arXiv:2409.12193v1 [cs.CV] 18 Sep 2024

Vista3D: Unravel the 3D Darkside of a Single Image

Qiuhong Shen[1], Xingyi Yang[1], Michael Bi Mi[2], and Xinchao Wang[1]*

[1] National University of Singapore [2] Huawei Technologies Ltd
{qiuhong.shen,xyang}@u.nus.edu xinchao@nus.edu.sg

Abstract. We embark on the age-old quest: unveiling the hidden dimensions of objects from mere glimpses of their visible parts. To address this, we present **Vista3D**, a framework that realizes swift and consistent 3D generation within a mere 5 minutes. At the heart of Vista3D lies a two-phase approach: the coarse phase and the fine phase. In the coarse phase, we rapidly generate initial geometry with Gaussian Splatting from a single image. In the fine phase, we extract a Signed Distance Function (SDF) directly from learned Gaussian Splatting, optimizing it with a differentiable isosurface representation. Furthermore, it elevates the quality of generation by using a disentangled representation with two independent implicit functions to capture both visible and obscured aspects of objects. Additionally, it harmonizes gradients from 2D diffusion prior with 3D-aware diffusion priors by angular diffusion prior composition. Through extensive evaluation, we demonstrate that Vista3D effectively sustains a balance between the consistency and diversity of the generated 3D objects. Demos and code will be available at https://github.com/florinshen/Vista3D.

Keywords: 3D Generation · 3D Reconstruction · Score Distillation

1 Introduction

Since the earliest times, our ancestors gazed upon the luminous moon, a symbol of mystery and wonder. Its bright facade, an elegant sphere in the cosmos, has always made us think about what remains hidden: the moon's obscure and elusive dark side. This curiosity, as ancient as human history itself, represents our innate desire to uncover the concealed dimensions that exist beyond the visible.

This quest, once purely philosophical, has now ventured into the realm of practicality, propelled by the advancements in 3D generative model [29,34,42,45,48]. These technologies enable a broad range of applications, especially in gaming and virtual reality, allowing for the creation of rich, detailed environments and objects without extensive modeling.

Nevertheless, the development of robust large-scale 3D generative models remains a formidable challenge, predominantly due to the limited availability of 3D data. Numerous attempts [1,13,27] have been made to train 3D diffusion models on relatively small 3D datasets, condition on textual or visual prompts; Yet,

* Corresponding Author.

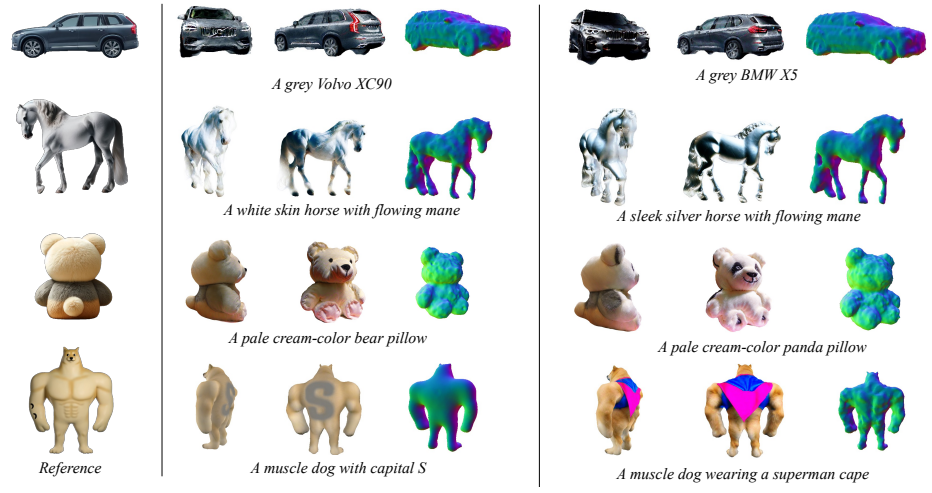

Fig. 1: 3D Darkside of Single Image. By employing various text prompts, Vista3D is capable of unveiling the diversity of unseen views while retaining 3D consistency and detail. Two novel views and the normal map are visualized for each text prompt.

these endeavors often fall short in creating 3D objects with structural integrity and textural consistency.

This challenge is further compounded in the context of reconstructing 3D objects from single images. In this context, two primary approaches emerge. The first considers the task as a problem of sparse-view reconstruction. However, this often leads to blurred 3D outputs due to the neglect of unseen elements, resulting in excessively blurred 3D objects [8, 52] as most views remain unseen.

On the other hand, the generative approach, which leverages large-scale 2D diffusion models [29, 42], introduces its own set of challenges. Efforts to develop 3D-aware 2D diffusion models [19, 21, 30, 32, 34, 39, 40, 51] involve fine-tuning 2D models with camera transformation modeling on 3D datasets [5, 6]. Nevertheless, the prevalence of synthetic objects in these datasets can lead to a compromise in 2D diversity. This often results in the generation of oversimplified geometries and textures.

