Hi3DGen: High-fidelity 3D Geometry Generation from Images via Normal Bridging

Chongjie Ye[1,2*] Yushuang Wu[2*] Ziteng Lu[1] Jiahao Chang[1]
Xiaoyang Guo[2] Jiaqing Zhou[2] Hao Zhao[3] Xiaoguang Han[1†]

[1]The Chinese University of Hong Kong, Shenzhen [2]ByteDance [3]Tsinghua University

Input Image a. Hunyuan3D-2.0 b. Clay c. Tripo-2.5 d. Trellis **e. Hi3DGen** (Ours)

Abstract

*With the growing demand for high-fidelity 3D models from 2D images, existing methods still face significant challenges in accurately reproducing fine-grained geometric details due to limitations in domain gaps and inherent ambiguities in RGB images. To address these issues, we propose **Hi3DGen**, a novel framework for generating high-fidelity 3D geometry from images via normal bridging. Hi3DGen consists of three key components: (1) an image-to-normal estimator that decouples the low-high frequency image pattern with noise injection and dual-stream training to achieve generalizable, stable, and sharp estimation; (2) a normal-to-geometry learning approach that uses normal-regularized latent diffusion learning to enhance 3D geometry generation fidelity; and (3) a 3D data synthesis pipeline that constructs a high-quality dataset to support training. Extensive experiments demonstrate the effectiveness and superiority of our framework in generating rich geometric details, outperforming state-of-the-art methods in terms of fidelity. Our work provides a new direction for high-fidelity 3D geometry generation from images by leveraging normal maps as an intermediate representation.*

1. Introduction

With the rapid advancement of computer vision and graphics technologies, the task of generating 3D models from 2D images has garnered significant attention in both academic and industrial domains. Despite significant advancements in recent years, existing methods remain inadequate in generating 3D models that sufficiently reflect the geometric details present in the input images, especially when dealing with real-world input images, which typically exhibit complex and rich geometric characteristics. Nevertheless, the ability to faithfully reproduce these geometric details in 3D generations is of paramount importance, as it directly influences the models' realism, precision, and overall applicability in practical scenarios.

Current state-of-the-art techniques for 3D generation from 2D images often rely on deep learning models to learn the direct mapping from the 2D RGB image to the 3D geometry. While these methods have shown promising results [37, 70, 84], their ability in producing fine-grained geometric details is inherently limited by several key factors. First, the scarcity of high-quality 3D training data restricts the model's ability to learn detailed geometric features. Second, there exists a significant domain gap between the training images (often rendered from synthetic 3D meshes) and test images of various possible styles, leading to suboptimal

* Equal Contribution.
† Corresponding author: hanxiaoguang@cuhk.edu.cn.

performance in practical applications. Third, the inherent ambiguity in RGB images, caused by lighting, shading, or complex object textures, further complicates the extraction of fine-grained geometric information.

To address these limitations, we propose to leverage normal maps as an intermediate representation to bridge the mapping from 2D RGB images to 3D geometry. Normal maps, which encode surface orientation information, offer several advantages for this task. First, by introducing strong 2D priors to process RGB images into normal maps, we can effectively alleviate the domain gap between synthetic training data and real-world applications, which eases the 2D-to-3D mapping learning. Second, normal maps, as a 2.5D representation, provide clearer geometric cues compared to RGB images, thereby having the potential of guiding the geometry learning more effectively, especially in producing fine-grained geometric details.

In this paper, we introduce **Hi3DGen**, a novel framework for high-fidelity 3D geometry generation from images via normal bridging. The framework consists of three key components: (i) an image-to-normal estimator (NiRNE) that achieves generalizable, stable, and sharp normal estimation through a noise-injected regressive network with dual-stream training to decouple the representation learning of low- and high-frequency image patterns; (ii) a normal-to-geometry learning approach (NoRLD) that employs normal-regularized latent diffusion learning to provide explicit 3D geometry supervision during training, significantly enhancing generation fidelity; and (iii) a 3D data synthesis pipeline that constructs the DetailVerse dataset, containing high-quality synthesized 3D assets, serving as important complementary of humman-created ones, to support the training of NiRNE and NoRLD. Our framework generates rich, fine-grained geometric details, surpassing state-of-the-art (SOTA) approaches in terms of generation fidelity, as shown in teaser figure.

Contributions Our key contributions are as follows:

- We propose Hi3DGen, the first framework that leverages normal maps as an intermediate representation to bridge the gap between 2D images and 3D geometry, addressing the limitations of existing methods in generating fine-grained details;
- We introduce NiRNE, which decouples the low-high frequency learning with noise-injected dual-stream training to achieve robust, stable, and sharp normal estimation from input images;
- We develop a data synthesis pipeline and construct the DetailVerse dataset, which contains high-quality synthesized 3D assets to support the training of our framework. We will also release this dataset and hope it can inspire related research;

2. Related Work

Datasets for 3D Generation Early 3D datasets typically encompass small-scale objects from a limited category range [8, 11, 73]. To address this limitation, researchers endeavor to expand 3D data repositories through scanning or multi-view photography [16, 31, 51, 62, 71]. This approach leads to the creation of large-scale datasets such as MVImgNet [72, 81]. However, the quality of the constructed data often falls short of the requirements for direct application in 3D generation tasks. Recently, larger-scale datasets have been constructed by aggregating available human-created 3D assets from a wide range of online sources [13, 14]. However, among the 10 million 3D assets in Objaverse-XL [14], 5.5 million are from GitHub [22], raising license concerns and high quality variety necessitating costly data cleaning, and another 3.5 million from Thingiverse [57] lack textures required by existing 3D generation pipelines. The remaining objects, mainly from Objaverse-1.0 [13], exhibit a severe imbalance, characterized by a scarcity of high-quality assets with complex geometric structures and rich surface details. This imbalance is a common issue in datasets of human-created 3D meshes, resulting in networks generating simplistic 3D models with significant loss of detail. To address this gap, this paper explores synthesizing 3D data with high semantic variety, geometric structure diversity, and surface detail richness, and utilizes them in the context of 3D generation as a non-trivial complement to human-created 3D assets.

Normal Estimation Monocular methods can be primarily divided into diffusion-based and regression-based approaches. Regression-based methods have advanced from early handcrafted features [27, 28] to deep learning techniques [18, 65, 85]. Recent progress includes leveraging large-scale data [17], estimating per-pixel normal probability distributions [2], adopting vision transformers [50], and conducting inductive bias modeling [1]. Though conducting deterministic prediction that ensures higher stability, regression-based methods struggle with generating fine-grained sharp details. Diffusion-based normal estimation has emerged with the adaptation of powerful text-to-image models [47, 52, 83]. For instance, Geowizard [21] incorporates a geometry switcher to handle diverse data distributions. Considering high-variance results caused by the inherent stochastic nature of diffusion processes [19], strategies such as affine-invariant ensembling [21, 32] and one-step generation [77] have been explored but come with computational intensity and oversmoothing issues. StableNormal [80] improves estimation stability by reducing diffusion inference variance via a coarse-to-fine strategy, but it remains challenged by imperfect stability. Differently, by deeply exploring the root causes of the sharpness produced by diffusion-based methods, we novelly propose a noise-

