Evaluating Perceptual fidelity of Text to 3D Models

Anonymous CVPR submission

Paper ID *****

Abstract

The field of text-to-3D generative methods has seen remarkable progress in recent times, driven by a series of breakthroughs. Despite this progress, the existing evaluation metrics often focus on a single criterion, such as the alignment between the input text and the generated 3D models, but they do not comprehensively evaluate the quality of the generated 3D model itself. Traditional methods for evaluating 3D models typically measure the distance between generated and reference shape distributions. However, these methods are not readily applicable to text-conditioned generative tasks due to the difficulty in obtaining a comprehensive reference set, given the vast range of natural language inputs. In this work, we propose a novel approach to evaluate the visual perception of generated 3D models using surface normal and visual feature analysis. Surface normals provide crucial information about the geometry of a surface, describing aspects such as surface orientation, curvature, and shape. Visual features provide a comprehensive understanding of the image's content and context.

†

1. Introduction

Based on the recent traction in the area of Text-to-3D models, there have also been many methods introduced to evaluate the generated 3D models based on the input query. These evaluation methods check againt the fidelity of 3D model based on the text input GPT-4V(ision)[11], T3Bench[3], To evaluate the geometric consistency of the generated 3D models, we use surface normal analysis as a key metric. First, we generate 3D models from text inputs using a state-of-the-art text-to-3D model, represented as triangular meshes. Surface normals are then computed directly from the mesh geometry, serving as ground truth for comparison. To analyze the models from different perspectives, we capture 2D images from both canonical (e.g., front, side) and non-canonical (e.g., oblique, tilted) viewing angles. For surface normal prediction, we utilize StableNormal[12], a robust model designed to predict surface normals from images under complex lighting and geometric conditions. The predicted normals from StableNormal are compared with the mesh-derived normals using cosine difference as the primary metric, which measures the angular discrepancy between the two sets of normals. To ensure that only valid regions of the model are evaluated, a masking procedure is applied to exclude irrelevant pixels from the background. This approach allows us to assess the geometric fidelity of the 3D models across multiple views and varying levels of complexity, providing insight into the performance of text-to-3D generative models. We have also taken inspiration from text-to-Image evaluation methods [5], text-to-3DModel evaluation methods [7], [2].

2. Methodology

Our proposed methodology evaluates the fidelity of 3D surface reconstruction by combining quantitative metrics with qualitative visualizations. The framework begins with mesh preprocessing, where vertex and face data are extracted, followed by the projection of image-based features onto the mesh. Normal maps generated by the model are compared with ground truth using multiple evaluation metrics. Cosine similarity is computed for pixel-wise normal vector alignment, capturing directional differences, while the structural similarity index (SSIM) quantifies perceptual similarities. Additionally, learned perceptual image patch similarity (LPIPS)[13] is employed to measure perceptual fidelity using pre-trained neural networks such as AlexNet and VGG. We also consider using a more recent method[4] to compute FID score used specifically for Image generation.To enhance evaluation reliability, masked regions are incorporated, focusing computations only on valid, unoccluded areas of the normal maps. The variance of surface features, such as mean, standard deviation, and variance, is quantified and visualized on the 3D mesh using Open3D, providing insights into spatial feature distribution. Heatmaps visualize cosine similarity and SSIM metrics, while statistical summaries, including variance statistics, are generated. The implementation integrates Python libraries like PyTorch, Scikit-image, and Matplotlib for metric computations and visualizations, ensuring an efficient pipeline for comprehensive evaluation. This multi-faceted approach enables a robust analysis of reconstructed surfaces, blending traditional image-level metrics with 3D geometric insights to support meaningful comparisons and advancements in 3D reconstruction techniques. While new Gaussian Splatting methods like LGM[9], DreamBeast[6], we evaluate the 3D models generated by ProlificDreamer[10]. We evaluate the prompt "A 3D model of an adorable cottage with a thatched roof"

2.1. Texture Feature point Analysis

Texture Feature Point Analysis Texture feature point analysis is a key part of evaluating the spatial distribution and consistency of features across the reconstructed 3D surface. This analysis focuses on projecting image-based DINO-V2[8] features onto the mesh and quantifying their variance, standard deviation, and mean to capture feature stability and alignment. The process enables a deeper understanding of the texture fidelity in the reconstructed model, highlighting areas where feature representations may vary significantly across different views or reconstructions.

Feature Projection and Mapping Feature extraction begins by identifying and projecting relevant texture points from input images onto the corresponding 3D mesh vertices. These features, derived from image patches, are mapped to the closest vertices using a KD-tree-based nearest neighbor search, which efficiently matches 2D image locations to 3D surface points. Each vertex is then assigned a feature vector, allowing a consistent texture representation across the surface.

Variance and Consistency Quantification For each feature point on the mesh, the variance, standard deviation, and mean of feature values across different views are computed. These metrics are used to assess the consistency of the features, indicating the stability and reliability of texture information for each vertex. High variance suggest areas where feature points lack stability, potentially due to occlusions or inconsistent texture mapping across images, while lower variance reflects a stable and uniform feature representation.