In this paper, we present Vista3D, a framework designed for reconstructing the unseen view (or "darkside") from a single image. Central to Vista3D is a dual-phase strategy: a coarse phase followed by a fine phase.

In **the coarse phase**, we leverage 3D Gaussian splatting [14] to swiftly create basic geometry and textures. To stabilize Gaussian Splatting optimization, we employ a gradient-based Top-K densification strategy, focusing on Gaussian points with the highest gradients. Additionally, we introduce two novel regularization terms targeting the Gaussian scale and transmittance values, significantly enhancing the convergence speed.

The fine phase then transforms this initial geometry into signed distance fields (SDF) for further optimization. Here, we employ FlexiCubes [38], an advanced differentiable isosurface technique, to refine the geometry. This refine-

ment aids in learning the signed distance fields (SDFs), deformation, and interpolation weights. The parameters are optimized by ensuring fidelity to the original image and guided by a score function derived from diffusion priors.

Despite these advancements, a unified representation and supervision across all views, both seen and unseen, prove insufficient for capturing the unique characteristics of different viewpoints and generating diverse, consistent 3D objects. To address this, we enhance the representation by implementing *Disentangled Texture Representation*, using two angularly disentangled networks for accurate texture prediction. Furthermore, our *Angular-based Composition* method amalgamates different diffusion priors, adjusting their gradients within specific angular bounds according to their gradient magnitudes. This strategic adjustment assures 3D consistency while promoting diversity in the unseen views.

Vista3D excels in efficiently generating diverse and consistent 3D objects from a single image within five minutes. Our extensive evaluations demonstrate its ability to maintain a flexible balance between the consistency and diversity of the generated 3D objects.

We summarize our contribution as follows:

- We present Vista3D, a framework for revealing the 3D darkside of single images, efficiently generating diverse 3D objects using 2D priors.
- We develop a transition from Gaussian Splatting to isosurface 3D representations, refining coarse geometry with a differentiable isosurface method and disentangled texture for textured mesh creation.
- We propose an angular composition approach for diffusion priors, constraining their gradient magnitudes to achieve diversity on the 3D darkside without sacrificing 3D consistency.

2 Related-works

2.1 3D Generation Conditioned on a Single Image

The objective of image-to-3D generation is to create 3D objects from a single reference image. Initial methods [8, 52] approached this challenge as a variant of sparse view 3D reconstruction. However, these methods often resulted in blurred object outputs due to insufficient priors. Recently, drawing inspiration from text-to-3D initiatives that utilize Score Distillation Sampling (SDS) to elevate 2D diffusion priors into 3D generative models, image-to-3D works [24, 33, 34, 40, 42] have adopted a similar approach for 3D object generation based on a single image. However, 2D diffusion priors alone cannot ensure 3D consistency, as they are typically trained solely on image datasets. To address this, several studies [19–21, 39] have attempted to refine 2D diffusion priors with 3D data [5, 6], enhancing their ability to model 3D consistency. A notable example is Zero-1-to-3, which can generate novel views condition on single image and camera position. Integrating this refined model with SDS [30, 41] allows for the reconstruction of coherent 3D objects. Moreover, another stream of works [9, 17, 36, 46, 47, 50, 55] pretrained on large-scale 3D dataset [5] directly predicting the representation

of a 3D object from a single image. Diverging from previous works, our work does not solely view this as a 3D reconstruction issue. We redefine it as a 3D generation task aimed at uncovering the unseen 3D aspects behind a single image. Through a meticulously crafted framework, our method efficiently generate diverse and consistent 3D objects.

2.2 3D Representations for Generation

Presently, most zero-shot text-to-3D and image-to-3D models utilize an optimization based pipeline, parameterizing the 3D object as a differentiable representation, which varies among different methods. The most prevalent representation in groundbreaking works like dreamfields [12], dreamfusion [29], and SJC [43] is Neural Radiance Fields (NeRF) [25]. However, training a NeRF is computationally intensive and takes long time to convergence. Magic3D [16] introduced a two-stage representation, initially learning a coarse NeRF, followed by refining the polygon mesh using a differentiable isosurface method, DMTet [37]. Fantasia3D [2] suggested directly optimizing DMTet [37] in separate phases for geometry and texture, but this often leads to mode collapse in the geometry phase and extends training time beyond NeRF. Gaussian Splatting [10, 14, 35, 44, 53] has gained attention for its efficiency in various 3D tasks, with several 3D generative models [3, 4, 41, 49] incorporating it for effective generation. However, as a point-based representation, it cannot yield high-fidelity meshes. In our approach, we employ Gaussian Splatting exclusively to create coarse geometry. This coarse geometry is then transformed into SDF, optimized with a hybrid isosurface representation, FlexiCubes [38], to produce high-fidelity meshes. Additionally, we propose an angular disentangled texture representation, tailored to the specifics of this task.