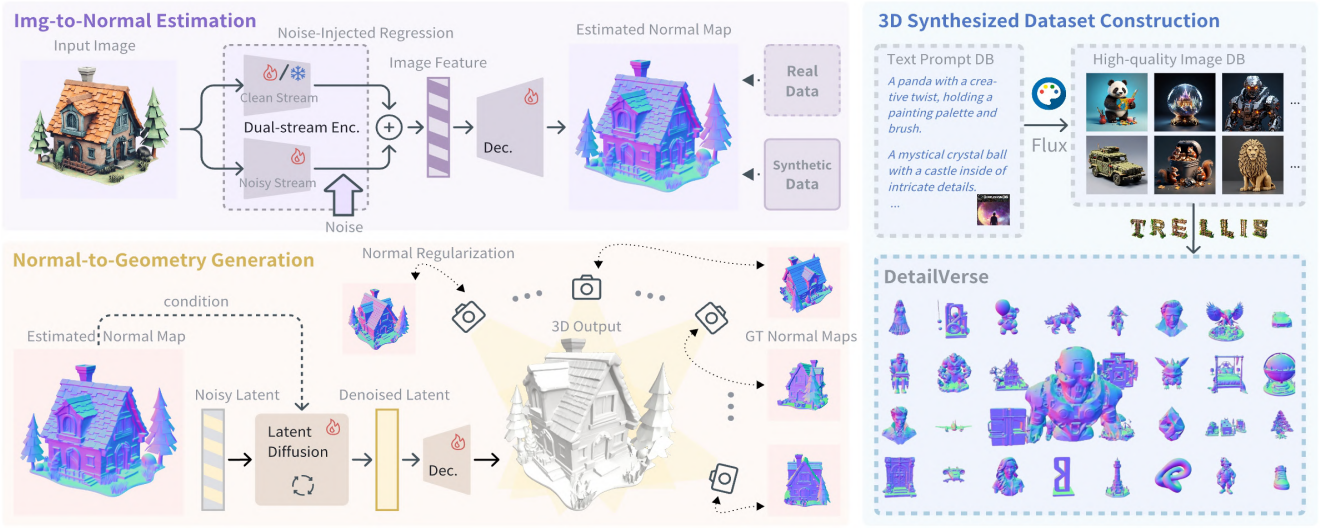

Figure 1. Overview of the proposed normal-bridged 3D geometry generation method. Our Hi3DGen comprises three components: an image-to-normal estimator, a normal-to-geometry generator, and a synthesized dataset (DetailVerse) construction pipeline.

injected regressive method to enable both sharp and stable estimations, with a dual-stream training strategy to fully utilize training data from different domains.

Normal Maps in 3D Generation Normal maps, which provide detailed geometric cues, have been widely used to enhance the fidelity and consistency of 3D reconstructions [5, 6, 55, 64, 68, 75, 76, 82]. Recently, they have also been explored in 3D generation. SDS-based methods [48, 63] render normal maps alongside RGB images in optimization to regularize geometry [23, 29, 49]. Other works use multi-view diffusion followed by reconstruction or fusion, generating normal images to complement RGB data and improve accuracy [3, 42, 44, 46, 58, 69, 79]. Though suffering from generating smooth surface details due to multi-view inconsistency, they have shown the significant potential of normal maps in enhancing 3D generation. In parallel, methods conducting 3D native diffusion based on 3D representations such as feature volumes, Triplane [7], 3D Gaussians [33] have leveraged normal maps by decoding them into meshes and applying normal rendering losses to regularize surfaces [15, 40, 78, 87]. However, these approaches often face limitations due to high memory requirements for high-resolution 3D representations. Meanwhile, methods focusing on latent code diffusion have achieved state-of-the-art performance [37, 38, 70, 74, 84]. However, the use of normal maps in this paradigm remains underexplored, where normal maps can not directly regularize the diffusion learning in the highly abstract latent space. Notable examples include CraftsMan [37], which uses normal map refinement as a post-processing step, and Trellis [74], which incorporates normal rendering loss during VAE training. Our approach uniquely emphasizes the critical role of normal maps in bridging image-to-3D generation and introduces a novel method to effectively integrate normal supervision into the diffusion learning of 3D latent codes, addressing limitations of prior work.

3. Method

This section outlines the proposed Hi3DGen framework, which aims to bridge the learning of 2D-to-3D lifting with 2.5D representation, normal map. Dividing the image-to-geometry generation into two parts, image-to-normal estimation and normal-to-geometry mapping, our framework consists of a dual-stream normal estimator for prediction stability and sharpness (Sec. 3.1) and an online normal regularizer for fine-grained generation details and image-geometry consistency in diffusion training (Sec. 3.2). We further propose a synthesized 3D dataset, which contains numerous generated 3D data of complex geometry structure and rich surface details, to facilitate sharp normal estimation and detailed 3D geometry generation (Sec. 3.3). An overview of the whole framework is visualized in Fig. 1.

3.1. Noise-Injected Regressive Normal Estimation

SOTA monocular normal estimation methods are mainly divided into diffusion-based and regression-based approaches. The former produces sharper results yet suffers from instability and spurious details due to their inherent probabilistic nature, while the latter offers stable one-step predictions but lacks sharpness. We first analyze the reason for sharper estimations of diffusion-based methods from the viewpoint of frequency domain. Then we propose integrating noise injection, the key mechanism in diffusion learning, into a regressive framework to enhance its sensitivity to high-frequency patterns, as illustrated in Fig. 2. Based on this, we further develop a dual-stream architecture to decou-

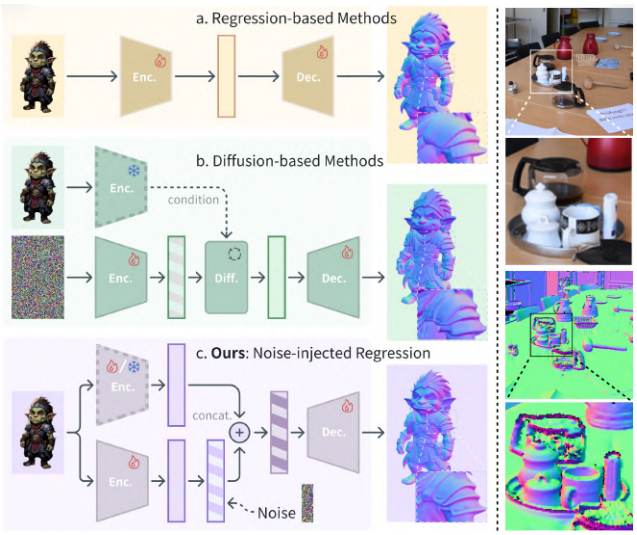

Figure 2. Left part: Illustration of Noise-injected Regressive Normal Estimation; Right part: Noisy label at high-frequency regions in real-domain data.

ple the low- and high-frequency representation learning for both generalizability and sharpness, with a domain-specific training strategy to stimulate the decoupled learning.

Noise Injection Considering normal sharpness usually appears at high-frequency image regions like edges and cavities, we begin by analyzing from the frequency domain the underlying mechanisms that enable sharp normal estimation results of diffusion-based methods. Defining the diffusion process with a stochastic differential equation:

$$x_t = x_0 + \int_0^t g(s) dw_t, \tag{1}$$

where the initial state X_0 evolves over time $t \in [0, T]$ to become x_t and w_t is a Wiener process (Brownian motion) representing injected random noise. By conducting Fourier transformation to this process, we can obtain the signal-to-noise ratio (SNR) of any frequency component ω at timestep t:

$$\text{SNR}(\omega, t) = \frac{|\hat{x}_0(\omega)|^2}{\int_0^t |g(s)|^2 ds}, \tag{2}$$

which is only subject to $\hat{x}_0(\omega)$ because the power of noise is equal over all ω. Since natural images exhibit low-pass characteristics, *i.e.*, $|\hat{x}_0(\omega)|^2 \propto |\omega|^{-\alpha}$ where $\alpha > 0$ represents the attenuation coefficient, the high-frequency components in x_t has a faster SNR degradation than low-frequency ones as the diffusion process progresses. This prompts that the model gets a stronger supervision at high-frequency regions in x_t, which encourages the model to focus more on capturing and predicting sharp details. Inspired by this, we integrate the noise injection technique into regression-based methods to encourage learning more high-frequency information.

Dual-Stream Architecture Compared to high-frequency features influencing the prediction sharpness, low-frequency features, conveying more overall structure information [9, 24], are important for the generalizability in low-level vision tasks [39]. To decouple these two kinds of features, we encode the input image through two independent streams: one processes the original image without noise injection to robustly capture low-frequency details (clean stream), while the other processes the noise-injected image to focus on high-frequency details (noisy stream). The latent representations from both streams are concatenated in a ControlNet-style manner [83] and fed into the decoder for final predictions, in a regression manner. This design uses noise injection in one stream to encourage high-frequency representation learning, and also maintains another clean stream to perceive the original image for regression, which effectively integrates the strengths of diffusion-based methods into a regressive method. An illustration of the method is presented in Fig. 2(c).

