 2D image generation to be used instead. The system is also designed to allow text or 2D image conversions to 3D, along with editing 3D assets. This method improves the efficiency of high quality 3D asset generation, increasing the technology’s usability and practicality. This paper cites LucidDreamer as a method that has promising results, but is often time consuming because of the iterative optimization required to update the 3D representation.

Poole, B., Jain, A., Barron, J. T., & Mildenhall, B. (2023). DreamFusion: Text-to-3D using 2D diffusion. International Conference on Learning Representations (ICLR). https://openreview.net/pdf?id=FjNys5c7VyY

This paper introduced a method to generate 3D models from text by using large, trained 2D text to image diffusion models. This solved a big issue of needing to train models from 3D datasets. The process uses SDS (as mentioned below) to serve as a probability density distillation mechanism (a mechanism for transferring complex data distributions from large models to smaller, more efficient models). The process works like a 3D version of “Deep Dream”, where a

randomly-initialized 3D model is optimized using gradient descent. The model renders into multiple 2D images from random viewing angles and the SDS uses the 2D diffusion model to adjust the 3D models' parameters so that the views are consistent with the text prompt. The results were objects that could be viewed from any angle and easily composited into 3D environments.

Zhao, W., Cao, Y.-P., Xu, J., Dong, Y., & Shan, Y. (2025). DI-PCG: Diffusion-based Efficient Inverse Procedural Content Generation for High-quality 3D Asset Creation. In Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://openaccess.thecvf.com/content/CVPR2025/papers/Zhao_DI-PCG_Diffusion-based_Efficient_Inverse_Procedural_Content_Generation_for_High-quality_3D_CVPR_2025_paper.pdf

In this paper, the authors describe DI-PCG, which is a method for inverse PCG (Procedural Content Generation) from images. Inverse PCG is a way to automate finding the parameters needed by a PCG program to create a 3D asset. Normally, PCG is very tricky to deal with, because it's a complex process of adjusting parameters to be generated correctly, but DI-PCG helps to automate it. This method uses a lightweight diffusion transformer model that learns the PCG parameters by treating them as the denoising target (a clean, original data that a denoising model is trained on to recontrust/predict). The paper found that DI-PCG is very efficient, needing only 7.6 million network parameters and a short training time, as well as generating 3D assets in only a few seconds. This, as well as the fact that the outputs are high quality editable 3D assets, makes the DI-PCG a flexible tool for making any existing procedural generator better. This paper cites DreamFusion as one of the earlier methods used for 3D asset creation using 2D diffusion models, using similar approaches but receiving better results.