3 Methodology

In this section, we outline our framework to generate detailed 3D object from single image with 2D diffusion priors. As depicted in Figure 2, our exploration of the 3D darkside of a single image commences with the efficient generation of basic geometry (Section 3.1), represented through 3D Gaussian Splatting. In refinement stage (Section 3.2), we devise a method for transforming the rudimentary 3D Gaussian geometry into signed distance fields, and thereafter, we introduce a differentiable isosurface representation to further enhance the geometry and textures. To enable diverse 3D darkside of given single image, we present a novel approach to constrain two diffusion priors (Section 3.3), enabling the creation of varied yet coherent darkside textures by bounding gradient magnitude. With these approaches, our method can efficiently generate diverse, high-fidelity meshes from a single image.

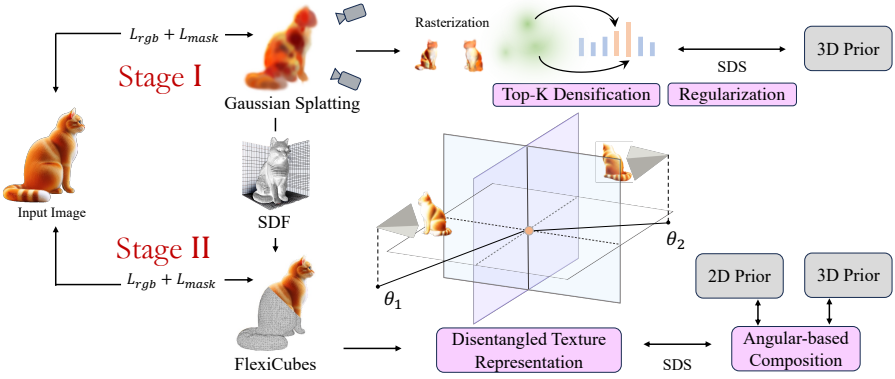

Fig. 2: Overview of Vista3D. We generate high-fidelity mesh from single image input in a coarse-to-fine manner. In the coarse stage, we utilize Gaussian Splatting to learn a coarse geometry with a 3D-aware 2D diffusion prior. We further extract sign distance fields from Gaussian Splatting for refinement. Another 2D diffusion prior is enabled with an angular-based composition to explore diverse darkside while retain 3D consistency in refinement stage.

3.1 Coarse geometry from Gaussian Splatting

In the coarse stage of our framework, we focus on constructing a basic object geometry using Gaussian Splatting. This technique, as described in [14], represents 3D scenes as set of anisotropic 3D Gaussians. Compared to other neural inverse rendering methods, such as NeRF [25, 26], Gaussian Splatting demonstrates a notably faster convergence speed in inverse rendering tasks.

Some works [3, 41, 49] has attempted to introduce Gaussian Splatting into 3D generative models. In these methods, we found that directly using Gaussian splatting to generate detailed 3D objects requires optimizing a large number of 3D Gaussians, necessitating significant time for optimization and densification, which is still time-consuming. However, Gaussian Splatting can quickly create a coarse geometry from a single image using a limited number of 3D Gaussians within just one minute. Therefore, in our approach, we utilize Gaussian Splatting solely for the initial coarse geometry generation.

Specifically, each 3D Gaussians is parameterized by its central position $x \in \mathbb{R}^3$, scaling $r \in \mathbb{R}$, rotation quaternion $q \in \mathbb{R}^4$, opacity $\alpha \in \mathbb{R}$, and spherical harmonics $c \in \mathbb{R}^3$ to represent color. To generate a coarse 3D object, we optimize a set of these Gaussian parameters $\Psi = \{\Phi_i\}$, where $\Phi_i = \{x_i, r_i, q_i, \alpha_i, c_i\}$. To render 3D Gaussians to 2D images, we utilized the highly-optimized tile based rasterization implementation [14].

To generate the coarse geometry of given single image I_{ref}, we adopt Zero-1-to-3 XL [5, 19] as 2D diffusion priors ϵ_ϕ with pretrained parameters ϕ. This prior enables denoising of novel views based on the given image I_{ref} and relative camera pose $\Delta\pi$. Accordingly, we optimize the 3D Gaussians Ψ with SDS [29]:

$$\nabla_\Psi \mathcal{L}_{SDS} = \mathbb{E}_{t,\epsilon}\left[\left(\epsilon_\phi\left(I_R^\pi; t, I_{ref}, \Delta\pi\right) - \epsilon\right)\frac{\partial I_R^\pi}{\partial \Psi}\right] \tag{1}$$

where π denotes the camera pose sampled around the object with fixed camera radius and FoV, I_R^π is the rendered image from 3D Gaussian set Ψ with camera pose π, timestep t is annealed to weight the gaussian noise ϵ added to the rendered image. Beyond this basic approach, we introduce a Top-K Gradient-based Densification strategy to accelerate convergence and add two regularization terms to enhance the reconstructed geometry.

Top-K Gradient-based Densification. In the optimization process, we find the periodical densification [14] with naive gradient threshold is hard to tune due to the nature randomness of SDS. So we instead use a more robust densification strategy. Only gaussians points with top-k gradients will be densified during each interval, this simple strategy can stablize training cross various given images.