Domain-Specific Training To encourage the decoupled representation learning in two streams, we design a domain-specific training strategy to optimize the network by delicately utilizing training data from different domains. Previous methods mix real- and synthetic-domain data in training to enhance the generalizability. However, real-domain data, limited by the collection environment and the precision of scanners, suffer from noisy labels especially at object edges (see a visualized example in Fig. 2 right part), which hinder accurate learning at high-frequency details. In contrast, synthetic domain data, constructed via rendering from 3D ground truth, can provide precise high-frequency labels, while it is limited by the domain gap with real images in application. Therefore, we first train the network using real-domain data to capture low-frequency information for strong generalizability. In the second stage, we finetune the noisy stream using synthetic-domain data while freezing the parameters of the other stream. This allows the noisy stream to focus on learning high-frequency details as a residual component of outputs by the clean stream. The domain-specific training not only well utilizes the training data from real and synthetic domains according to their strengths, but also properly encourages the optimization of dual streams for decoupled representation learning.

3.2. Normal-Regularized Latent Diffusion

State-of-the-art 2D-to-3D generation methods rely on 3D latent diffusion, which represents 3D geometries in a compact latent space so that the 2D-to-3D mapping can be learned more efficiently [37, 70, 74, 84]. However, these methods suffer from easy loss of details or detail-level inconsistency with the input images (see examples in Fig. **??**).

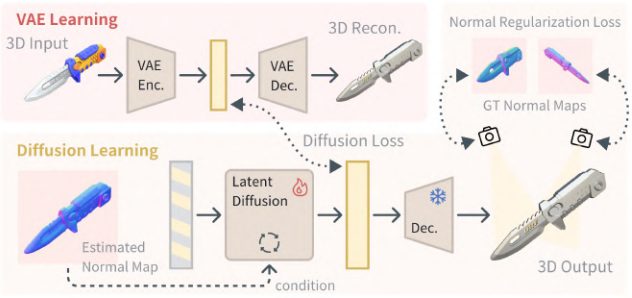

Figure 3. An illustration of Normal-Regularized Latent Diffusion.

Except for the ambiguity of RGB image input in indicating fine-grained geometries, another important reason is the indirect supervision from the latent space only, where the geometry information, especially fine-grained details, is usually greatly compressed to ensure compactness for diffusion learning.

Latent Diffusion We first formulate the typical latent diffusion process in 3D generation methods. A Variational Auto-Encoder (VAE) is trained to encode any 3D geometry X into a latent representation x_0 and decode it back to the original geometry $\hat{X}$:

$$x_0 = E(X), \quad \hat{X} = D(x_0), \tag{3}$$

where $E(\cdot)$ and $D(\cdot)$ denote the encoder and decoder, respectively. The reparameterization process is omitted for simplicity. The image-conditioned diffusion process constructs x_t by injecting noise into x_0 at a given timestep t and learns to recover x_0 from x_t. Flow matching is commonly used to address it, which aims to learn a continuous transformation by modeling the time-dependent velocity field, with the loss function formulated as:

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{t,x_0,x_t}\left[\left\|\mathbf{v}_\theta(x_t,t) - \mathbf{u}(x_t,t)\right\|^2\right], \tag{4}$$

where θ denotes the network parameters, $\mathbf{u}(x_t,t) = \nabla_{x_t}\log p(x_t|x_0)$ is the true velocity field, and the image/text condition is implicitly included.

Normal Regularization Regularization in the 3D geometry space allows for more precise supervision, especially over surface details. Therefore, we propose an enhanced loss function with explicit normal map regularization:

$$\mathcal{L}_{\text{Norld}} = \mathcal{L}_{\text{LDM}} + \lambda \cdot \mathcal{R}_{\text{Normal}}(\hat{x}_0), \tag{5}$$

where $\hat{x}_0$ represents the predicted clean sample and $\mathcal{R}_{\text{normal}}$ is the proposed regularization term:

$$\mathcal{R}_{\text{Normal}}(\hat{x}_0) = \mathbb{E}_v\left[\left\|R_v(D(\hat{x}_0)) - N_v\right\|^2\right], \tag{6}$$

where the predicted target latent $\hat{x}_0$ is decoded into explicit 3D geometry, R_v renders the normal map from viewpoint v, and N_v denotes the corresponding ground truth normal map. Note that this regularization is conducted online during the diffusion training process, as shown in Fig. 3, rather than in a post-processing stage. This actively guides the training of diffusion networks and aligns the predicted latent with a distribution that contains rich details consistent with the input images.

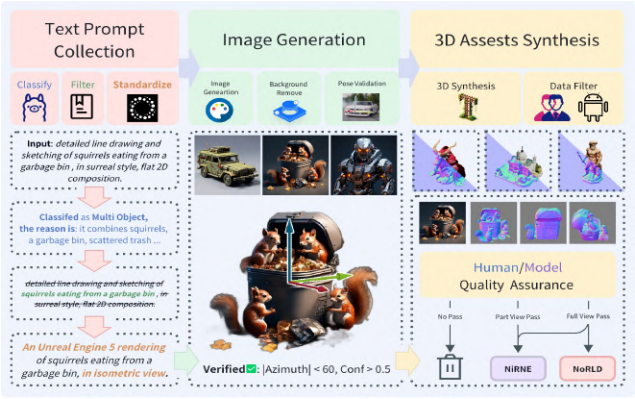

Figure 4. The procedure of DetailVerse Construction.

3.3. DetailVerse Dataset

High-quality 3D data is essential in the training of our NiRNE to provide clean normal labels and NoRLD for high-fidelity 3D generation. Although Objaverse [13, 14] provides a substantial number of image-normal and normal-geometry training pairs, the majority of the 3D assets exhibit simple structures and plain surface details, as shown in Tab. 1. This limitation restricts the generation capabilities of Hi3DGen. Given the prohibitive cost of manually creating high-quality 3D assets, we propose a 3D data synthesis pipeline to perform Text→Image→3D generation. By using advanced generators integrated with meticulous prompt engineering and data cleaning, this pipeline leads to a dataset DetailVerse of 700k synthesized 3D assets with considerably complex structures and rich details.

Dataset Construction We initiate the 3D data synthesis process with text prompts rather than image prompts because text prompts allow for more straightforward control of semantic diversity, thereby ensuring the variety of final geometries. we first sourced approximately 14M high-quality raw prompts from DiffusionDB [67]. A LLaMA-3-8B model [60] is adopted for classification to filter out complex scenes. Then, a rule-based filtering method to eliminate stylistic modifiers, together with a LLaMA-3-13B [60] for structural standardization to ensure consistent formatting. In the second step, we employ the SOTA Flux.1-Dev [35] for text-to-image generation. Additionally, we specify the text prompt condition to control the viewpoint and lightning in generation, and conduct pose validation using OrientAnything [66] to filter ones with special viewpoints, which is important to guarantee stable 3D generation. In the third step We employ Trellis [74], a

Table 1. Comparison of 3D object dataset statistics. The numbers X/Y in the third column means the Mean/Medium number.

Dataset	Obj #	Sharp Edge #	Source
GSO [16]	1K	3,071 / 1,529	Scanning
Meta [56]	8K	10,603 / 6,415	Scanning
ABO [11]	8K	2,989 / 1,035	Artists
3DFuture [20]	16K	1,776 / 865	Artists
HSSD [34]	6K	5,752 / 2,111	Artists
ObjV-1.0 [13]	800K	1,520 / 452	Mixed
ObjV-XL [14]	10.2M	1,119 / 355	Mixed
DetailVerse	700K	**45,773 / 14,521**	Synthesis

SOTA 3D generator, to conduct image-to-3D synthesis. Finally, a rigorous data cleaning process that combines expert evaluation with automated assessment preserves 700k high-quality meshes.