Visualization of Feature Variance To provide a spatial understanding of feature consistency, variance values are visualized directly on the 3D mesh. Each vertex is colored based on its variance, creating a visual map of texture stability across the surface. High-variance regions are highlighted to indicate areas with potential instability in texture representation, while low-variance regions show where texture mapping is consistent and reliable. This visualization is saved as a 3D .obj file, allowing easy inspection. Figure of variance is shown in image 3

Interpretation and Use Texture feature point analysis offers insights into the spatial consistency of textures on 3D surfaces, highlighting potential areas of improvement

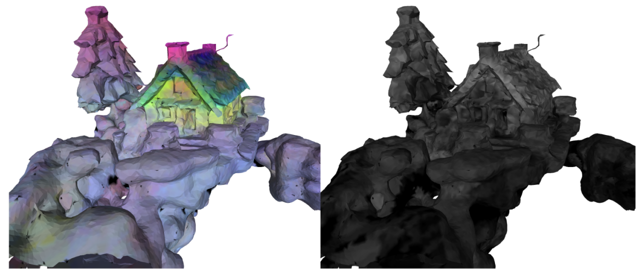

Figure 1. Left: shows the mean DINO-v2 features, Right: shows the standard deviation of the features.

in texture mapping and feature alignment. By integrating variance visualization and statistical reporting, this analysis serves as a robust tool for evaluating texture fidelity, enabling model developers to refine their approaches and enhance the visual realism of reconstructed surfaces.

2.2. Surface Normal Analysis

To evaluate the geometric consistency of the generated 3D models, we use surface normal analysis as a key metric. First, we generate 3D models from text inputs using a state-of-the-art text-to-3D model, represented as triangular meshes. Surface normals are then computed directly from the mesh geometry, serving as ground truth for comparison. To analyze the models from different perspectives, we capture 2D images from both canonical (e.g., front, side) and non-canonical (e.g., oblique, tilted) viewing angles. For surface normal prediction, we utilize StableNormal[12], a robust model designed to predict surface normals from images under complex lighting and geometric conditions. The predicted normals from StableNormal are compared with the mesh-derived normals using cosine difference as the primary metric, which measures the angular discrepancy between the two sets of normals. To ensure that only valid regions of the model are evaluated, a masking procedure is applied to exclude irrelevant pixels from the background. This approach allows us to assess the geometric fidelity of the 3D models across multiple views and varying levels of complexity, providing insight into the performance of text-to-3D generative models. We also considered to process the normal maps into 3D object inspired form [1].

The analysis of surface normals is a critical component of the proposed methodology, aiming to assess the accuracy and perceptual fidelity of reconstructed 3D surfaces. This process evaluates the alignment and similarity of normal maps generated by the reconstruction model against ground-truth normal maps using three complementary approaches: cosine similarity, structural similarity (SSIM), and learned perceptual image patch similarity (LPIPS).

Cosine Similarity Cosine similarity is employed to measure the directional alignment of surface normals on a per-pixel basis. Normal maps are first normalized to unit vectors, ensuring consistent magnitude across all normal vectors. The cosine similarity is then computed as the dot product of corresponding vectors, providing a scalar value between -1 and 1, where 1 indicates perfect alignment. The methodology further aggregates these values to compute average, variance, and median cosine similarity scores, enabling quantitative comparisons of directional accuracy.

Structural Similarity (SSIM) SSIM is used to evaluate the perceptual similarity between the reconstructed and ground-truth normal maps. By comparing luminance, contrast, and structural information, SSIM captures differences that are more aligned with human visual perception. This metric is computed pixel-wise across the entire normal map and visualized as a difference heatmap, highlighting areas with significant deviations.

Learned Perceptual Image Patch Similarity (LPIPS) LPIPS evaluates the perceptual quality of reconstructed normals using deep learning-based feature representations. By leveraging pre-trained networks such as AlexNet and VGG, LPIPS captures high-level perceptual differences that go beyond simple pixel-wise comparisons. The normal maps are resized and normalized to ensure compatibility with the network, and the LPIPS distance is computed for each pair of normal maps.

Mask Integration To ensure the robustness of the analysis, a mask is applied to exclude invalid or occluded regions of the normal maps. This focuses the evaluation on relevant areas, preventing noisy or undefined regions from skewing the results.

Visualization and Outputs The results of surface normal analysis are visualized through heatmaps that represent cosine similarity and SSIM metrics. These heatmaps provide an intuitive understanding of normal alignment and perceptual fidelity across the surface. Additionally, statistical metrics, including the mean and variance of cosine similarity and SSIM, are summarized in CSV files for quantitative comparison. The visualization of discrepancy in texture is shown in 2

This comprehensive analysis of surface normals enables a detailed assessment of reconstruction accuracy, combining traditional geometric alignment metrics with advanced perceptual measures. The integration of visualization and statistical reporting further facilitates a deeper understanding of model performance and areas for improvement.