Scale & Transmittance Regularization. Additionally, We add two regularization terms to encourage Gaussian Splatting to learn more detailed geometry in this phase. A scale regularization is introduced to avoid too large 3d gaussians, and another transmittance regularization is adopted to encourage the geometry learning from transparent to solid. The overall loss function in this stage can be written as:

$$\begin{aligned}\nabla_\Psi \mathcal{L}_{\text{coarse}} = \lambda_{SDS}\nabla_\Psi \mathcal{L}_{SDS} + \lambda_{rgb}\nabla_\Psi \mathcal{L}_{rgb} \\ + \lambda_{mask}\nabla_\Psi \mathcal{L}_{mask} + \underbrace{\lambda_{\text{scale}}\nabla_\Psi \sum_i \|s_i\|}_{\text{Scale Regularization}} \\ \underbrace{- \lambda_{\text{tr}}\nabla_\Psi \min(\tau, \frac{1}{N_{fg}}\sum_k T_k)}_{\text{Transmittance Regularization}};\end{aligned} \tag{2}$$

where $\mathcal{L}_{rgb}$ and $\mathcal{L}_{mask}$ are two MSE loss computed between the rendered reference view and the given image. The term $T_k = \sum_i \alpha_i \prod_{j=1}^{i-1}(1-\alpha_j)$ denotes the transmittance value for the k-th pixel in I_R^π, where N_{fg} is the total number of foreground pixels. Additionally, τ serves as a hyperparameter that is gradually annealed from 0.4 to 0.9, effectively regularizing transmittance over time.

3.2 Mesh refinement and texture disentanglement

In the refinement stage, our focus shifts to transforming the coarse geometry, produced via Gaussian splatting, into signed distance fields (SDF) and refining its parameters using a hybrid representation.

This stage is crucial for overcoming the challenges presented in the coarse stage, notably the surface artifacts frequently introduced by Gaussian splatting. Due to the inability of Gaussian splatting to provide direct estimates of surface normals, we cannot employ traditional smoothing methods to alleviate these artifacts. To counter this, our method incorporates a hybrid mesh representation, which entails modeling the 3D object's geometry as a differentiable isosurface and learning the texture using two distinct, disentangled networks. This dual approach not only smooths out the surface irregularities but also significantly improves the fidelity and overall quality of the 3D model.

Geometry representation. We utilize FlexiCubes to represent the geometry in our approach. FlexiCubes is a differentiable isosurface representation which allow local flexible adjustments to the extracted mesh geometry and connectivity [38]. The geometry of an object is depicted as a deformable voxel grid with learnable weights. Deformation $\delta_i \in \mathbb{R}^3$ and sign distance field (SDF) $s_i \in \mathbb{R}$ is learnt for every vertices v_i in the voxel grid. And interpolation weights $\beta \in \mathbb{R}^{20}$ and splitting weights $\gamma \in \mathbb{R}$ are learnt for each grid cell to position dual vertices and control quadrilaterals splitting. Triangle meshes can be extracted from it differentiablely through Dual Marching Cubes [28]. To bridge the gap between the learned coarse geometry and the isosurface representation, we initially extract a density field from Gaussian splattings using local density queries [41], followed by the application of marching cubes [22] to extract a base mesh M_{coarse}. Subsequently, we query this base mesh at grid vertices v_i to obtain the initial Signed Distance Field (SDF) $s(v_i)$. For stable optimization, the queried SDF is then scaled as follows:

$$s(v_i) = \frac{\xi \cdot s(v_i)}{\max\{|s_j| : s_j \in S, s_j < 0\}}, \text{ where } S = \{s_i\} \tag{3}$$

where $s_j < 0$ indicates the field within the object. The scale factor ξ linearly increases from 1 to 3 during the optimization process.

Disentangled Texture Representation. For texture learning, we employ hash encoding followed by a MLP to directly learn albedo. However, distinct from text-to-3D tasks, we recognize two primary supervision sources in this task: the provided reference image and the SDS gradient from 2D Diffusion priors. Typically, a substantial loss weight λ_{rgb} is assigned for the reference image. This dominant reference image supervision can decelerate the convergence of textures in unseen views, particularly when unseen views significantly differ from the reference view.

To address this, we separate the texture into two hash encoding, utilizing a ratio that combines with the relative azimuth angle $\Delta\theta = \theta_\pi - \theta_{ref}$, where θ_π represents the azimuth of the sampled camera pose π, and θ_{ref} is the azimuth of the reference image. The hash encoding for a given query point κ in the rasterized triangle mesh is expressed as:

$$E = (1-\eta)H_{back}(\kappa) + \eta H_{ref}(\kappa) \tag{4}$$

where H_{ref} and H_{back} denote learnable hash encoding facing forward and back, $\eta = (\cos(\Delta\theta) + 1)/2$ is the balance factor that varies with the sampled azimuth angle. Then the encoded feature E is fed into a MLP predict albedo values.