Dataset Statistics We present the model number in the dataset and the mean sharp edge number in each model in Tab. 1 to show the scale and geometric detail richness of our DetailVerse dataset. The sharp edge detection follows the implementation in Dora-Bench [10]. The synthesized assets in DetailVerse present rich surface details, as presented by the examples in the blue block of Fig. 1.

4. Experiments

4.1. Experiment Setup

Dataset For image-to-normal training, we utilize two complementary datasets. One is a diverse realistic dataset following Depth-pro[4]. Another contains synthetic data consisting of 20M RGB-to-normal pairs created by rendering 40 images per asset from 500k DetailVerse assets. For normal-to-geometry training, we curate a large-scale dataset comprising 170K cleaned 3D assets from Objaverse [13] and 700K synthesized 3D assets from our DetailVerse. We render 40 images per asset following Trellis [74]. For evaluation, the generalization ability of the image-to-normal estimator in real scenes is validated on the the reconstruction dataset LUCES-MV [41]. All images for visual comparison and user studies are collected from Hyper3D website [12], Hunyuan3D-2.0 project page [59], and Dora project page [54].

Implementation Details For image-to-normal, we adopt GenPercept [77] architecture for Normal Regression network. We initialize the encoder and decoder weights from the Stable Diffusion V2.1 [53], finetuned using the AdamW optimizer with a fixed learning rate of 3×10^{-5}. For normal-to-geometry, we build upon the Trellis [74], incorporating classifier-free guidance (CFG) [26] with a drop rate of 0.1 and AdamW [43] optimizer with a fixed learning rate of 1×10^{-4}. For the normal-to-geometry training stage, we finetune the Large variant of Trellis using 8 NVIDIA A800 GPUs (80GB each) for 50k steps with a batch size of 256. During inference, we set the CFG strength to 3.0 and use 50 sampling steps to achieve optimal results.

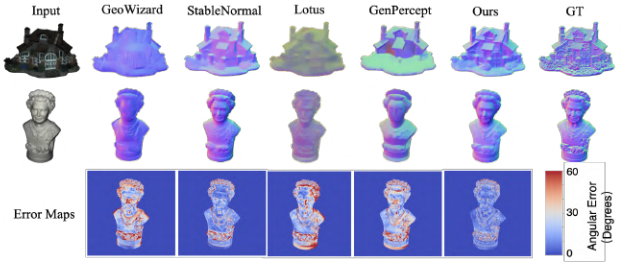

Figure 5. Normal estimation results comparison.

Evaluation Metrics For the evaluation of image-to-normal estimation, we basically use normal angle error (NE) to measure the overall prediction accuracy, measured in degrees. We additionally use the metric Sharp Normal Error (SNE) following Dora [10] to give emphasis on sharp edges where geometric details are most salient. For evaluating normal-to-geometry conversion, we render normal maps from 22 viewpoints around each object, which is used for compute NE and SNE to measure the overall and detailed geometry accuracy, respectively. More implementation details are included in the supplementary details.

Competitive methods We compare our *NiRNE* with SOTA normal estimators across different methodological categories. The comparison includes regression-based methods (Lotus [25] and GenPercept [77]), diffusion-based approaches (GeoWizard [21] and StableNormal [80]). Besides, *Hi3DGen* is compared with existing SOTA 3D generation methods including open-sourced CraftsMan-1.5 [37], Hunyuan3D-2.0 [86], Trellis [74], and close-sourced Clay[84], Tripo-2.5 [61], and Dora [10]. Note that Dora has not released its testing API, so we compare with Dora using the examples on its project page.

4.2. Image-to-Normal Estimation

Quantitative Results We provide the quantitative comparison between our NiRNE on LUCES-MV and other methods in Tab. 2. It validates that NiRNE gets significantly superior normal estimation performance to other regression- or diffusion-based methods, in both overall normal accuracy and sharp-region normal accuracy.

Qualitative Results A qualitative results is presented in Fig. 5, which shows that our NiRNE achieves superior estimation performance in (i) robustness with strong generalizability on human and object inputs; (ii) stability with less wrong details than diffusion-based methods (see error maps); and (iii) sharpness especially when compared with regression-based methods. These results further support our related claims in Sec. 3.1.

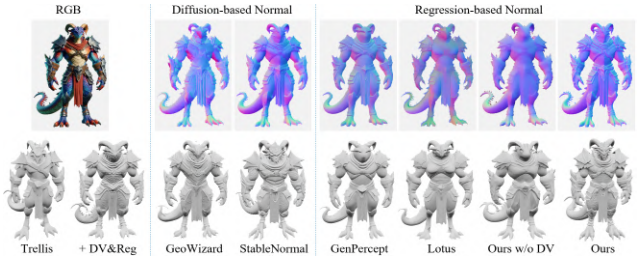

Figure 6. Ablations on the importance of normal bridging.

Figure 7. High-fidelity 3D results generated by our Hi3DGen.

Table 2. Performance comparison on image normal estimation. We use (Diff.) and (Regr.) to indicate diffusion- and regression-based methods, respectively. **Bold** indicates best results.

Method	NE ↓	SNE ↓
(Diff.) GeoWizard [21]	31.381	36.642
(Diff.) StableNormal [80]	31.265	37.045
(Regr.) Lotus [25]	53.051	52.843
(Regr.) GenPercept [77]	28.050	35.289
(Regr.) **NiRNE** (Ours)	**21.837**	**26.628**

4.3. Normal-to-Geometry Generation

Qualitative Results We give a qualitative comparison between the generated 3D geometries of the proposed Hi3DGen and other methods, as shown in Fig. 9. It impressively shows the superiority of our Hi3DGen in generating high-fidelity results with rich details that are consistent with the input images, which are easily lost by other methods. Besides, our Hi3DGen also produce robust generations with relatively smooth surface when less details presented in the input images (*e.g.* the first and third example in Fig. 9). We give more generation results of our Hi3DGen in Fig. 7, with more in supplementary materials.

User Study We conducted a user study to evaluate the 3D generation results of our Hi3DGen and 5 other methods including Hunyuan3D-2.0, Dora, Clay, Tripo-2.5, and Trellis. All 3D results for user study are randomly sampled from the 300×6 generations for visual comparison. The evaluation criteria focus on the fidelity of the generated 3D geometry to the input images, which is measured by the consistency in both overall shape and local details. For the parts of the input images that are not visible, we ask the evaluators to exercise their judgment and imagination to assess the plausibility of the generated results and their stylistic consistency with the visible portions. To ensure the comprehensiveness and professionalism of the user study, we invite two groups of evaluators. The first group consist of 50 amateur 3D users, who assess 100×6 randomly sampled results from the perspective of everyday applications, such as 3D printing. The second group includes 10 professional 3D artists, who evaluate 20×6 results from the standpoint of professional use, like 3D modeling and design. The results are presented in Fig. 8, which shows that our Hi3DGen achieves the highest generation quality for both amateur users and professional artists.

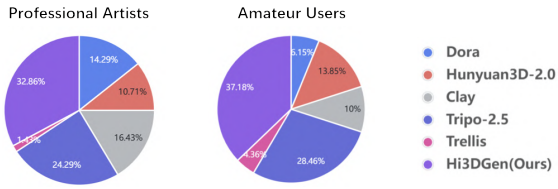

Figure 8. User study results.

4.4. Ablation Study

Normal Bridge We first validate the effectiveness of using normal maps to bridge 3D generation. A direct image-to-geometry generator based on Trellis [74] performs worse than our normal-bridged Hi3DGen, and when using the same normal regularization and training data as Hi3DGen, it produces fake details (see the first two columns v.s. the last column of Fig. 6). We also validate the influence of using normal conditions of different accuracy and sharpness to the final 3D generation quality. Smoother or wrong normal estimations by other methods lead to a performance drop, which also proves the importance of using accurate and sharp estimated normals as the bridge.