3. Results

3.1. Surface Normal Analysis

Our results on evaluating 20 3D models generated by 5 generative models, including the most recent work Prolificdreamer, plot shown in Figure 4, show that canonical views (Rear and Side_Left), demonstrated high geometric consistency, with the highest mean cosine difference reaching 0.93, indicating strong alignment between the predicted and ground truth surface normals. In contrast, non-canonical views (Side View from Non-Right Angles, Bottom-Up, and Top-Down), showed comparatively lower consistency, with the lowest mean cosine difference being 0.88. Although these non-canonical views also displayed relatively good consistency, these findings emphasize the importance of focusing on non-canonical views to enhance the overall geometric fidelity of text-to-3D generative models.

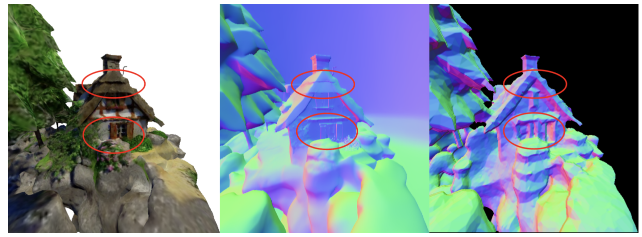

Figure 2. Shows mismatch in normals of the geometry and the texture. Left: Generated 3D model. Middle: 3D model's normal. Right: Normals generated using StableNormals using the Left image.

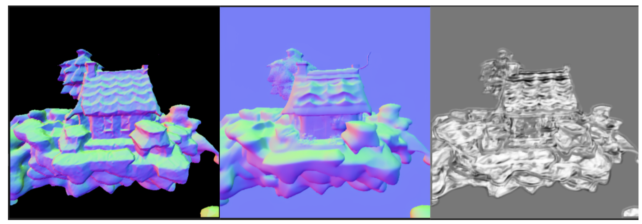

Figure 3. Left: Generated 3D model. Middle: 3D model's normal. Right: SSIM of the Normals image

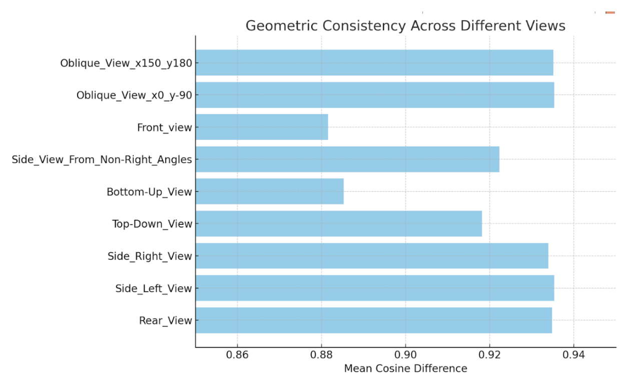

Figure 4. Plot of mean cosine differences of Surface Normal Analysis of 20 models, across various camera views.

References

[1] Xu Cao and Takafumi Taketomi. Supernormal: Neural surface reconstruction via multi-view normal integration. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 20581–20590, 2024. 2

[2] Yuze He, Yushi Bai, Matthieu Lin, Wang Zhao, Yubin Hu, Jenny Sheng, Ran Yi, Juanzi Li, and Yong-Jin Liu. T ˆ3 bench: Benchmarking current progress in text-to-3d generation. *arXiv preprint arXiv:2310.02977*, 2023. 1

[3] Yuze He, Yushi Bai, Matthieu Lin, Wang Zhao, Yubin Hu, Jenny Sheng, Ran Yi, Juanzi Li, and Yong-Jin Liu. T^3bench: Benchmarking current progress in text-to-3d generation, 2024. 1

[4] Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation, 2024. 1

[5] Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Teufel, Marco Bellagente, et al. Holistic evaluation of text-to-image models. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[6] Runjia Li, Junlin Han, Luke Melas-Kyriazi, Chunyi Sun, Zhaochong An, Zhongrui Gui, Shuyang Sun, Philip Torr, and Tomas Jakab. Dreambeast: Distilling 3d fantastical animals with part-aware knowledge transfer, 2024. 2

[7] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pages 366–384. Springer, 2025. 1

[8] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. 2

[9] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation, 2024. 2

[10] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation, 2023. 2

[11] Tong Wu, Guandao Yang, Zhibing Li, Kai Zhang, Ziwei Liu, Leonidas Guibas, Dahua Lin, and Gordon Wetzstein. Gpt-4v(ision) is a human-aligned evaluator for text-to-3d generation, 2024. 1

[12] Chongjie Ye, Lingteng Qiu, Xiaodong Gu, Qi Zuo, Yushuang Wu, Zilong Dong, Liefeng Bo, Yuliang Xiu, and Xiaoguang Han. Stablenormal: Reducing diffusion variance for stable and sharp normal. *ACM Transactions on Graphics (TOG)*, 2024. 1, 2

[13] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 1