With these geometry and texture representation, we can render the 3D object to images by memory-efficient rasterization coupled with lambertian shading. Above learnable parameters Θ is refined with $\nabla_\Theta \mathcal{L}_{\text{refine}}$:

$$\begin{aligned}\nabla_\Theta \mathcal{L}_{\text{refine}} = \lambda_{SDS}\nabla_\Theta \mathcal{L}_{SDS} \\ + \lambda_{SDF}\nabla_\Theta \mathcal{L}_{SDF} + \lambda_{\text{consistency}}\nabla_\Theta \mathcal{L}_{consistency} \\ + \lambda_{rgb}\lambda_{SDS}\nabla_\Theta \mathcal{L}_{rgb} + \lambda_{mask}\nabla_\Theta \mathcal{L}_{mask};\end{aligned} \tag{5}$$

where the $\mathcal{L}_{SDF}$ is a simple SDF regulariztion term to avoid floaters, $\mathcal{L}_{consistency}$ is a smooth loss applied on surface normals [16,24], $\mathcal{L}_{rgb}$ and $\mathcal{L}_{mask}$ are two MSE loss between the rendered reference view and the given image.

3.3 Darkside Diversity via Prior Composition

In implementing our pipeline, we encountered a key challenge related to the lack of diversity in unseen views. This issue largely stems from the reliance on the Zero-1-to-3 XL prior, a model trained on synthetic 3D objects from Objaverse-XL [5]. While this prior is adept at handling 3D-aware generation based on reference images and relative camera poses, it tends to produce oversimplified or overly smooth results in unseen views. This limitation becomes especially pronounced when dealing with objects captured in the real world.

To address this, we integrate an additional prior from Stable-Diffusion, known for its ability to synthesize diverse images.

Darkside diversification with 2D diffusion. We introduce a second prior, ϵ_ρ with pretrained parameters ρ, leading to two Score Distillation Sampling (SDS) loss terms $\nabla\mathcal{L}^{\phi}_{SDS}$ and $\nabla\mathcal{L}^{\rho}_{SDS}$ (Equation 1) for optimization. The optimal balance between these two priors remains relatively unexplored. While Magic123 [30] uses an empirical loss weight of 1/40 for the latter term, this approach may not fully harness the potential of the 2D prior. The key objective in introducing this 2D prior is to introduce greater diversity in unseen view. A small weight with $\nabla\mathcal{L}^{\rho}_{SDS}$ may largely limit its effect.

To enhance the diversity in the unseen aspects of the given image, we employ a gradient constrain method to merge these two priors. We reformulate the SDS loss as a score function [29], $\nabla_\Theta\mathcal{L}_{SDS}(\phi, \mathbf{x}) = -\mathbb{E}_{t,\mathbf{z}_t|\mathbf{x}}\nabla_\Theta \log p_\phi(\mathbf{z}_t|y)$, where t is the timestep and z_t is noise latent.

Here $\nabla\mathcal{L}^{\phi}_{SDS}$ is a 3D-aware term conditioned on $y = \{\Delta\pi, I_{ref}\}$, while $\nabla\mathcal{L}^{\rho}_{SDS}$ is a diverse text-to-image term conditioned on text prompt $y = P_T$. With different condition y, the score function of these two SDS term varies. To retain 3D consistency of unseen views, the magnitude of $\nabla_\Theta \log p_\rho(\mathbf{z}_t|y)$ need to be constrained with respect to the 3D-aware term $\nabla_\Theta \log p_\phi(\mathbf{z}_t|y)$. And to avoid the texture to be over-smoothed by the 3D-aware diffusion model, the magnitude of $\nabla_\Theta \log p_\phi(\mathbf{z}_t|y)$ is indeed to be constrained with the $\nabla_\Theta \log p_\rho(\mathbf{z}_t|y)$ term.

Angular-based Score Composition. Since the noise latents $\mathbf{z}_t$ in both priors have different encoding spaces, direct evaluation of their magnitudes using the predicted noise difference $\epsilon_\rho - \epsilon$ is not feasible. Instead, we evaluate the magnitude of these terms by observing their gradient on the rendered image $\mathbf{x}$, specifically $\nabla_\mathbf{x}\mathcal{L}_{SDS}$. Consequently, we establish upper and lower bounds for the gradient magnitude ratio of these two SDS terms, allowing for a more accurate and feasible evaluation method:

$$B_{lower}(\eta, \iota) \leq G = \frac{||\nabla_\mathbf{x}\mathcal{L}^{\rho}_{SDS}||_2}{||\nabla_\mathbf{x}\mathcal{L}^{\phi}_{SDS}||_2} \leq B_{upper}(\eta, \iota) \tag{6}$$

When this ratio exceeds B_{upper}, we adjust the magnitude of $\nabla_\mathbf{x}\mathcal{L}^{\rho}_{SDS}$ using the factor B_{upper}/G. Conversely, if the ratio falls below B_{lower}, we scale the

magnitude of $\nabla_{\mathbf{x}}\mathcal{L}^{\phi}_{SDS}$ using G/B_{lower}. And this B_{upper} and B_{lower} are regulated by the balance factor η, influenced by the camera pose, and by iterations ι, facilitating a balance between diversity and 3D consistency.