DetailVerse Data We also validate the value of the proposed DetailVerse dataset. By integrating image-normal training pairs rendered from DetailVerse data, our NiRNE can achieve 0.4 and 1.7 improvements in NE and SNE respectively, as shown in the first two rows of Tab. 3. By using additional normal-geometry training pairs from DetailVerse, our NoRLD can achieve higher-fidelity generation details, as shown in the 3rd and final columns in Fig. 10.

NiRNE Ablation We conduct ablative experiments to validate the three components in our NiRNE: the noise injection technique, the dual-stream architecture, and the domain-specific training. Results in Tab. 3 validates the ef-

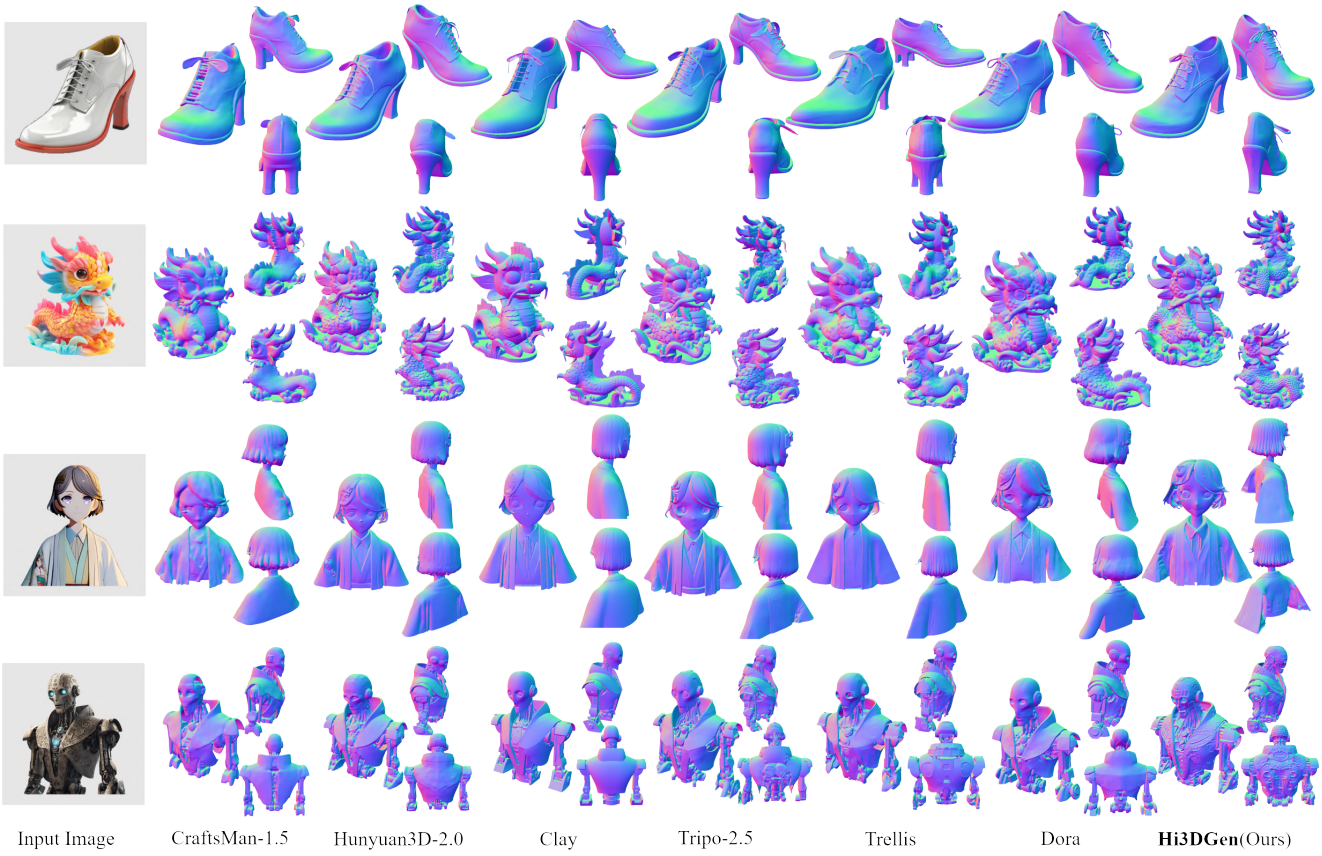

Figure 9. Qualitative 3D generation comparison on samples from Dora's project page [54].

Table 3. Ablation study on different components of NiRNE. "DV", "NI","DS", and "DST" denote DetailVerse data, Noise Injection technique, Dual-Stream architecture, and Domain-Specific Training strategy, respectively.

Method	NE ↓	SNE ↓
Ours (full model)	**21.837**	**26.628**
Ours w/o DV	22.209	28.324
Ours w/o DST	23.288	29.690
Ours w/o DS	21.918	29.520
Ours w/o all	22.507	35.997

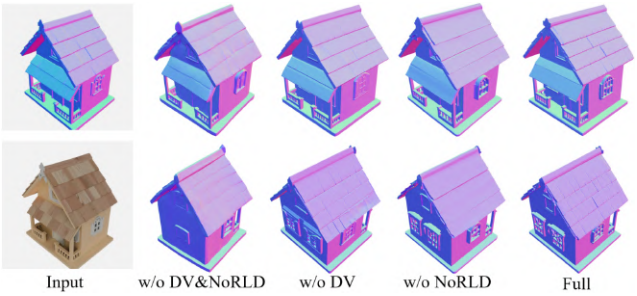

Figure 10. Ablation on the proposed NoRLD.

fectiveness of each component. More qualitative comparisons are included in the supplementary materials.

NoRLD Ablation We visualize the difference of not using the proposed online normal regularization in Fig. 10, which shows that whether using DetailVerse data for training, adopting normal regularization can greatly improve the generation fidelity (zoom in to see the roof details).

5. Conclusion

This paper propose Hi3DGen, a high-fidelity image-to-3D generation framework. It works by using normal map, a 2.5D representation, to bridge the 3D generation for rich details consistent with input images in generations. Hi3DGen consists of three components, a image normal estimator producing robust, stable, and sharp predictions, a normal-to-3D synthesis network with normal regularization for geometry consistency, and a 3D data synthesis pipeline providing detail-rich synthetic data for the training. Extensive experiments validate their effectiveness and superiority.

Limitations Although Hi3DGen generates detail-rich 3D results, part of them still have possible inconsistent or non-aligned details with the input, which is caused by the generative nature of the 3D latent diffusion learning. It is our future work to pursue reconstruction-level 3D generations.

References

[1] Gwangbin Bae and Andrew J. Davison. Rethinking inductive biases for surface normal estimation. In *CVPR*, 2024. 2

[2] Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. Estimating and exploiting the aleatoric uncertainty in surface normal estimation. In *ICCV*, 2021. 2

[3] Maciej Bala, Yin Cui, Yifan Ding, Yunhao Ge, Zekun Hao, Jon Hasselgren, Jacob Huffman, Jingyi Jin, JP Lewis, Zhaoshuo Li, et al. Edify 3d: Scalable high-quality 3d asset generation. *arXiv preprint arXiv:2411.07135*, 2024. 3

[4] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. *arXiv preprint arXiv:2410.02073*, 2024. 6

[5] Baptiste Brument, Robin Bruneau, Yvain Quéau, Jean Mélou, François Bernard Lauze, Jean-Denis Durou, and Lilian Calvet. Rnb-neus: Reflectance and normal-based multi-view 3d reconstruction. In *CVPR*, pages 5230–5239, 2024. 3

[6] Xu Cao and Takafumi Taketomi. Supernormal: Neural surface reconstruction via multi-view normal integration. In *CVPR*, pages 20581–20590, 2024. 3

[7] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *CVPR*, pages 16123–16133, 2022. 3

[8] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 2

[9] Genggeng Chen, Kexin Dai, Kangzhen Yang, Tao Hu, Xiangyu Chen, Yongqing Yang, Wei Dong, Peng Wu, Yanning Zhang, and Qingsen Yan. Bracketing image restoration and enhancement with high-low frequency decomposition. In *CVPR*, pages 6097–6107, 2024. 4

[10] Rui Chen, Jianfeng Zhang, Yixun Liang, Guan Luo, Weiyu Li, Jiarui Liu, Xiu Li, Xiaoxiao Long, Jiashi Feng, and Ping Tan. Dora: Sampling and benchmarking for 3d shape variational auto-encoders. *arXiv preprint arXiv:2412.17808*, 2024. 6, 1

[11] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, et al. Abo: Dataset and benchmarks for real-world 3d object understanding. In *CVPR*, pages 21126–21136, 2022. 2, 6

[12] Deemos. Rodin gen-1: A production-level 3d generation model. https://hyper3d.ai, 2024. 6

[13] 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. In *ICCV*, pages 13142–13153, 2023. 2, 5, 6, 1

[14] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *NeurIPS*, 36, 2024. 2, 5, 6

[15] Zijian Dong, Xu Chen, Jinlong Yang, Michael J Black, Otmar Hilliges, and Andreas Geiger. Ag3d: Learning to generate 3d avatars from 2d image collections. In *ICCV*, pages 14916–14927, 2023. 3

[16] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *ICRA*, pages 2553–2560, 2022. 2, 6

[17] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *ICCV*, pages 10786–10796, 2021. 2

[18] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *ICCV*, pages 2650–2658, 2015. 2

[19] Martin Nicolas Everaert, Athanasios Fitsios, Marco Bocchio, Sami Arpa, Sabine Süsstrunk, and Radhakrishna Achanta. Exploiting the signal-leak bias in diffusion models. In *WACV*, pages 4025–4034, 2024. 2

[20] Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3d-future: 3d furniture shape with texture. *IJCV*, 129:3313–3337, 2021. 6

[21] Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *ECCV*, pages 241–258. Springer, 2024. 2, 6, 7

[22] Inc. GitHub. Github: Where the world builds software. https://www.github.com, 2025. Accessed: 2025-03-07. 2

[23] Chun Gu, Zeyu Yang, Zijie Pan, Xiatian Zhu, and Li Zhang. Tetrahedron splatting for 3d generation. *NeurIPS*, 37:80165–80190, 2025. 3

[24] Guang Han, Kang Wu, Fanyu Zeng, Jixin Liu, and Sam Kwong. Dual-stream adaptive convergent low-light image enhancement network based on frequency perception. *IEEE Transactions on Computational Imaging*, 9:1152–1164, 2023. 4

[25] Jing He, Haodong Li, Wei Yin, Yixun Liang, Leheng Li, Kaiqiang Zhou, Hongbo Liu, Bingbing Liu, and Ying-Cong Chen. Lotus: Diffusion-based visual foundation model for high-quality dense prediction. *arXiv preprint arXiv:2409.18124*, 2024. 6, 7

[26] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPSW*, 2021. 6

[27] Derek Hoiem, Alexei A. Efros, and Martial Hebert. Automatic photo pop-up. *TOG*, page 577–584, 2005. 2

[28] Derek Hoiem, Alexei A. Efros, and Martial Hebert. Recovering surface layout from an image. *IJCV*, page 151–172, 2007. 2

[29] Xin Huang, Ruizhi Shao, Qi Zhang, Hongwen Zhang, Ying Feng, Yebin Liu, and Qing Wang. Humannorm: Learning normal diffusion model for high-quality and realistic 3d human generation. In *CVPR*, pages 4568–4577, 2024. 3

[30] Satoshi Ikehata. Scalable, detailed and mask-free universal photometric stereo. In *CVPR*, pages 13198–13207, 2023. 3

[31] Varun Jampani, Kevis-Kokitsi Maninis, Andreas Engelhardt, Arjun Karpur, Karen Truong, Kyle Sargent, Stefan Popov, André Araujo, Ricardo Martin Brualla, Kaushal Patel, et al. Navi: Category-agnostic image collections with high-quality 3d shape and pose annotations. *NeurIPS*, 36:76061–76084, 2023. 2

[32] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *CVPR*, pages 9492–9502, 2024. 2

[33] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *TOG*, 42(4):139–1, 2023. 3

[34] Mukul Khanna, Yongsen Mao, Hanxiao Jiang, Sanjay Haresh, Brennan Shacklett, Dhruv Batra, Alexander Clegg, Eric Undersander, Angel X Chang, and Manolis Savva. Habitat synthetic scenes dataset (hssd-200): An analysis of 3d scene scale and realism tradeoffs for objectgoal navigation. In *CVPR*, pages 16384–16393, 2024. 6

[35] Black Forest Labs. Flux.1 dev: Open-weight text-to-image generation model. https://flux1ai.com/dev, 2025. 5, 2

[36] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. *ACM Transactions on Graphics*, 39(6), 2020. 2

[37] Weiyu Li, Jiarui Liu, Rui Chen, Yixun Liang, Xuelin Chen, Ping Tan, and Xiaoxiao Long. Craftsman: High-fidelity mesh generation with 3d native generation and interactive geometry refiner. *arXiv preprint arXiv:2405.14979*, 2024. 1, 3, 4, 6

[38] Yangguang Li, Zi-Xin Zou, Zexiang Liu, Dehu Wang, Yuan Liang, Zhipeng Yu, Xingchao Liu, Yuan-Chen Guo, Ding Liang, Wanli Ouyang, et al. Triposg: High-fidelity 3d shape synthesis using large-scale rectified flow models. *arXiv preprint arXiv:2502.06608*, 2025. 3

[39] Zhaowen Li, Xu Zhao, Chaoyang Zhao, Ming Tang, and Jinqiao Wang. Transfering low-frequency features for domain adaptation. In *ICME*, pages 01–06, 2022. 4

[40] Minghua Liu, Chong Zeng, Xinyue Wei, Ruoxi Shi, Linghao Chen, Chao Xu, Mengqi Zhang, Zhaoning Wang, Xiaoshuai Zhang, Isabella Liu, Hongzhi Wu, and Hao Su. Meshformer: High-quality mesh generation with 3d-guided reconstruction model. *arXiv preprint arXiv:2408.10198*, 2024. 3

[41] Fotios Logothetis, Ignas Budvytis, Stephan Liwicki, and Roberto Cipolla. Luces-mv: A multi-view dataset for near-field point light source photometric stereo, 2024. 6

[42] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *CVPR*, pages 9970–9980, 2024. 3

[43] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[44] Yuanxun Lu, Jingyang Zhang, Shiwei Li, Tian Fang, David McKinnon, Yanghai Tsin, Long Quan, Xun Cao, and Yao Yao. Direct2.5: Diverse text-to-3d generation via multi-view 2.5 d diffusion. In *CVPR*, pages 8744–8753, 2024. 3

[45] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research Journal*, pages 1–31, 2024. 2

[46] Yatian Pang, Tanghui Jia, Yujun Shi, Zhenyu Tang, Junwu Zhang, Xinhua Cheng, Xing Zhou, Francis EH Tay, and Li Yuan. Envision3d: One image to 3d with anchor views interpolation. *arXiv preprint arXiv:2403.08902*, 2024. 3

[47] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pages 4195–4205, 2023. 2

[48] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 3

[49] Lingteng Qiu, Guanying Chen, Xiaodong Gu, Qi Zuo, Mutian Xu, Yushuang Wu, Weihao Yuan, Zilong Dong, Liefeng Bo, and Xiaoguang Han. Richdreamer: A generalizable normal-depth diffusion model for detail richness in text-to-3d. In *CVPR*, pages 9914–9925, 2024. 3, 1

[50] Rene Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *ICCV*, 2021. 2

[51] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *ICCV*, pages 10901–10911, 2021. 2

[52] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. 2021. 2

[53] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 6

[54] Rui Chen. Dora: Sampling and benchmarking for 3d shape variational auto-encoders. https://aruichen.github.io/Dora, 2024. 6, 8

[55] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *CVPR*, pages 84–93, 2020. 3