4 Experiments

4.1 Implementation Details

Coarse geometry learning. In this phase, the input image undergoes preprocessing with SAM [15, 23, 34], where the object is extracted and recentered. We initialize all 3D Gaussians with an opacity of 0.1 and a grey color, confined within a sphere of radius 0.5. The rendering resolution is progressively increased from 64 to 512. This stage involves a total of 500 optimization steps, with the densification and pruning of 3D Gaussians occurring every 100 iterations. The top-K densification starts at a ratio of 0.5 and gradually anneals to 0.1, while the pruning opacity remains constant at 0.1. After the first densification, transmittance regularization is activated and selectively applied to the top-80% opacity values of 3D Gaussians to avoid affecting transparent Gaussians. Scale regularization is enforced using L_1 norm. The weights of λ_{scale} and λ_{tr} are maintained at 0.01 and 1, respectively, throughout the optimization, whereas λ_{rgb} and λ_{mask} are gradually increased from 0 to 10000 and 1000, respectively. The timestep for SDS is linearly annealed from 980 to 20. For camera pose sampling, the azimuth is sampled in the range of $[-180, 180]$ and elevation in $[-45, 45]$, with a fixed radius of $r = 2$. This phase of optimizing the coarse geometry takes about 30 s.
Mesh refinement. In the refinement phase, we configure the grid size of FlexiCubes to 80^3 within the space $[-1, 1]^3$. The coarse geometry obtained from the initial stage is recentered and rescaled to initialize the Signed Distance Field (SDF) for the vertices of this grid. Interpolation weights are set to 1, and all deformations start at 0. For texture, we use two hash encodings with a two-layer Multilayer Perceptron (MLP). The batch size is maintained at 4. The learning rate for deformation and interpolation weights is 0.005, while it's 0.001 for SDF, and 0.01 for texture parameters. The rendering resolution is gradually increased from 64 to 512. In Equation 5, the loss weights are set as follows: $\lambda_{rgb} = 1500$, $\lambda_{mask} = 5000$, $\lambda_{sdf} = 1$, and $\lambda_{SDS} = 1$. We develop two versions for optimization: **Vista3D-S** and **Vista3D-L**. **Vista3D-S** performs 1000 steps of optimization solely with the 3D-aware prior, aiming to generate 3D mesh within 5 minutes. **Vista3D-L** undergoes 2000 steps of optimization with two diffusion priors to create more detailed 3D objects. The entire optimization process for Vista3D ranges from 15 to 20 minutes. In this stage, camera poses are sampled using a 3D-aware Gaussian unsampling strategy to expedite convergence (additional details are provided in the supplementary material). All experiments are conducted on an RTX3090 GPU.
Score distillation sampling. In SDS optimization, the practice of linearly annealing the timestep t to adjust the noise level has been established as effective for producing higher-quality 3D objects [11]. However, in our experiments, we observed that linear annealing may not be the optimal strategy. Consequently, we

have implemented an interval annealing approach. In this approach, the timestep t is randomly sampled from an annealing interval rather than adhering to a fixed linear progression. This strategy has been found to effectively mitigate the artifacts commonly observed with linear annealing.

Fig. 3: Qualitative Comparison on image-to-3D generation. We compare our Vista3D-S with DreamGaussian [41], and Magic123 [30]. Vista3D-S only takes 5 minutes to reconstruct single 3D object, yielding competitive geometry and more consistent textures compared to Magic123 [30] with 20× speedup.

Angular diffusion prior composition. In our model, we utilize two diffusion models: Zero-1-to-3 XL [5,19] and the Stable-Diffusion model [31]. For the Stable-Diffusion model, the timestep t is scaled by the factor η to ensure consistency with the reference view. When editing with both diffusion priors, we start with a large initial upper bound $B_{upper} = 100$, which is linearly annealed to 10 across optimization iterations. For front-facing views, where $\eta > 0.75$, we adjust the upper bound using the factor $(1-\eta)$. The lower bound is specifically implemented for unseen views with $\eta < 0.5$, and its range is gradually reduced from 10 to 1 during the optimization process. For enhancements using the diffusion prior, we apply tighter constraints, with B_{upper} being reduced from 2 to 0.5. The text prompts utilized for the Stable-Diffusion model are derived from the image captions generated by GPT-4.