[56] Yawar Siddiqui, Tom Monnier, Filippos Kokkinos, Mahendra Kariya, Yanir Kleiman, Emilien Garreau, Oran Gafni, Natalia Neverova, Andrea Vedaldi, Roman Shapovalov, et al. Meta 3d assetgen: Text-to-mesh generation with high-quality geometry, texture, and pbr materials. *arXiv preprint arXiv:2407.02445*, 2024. 6

[57] Zach Smith. Thingiverse. https://www.thingiverse.com, 2025. Accessed: 2025-03-07. 2

[58] Jingxiang Sun, Cheng Peng, Ruizhi Shao, Yuan-Chen Guo, Xiaochen Zhao, Yangguang Li, Yanpei Cao, Bo Zhang, and Yebin Liu. Dreamcraft3d++: Efficient hierarchical 3d generation with multi-plane reconstruction model. *arXiv preprint arXiv:2410.12928*, 2024. 3

[59] Tencent Hunyuan3D Team. Hunyuan3d 2.0: High-resolution 3d asset generation. https://3d-models.hunyuan.tencent.com, 2025. 6

[60] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 5, 2

[61] Tripo AI. Tripo ai - create your first 3d model with text. https://www.tripo3d.ai, 2025. 6

[62] Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Thanh Nguyen, and Sai-Kit Yeung. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *ICCV*, pages 1588–1597, 2019. 2

[63] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *CVPR*, pages 12619–12629, 2023. 3

[64] Jiepeng Wang, Peng Wang, Xiaoxiao Long, Christian Theobalt, Taku Komura, Lingjie Liu, and Wenping Wang. Neuris: Neural reconstruction of indoor scenes using normal priors. In *ECCV*, pages 139–155. Springer, 2022. 3

[65] Rui Wang, David Geraghty, Kevin Matzen, Richard Szeliski, and Jan-Michael Frahm. Vplnet: Deep single view normal estimation with vanishing points and lines. In *CVPR*, 2020. 2

[66] Zehan Wang, Ziang Zhang, Tianyu Pang, Chao Du, Hengshuang Zhao, and Zhou Zhao. Orient anything: Learning robust object orientation estimation from rendering 3d models. *arXiv preprint arXiv:2412.18605*, 2024. 5, 2

[67] Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. In *ACL*, pages 893–911, 2023. 5, 2

[68] Meng Wei, Qianyi Wu, Jianmin Zheng, Hamid Rezatofighi, and Jianfei Cai. Normal-gs: 3d gaussian splatting with normal-involved rendering. *arXiv preprint arXiv:2410.20593*, 2024. 3

[69] Kailu Wu, Fangfu Liu, Zhihan Cai, Runjie Yan, Hanyang Wang, Yating Hu, Yueqi Duan, and Kaisheng Ma. Unique3d: High-quality and efficient 3d mesh generation from a single image. *arXiv preprint arXiv:2405.20343*, 2024. 3

[70] Shuang Wu, Youtian Lin, Feihu Zhang, Yifei Zeng, Jingxi Xu, Philip Torr, Xun Cao, and Yao Yao. Direct3d: Scalable image-to-3d generation via 3d latent diffusion transformer. *arXiv preprint arXiv:2405.14832*, 2024. 1, 3, 4

[71] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, et al. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *CVPR*, pages 803–814, 2023. 2

[72] Yushuang Wu, Luyue Shi, Haolin Liu, Hongjie Liao, Lingteng Qiu, Weihao Yuan, Xiaodong Gu, Zilong Dong, Shuguang Cui, and Xiaoguang Han. Mvimgnet2. 0: A larger-scale dataset of multi-view images. *TOG*, 43(6):1–16, 2024. 2

[73] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *CVPR*, pages 1912–1920, 2015. 2

[74] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. *arXiv preprint arXiv:2412.01506*, 2024. 3, 4, 5, 6, 7, 2

[75] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. Icon: Implicit clothed humans obtained from normals. In *CVPR*, pages 13286–13296, 2022. 3

[76] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J Black. Econ: Explicit clothed humans optimized via normal integration. In *CVPR*, pages 512–523, 2023. 3

[77] Guangkai Xu, Yongtao Ge, Mingyu Liu, Chengxiang Fan, Kangyang Xie, Zhiyue Zhao, Hao Chen, and Chunhua Shen. What matters when repurposing diffusion models for general dense perception tasks? *arXiv preprint arXiv:2403.06090*, 2024. 2, 6, 7, 1

[78] Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024. 3

[79] Yuezhi Yang, Qimin Chen, Vladimir G. Kim, Siddhartha Chaudhuri, Qixing Huang, and Zhiqin Chen. Genvdm: Generating vector displacement maps from a single image, 2025. 3

[80] 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. *TOG*, 43(6):1–18, 2024. 2, 6, 7

[81] Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Chenming Zhu, Zhangyang Xiong, Tianyou Liang, et al. Mvimgnet: A large-scale dataset of multi-view images. In *CVPR*, pages 9150–9161, 2023. 2

[82] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *NeurIPS*, 35:25018–25032, 2022. 3

[83] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. 2, 4, 1

[84] Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. Clay: A controllable large-scale generative model for creating high-quality 3d assets. *TOG*, 43(4):1–20, 2024. 1, 3, 4, 6

[85] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Yan Yan, Nicu Sebe, and Jian Yang. Pattern-affinitive propagation across depth, surface normal and semantic segmentation. In *CVPR*, pages 4106–4115, 2019. 2

[86] Zibo Zhao, Zeqiang Lai, Qingxiang Lin, Yunfei Zhao, Haolin Liu, Shuhui Yang, Yifei Feng, Mingxin Yang, Sheng Zhang, Xianghui Yang, et al. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation. *arXiv preprint arXiv:2501.12202*, 2025. 6

[87] Xin-Yang Zheng, Hao Pan, Yu-Xiao Guo, Xin Tong, and Yang Liu. Mvdˆ 2: Efficient multiview 3d reconstruction for multiview diffusion. In *SIGGRAPH*, pages 1–11, 2024. 3

Hi3DGen: High-fidelity 3D Geometry Generation from Images via Normal Bridging

Supplementary Material

6. More Details for the Method

More Implementation Details We reimplement GenPercept [77] using StableDiffusion 2.1 by replacing the input noise with VAE-encoded images. We implement noise injection using an EDM-style noise sampler, which randomly adds noise to the encoder output latents before they are input to the decoder. Specifically, we follow EDM to use standard parameters with $\sigma_{\min} = 0.002$ and $\sigma_{\max} = 80.0$. We guarantee the SNR of the features by selecting timestamps from 0 to 400, which we empirically found maintains coarse shape knowledge — this approach aligns with Instruct-Pix2Pix, which also adds noise in a middle timerange to avoid structural changes. To better preserve coarse knowledge while instructing the encoder to focus on detailed information, we follow ControlNet [83] to add a secondary encoder by copying the weights of the SD2.1 encoder and concatenating multi-layer features with the decoder for dual-stream training. Specifically, an image is first processed by the VAE before entering the two encoders. The encoders do not share weights since they are designed to learn different frequency information from the images.

Training Details For training our I2N method, we input identical image latents to the dual encoders, where the coarse encoder remains noise-free with no modifications to any layers, while for the fine-grained encoder, we follow the approach described above to inject noise on the encoder output. Notably, we feed the noised features to the decoder layers rather than back to the encoder layers. For domain-specific training, we first train on real-world data from the Depth-pro dataset for 50,000 steps with a batch size of 256 at 768px resolution. We randomly crop images at varying aspect ratios before resizing to 768px. Subsequently, we train our model on rendered images from DetailVerse and Objaverse[13]. For DetailVerse, we render 40 spherical views per object using nvdiffrast, varying the radius and field of view. For Objaverse training, we use the 40-view renders from GObjaverse, applying the filter criteria from RichDreamer[49] to select 170K high-quality samples. We fine-tune in this second stage on the synthetic dataset while freezing the coarse encoder. For training our N2G method, we follow Trellis to employ rectified flow for model fine-tuning. We reuse the Sparse Structure VAE and Structured Latent VAE without modification, as DetailVerse is generated using Trellis (ensuring domain compatibility), and our selected Objaverse subset is already included in the original Trellis training dataset.

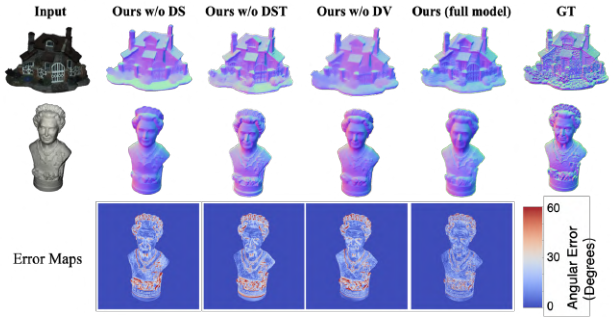

Figure S11. Ablations on image-to-normal estimation.

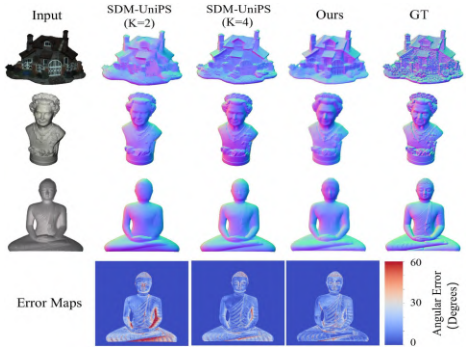

Figure S12. Qualitative comparison of image-to-normal estimation with SOTA Photometric Stereo-based Method, SDM-UniPS.

Inference Details During inference of Hi3DGen, we first utilize an off-the-shelf background removal model to isolate the foreground object. We crop the foreground and pad the image to a square format, then resize it to 768×768 resolution before inputting it to our NiRNE model. During the inference of NiRNE, we do not inject any noise into the encoder features to ensure stable inference and maximize the preservation of detail information captured by the fine-grained encoder. Given the estimated normal map, we set the background to white and input the normal map to NoRLD. Trellis employs a two-stage generation pipeline to produce structured latents, which first generates the sparse structure, followed by the local latents attached to it. Following the same approach as Trellis, we first generate the sparse structure represented by sparse voxels, then initialize noise on the sparse voxel representation to generate the final structure latents using our fine-tuned structure latents flow model. The final mesh is generated using the pre-trained mesh decoder.

Metric Explanation For comprehensive evaluation of image-to-normal, we adopt metrics from Dora [10] to quantify normal map accuracy, with particular emphasis on sharp

Table S4. Image-to-Normal estimation evaluation on Luces-MV (SNE). Comparisons of NiNRE with SOTA photometric stereo techniques. **Bold** indicates the second best results and **Red** indicates best results.

Method	Bowl	Buddha	Bunny	Cup	Die	Hippo	House	Owl	Queen	Squirrel	Ave.
SDM-UniPS (K=2)	37.65	26.24	**29.02**	23.70	**26.32**	31.45	40.68	**24.56**	27.14	26.10	29.286
SDM-UniPS (K=4)	**31.64**	**20.59**	**23.23**	**23.39**	**25.58**	**21.91**	**38.61**	**22.26**	**25.97**	**24.04**	**25.722**
Ours	**34.55**	**21.13**	30.45	**17.47**	27.20	**24.64**	**34.58**	25.15	**26.82**	**24.29**	**26.628**

edges where geometric details are most salient. Specifically, we compute the Sharp Normal Error (SNE) through a three-step process: Firstly, we detect salient regions in the ground truth normal maps through canny. Secondly, we dilate these masked regions to ensure complete coverage of edge features. Finally, we calculate the normal angle error within these masked regions. For completeness and fair comparison with existing methods, we also report the Normal Error (NE) across the entire normal map, measured in degrees. For evaluating normal-to-geometry conversion, we render normal maps from 22 fixed, evenly spaced viewpoints around each object using nvdiffrast [36], which is used to compute SNE and NE.

7. More Details for the DetailVerse

To ensure the quality of our synthesized meshes, we implement a rigorous multi-stage data generation and filtering pipeline that combines expert evaluation with automated assessment techniques.

Step 1: Semantic Text Prompt Curation We initiate the 3D data synthesis process with text prompts rather than image prompts, as textual descriptions enable more precise control over semantic diversity, thereby ensuring variety in the resulting geometries. To collect high-quality text prompts with semantic diversity, we first sourced approximately 14M raw prompts from DiffusionDB [67], covering a wide range of topics relevant to AI generation applications. We employed a LLaMA-3-8B model [60], fine-tuned with manually annotated examples, to categorize these prompts into four distinct classes: (i) Single Objects; (ii) Multiple Objects; (iii) Scenes; and (iv) Others. Only prompts from classes (i) and (ii) were retained, yielding approximately 1M high-fidelity prompt candidates.

Next, we applied rule-based filtering to preserve geometric and semantic attributes while eliminating stylistic modifiers. Empirically, we observed that input images with near-isometric viewpoints and CGI-rendered aesthetics significantly enhance the fidelity of 3D synthesis. Thus, we implemented structural prompt standardization to prompting the image generation. Specifically, we applying domain-specific prompt templates to enforce explicit geometric cues and structural clarity (e.g., "isometric perspective", "Unreal Engine 5 Rendering", "4K", "MasterPiece"). This comprehensive process yielded approximately 1.5 million well-curated and natural prompts.

Step 2: High-Quality Image Generation With our diverse text prompt collection established, the next step involved generating corresponding images suitable for 3D asset synthesis. The key requirements for these images were: (i) high visual fidelity with rich details that accurately reflect the textual descriptions; and (ii) specific viewpoints and styles that facilitate robust 3D reconstruction.

We integrated the state-of-the-art Flux.1-Dev [35] as our image generator. To ensure detailed output, we filtered the generated images by ranking their sharpness according to the number of sharp pixels, as calculated using Canny edge detection, and retained only the top 50%. For each prompt, we randomly selected a seed to encourage variety, generating exactly one image per prompt.

To mitigate geometry distortion in the resulting 3D models, we utilized OrientAnything [66], a robust object orientation estimation model, to measure the alignment between the camera view and canonical object orientation. Images with angular deviations exceeding 60° were rejected to prevent structural distortions and preserve geometric fidelity. Through this filtering process, we preserved 1 million high-quality images for the subsequent 3D synthesis stage.

Step 3: Robust Image-to-3D Synthesis We employed Trellis [74], a state-of-the-art two-stage 3D generator, to produce high-fidelity 3D objects from the prepared images. Given its superior performance with high-quality inputs, we initially generated a set of preliminary meshes.

To ensure mesh quality, we implemented a rigorous data cleaning process combining expert evaluation with automated assessment. We randomly sampled 10K meshes and engaged 10 trained experts to conduct triple-blind quality assessments. The evaluation criteria primarily focused on surface quality, specifically examining whether the rendered normal maps contained holes or noise artifacts.

Based on these expert annotations, we trained a quality assessment network using DINOv2 [45] features. Specifically, we extracted features from four equiangular rendered normal maps of each mesh and trained a three-layer MLP classifier for quality scoring. This trained network was then applied to evaluate the entire dataset. Models that received positive classifications across all four views were selected for training our NoRLD model. Through this comprehensive quality assurance process, we retained 700K high-quality object meshes to form our DetailVerse dataset. A data gallery is shown in Fig. S13, and better visualizations

are presented in the demo video.

8. More Ablation Studies

NiRNE Ablation We provide qualitative results to supplement the ablation studies on the proposed NiRNE. As shown in Fig. S11, each component makes positive role in the final performance.

9. More Results

More Image-to-Normal Results We compare NiRNE with SOTA photometric stereo technique (SDM-UniPS [30]), which works in a different setup that requires input images under K different lightning conditions. (As shown in Fig. S12).

More Comparisons We give more qualitative comparisons in Fig. S14, which shows our normal-bridged Hi3DGen can achieve more consistent 3D detailed geometries with input images than existing methods. Better visualizations are presented in the demo video.

Figure S13. More DetailVerse data exhibition.

Figure S14. More 3D generation results comparison.