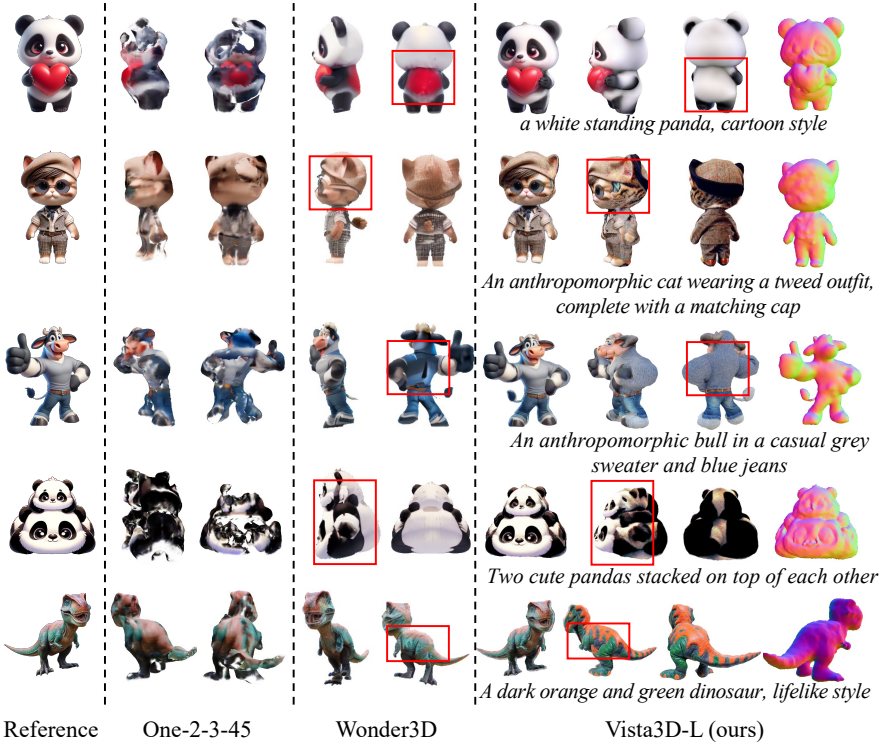

Fig. 4: Qualitative Comparison with One-2-3-45 [18] and Wonder3D [21]. In this comparison, we render two views of each 3D object as generated by One-2-3-45 and Wonder3D. For Vista3D-L, we detail the text prompts utilized for the generation of each 3D object, showcasing three rendered views alongside a single normal map for a comprehensive comparison.

4.2 Qualitative Comparison

In Figure 3, we show our efficient Vista3D-S is capable of generating competitive 3D objects with a 20× speedup compared to existing coarse-to-fine methods. For Vista3D-L, as depicted in Figure 1 and Figure 4, we highlight our angular gradient constraint which distinguishes our framework from previous image-to-3D methods, as it can explore the diversity of the backside of single images without sacrificing 3D consistency. In Figure 3, we primarily compare our Vista3D-S with two baselines, Magic123 [30] and DreamGaussian [41], for generating 3D objects from a single reference view. Regarding the quality of generated 3D objects, our method outperforms these two methods in terms of both geometry and texture. Regarding Vista3D-L, we compare it with two inference-only single view reconstruction models, specifically One-2-3-45 [18] and Wonder3D [21]. As shown in Fig. 4, One-2-3-45 tends to produce blurred texture and may result in incomplete geometry for more complex objects, while our Vista3D-L achieves more refined textures, particularly on the backside of 3D objects, using user-specified text prompts. And Wonder3D often resorts to simpler textures due to its primary training on synthetic datasets [5], which occasionally leads to out-of-

distribution issues for certain objects. In contrast, Vista3D-L offers zero-shot 3D object reconstruction by controlling two diffusion priors, enabling more detailed and consistent textural. Moreover, given that only a single reference view of the object is provided, we posit that the object should be amenable to editing during optimization with user-specified prompts. To illustrate this, we display several results in Figure 1 that emphasize the potential for editing.

	Type	CLIP-Similarity ↑	Time Cost ↓
One-2-3-45 [18]	Inference	0.594	45 s
Point-E [27]	Inference	0.587	78 s
Shape-E [13]	Inference	0.591	27 s
Zero-1-to-3 [19]	Optimization	0.778	30 min
DreamGaussian [41]	Optimization	0.738	2 min
Magic123 [30]	Optimization	0.802	2 h
DreamCraft3D [40]	Optimization	0.842	3.5 h
Vista3D-S	Optimization	0.831	5 min
Vista3D-L	Optimization	0.868	15 min

Table 1: Quantitative Comparisons on generation quality in terms of CLIP-Similarity for image-to-3D task. Average generation time is reported.

4.3 Quantitative Comparison

In our evaluation, we employ the CLIP-similarity metric [19, 24, 30] to assess the performance of our method in 3D reconstruction using the RealFusion [24] dataset, which comprises 15 diverse images. Consistent with the settings used in previous studies, we sample 8 views evenly across an azimuth range of $[-180, 180]$ degrees at zero elevation for each object. The cosine similarity is then calculated using the CLIP features of these rendered views and the reference view. Table 1 highlights that Vista3D-S attains a CLIP-similarity score of 0.831, with an average generation time of just 5 minutes, thereby surpassing the performance of the Magic123 [30]. Furthermore, when compared to another optimization-based method, DreamGaussian [41], Vista3D-S may take longer at 5 minutes, but it significantly improves consistency, as evidenced by the higher CLIP-Similarity score. For Vista3D-L, we apply an enhancement-only setting. By employing angular diffusion prior composition, our method achieves a higher CLIP-Similarity of 0.868. The capabilities of Vista3D-L, especially in generating objects with more detailed and realistic textures through prior composition, are demonstrated in Figure 4. Additionally, we conduct quantitative experiments on the Google Scanned Object (GSO) [7] Dataset, following the setting in SyncDreamer [20]. We evaluate each method using 30 objects and computed PSNR, SSIM, and LPIPS [54] between the rendered views of the 3D object and 16 ground-truth anchor views. The results, as shown in Tab. 2, reveal that our Vista3D-L achieves SOTA performance among these methods with a large margin. Vista3D-S also demonstrates competitive performance, albeit with a single diffusion prior.

	PSNR ↑	SSIM ↑	LPIPS ↓
RealFusion [24]	15.26	0.722	0.283
Make-it-3D [42]	15.79	0.741	0.245
Zero-1-to-3 [19]	18.93	0.779	0.166
One-2-3-45 [18]	17.47	0.768	0.184
SyncDreamer [20]	20.05	0.798	0.146
DreamGaussian [41]	23.43	0.832	0.092
Magic123 [30]	24.89	0.875	0.084
Vista3D-S	25.42	0.912	0.073
Vista3D-L	26.31	0.929	0.062

Table 2: Quantitative Comparison on the GSO [7] dataset

4.4 User study

In our user study, we evaluate reference view consistency and overall 3D model quality [41]. The evaluation encompasses four methods: DreamGaussian [41], Magic123 [30], and our own Vista3D-S and Vista3D-L. We recruited 10 participants for this user study. Each was asked to sort generated 3D object from different methods in terms of view consistency and overall quality respectively. Thus, the scores presented for each metric range from 1 to 4. The results, presented in Table 3, reveal that our Vista3D-S outperforms the previous methods in both view consistency and overall quality. Furthermore, the adoption of the angular prior composition in Vista3D-L leads to additional improvements in both the consistency and quality of the generated 3D objects.

	DreamGaussian [41]	Magic123 [30]	Vista3D-S	Vista3D-L
View Consistency ↑	1.78	2.11	2.87	3.24
Overall Quality ↑	2.02	1.83	2.81	3.33

Table 3: User study of Vista3D. We conduct user study in terms of view consistency and overall quality, the score ranges from 1 to 4, the higher the better.

4.5 Ablation Study

Coarse-to-fine framework. Our framework integrates a coarse stage to learn initial geometry then a fine stage to refine geometry and shade textures. We validate the necessity of such a coarse-to-fine pipeline in Figure 5 (a). We first commence with isosurface representation to learn geometry directly, finding the geometry optimization is prone to collapse without preliminary geometry initialization. Thus, a coarse initialization becomes imperative. Beside, we present the normal map of a rough mesh extracted from 3DGS from the coarse stage. It is observed that the coarse stage tends to generate rough even non-watertight geometry, both difficult to mitigate. These findings demonstrate that combining both stages is crucial for the optimal performance of Vista3D.

Disentangled Texture. For validating the effectiveness of the disentangled texture, we compare adopting both hash encodings with single hash encoding

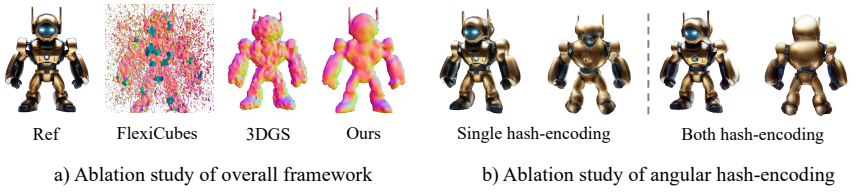

Fig. 5: Ablation study of overall framework and disentangled texture.

in Figure 5 (b). With both hash-encodings, the artifacts on the reconstructed robot are notably reduced, especially at the backside. Further, we visualize the disentangled texture in supplementary Figure 6(b). Specifically, when visualizing H_{ref}, H_{back} is set as 0 in Equation 4, and vice versa. From the shown visualization, we can clearly find that the facing-forward hash encoding H_{ref} mainly encodes the detail features consistent with the given reference view. While the back hash encoding H_{back} mainly encodes the features in the unseen views. The textures of the facing-forward view and back views are disentangled and learned in two separate hash encodings, which can facilitate learning better textures near the reference view and in unseen views.

5 Conclusion

In this paper, we present a coarse-to-fine framework Vista3D to delve into the 3D darkside of a single input image. This framework facilitates user-driven editing through text prompts or enhances generation quality using image captions. The generation process begins with a coarse geometry obtained through Gaussian Splatting, which is subsequently refined using an isosurface representation complemented by disentangled textures. The design of these 3D representations enables the generation of textured meshes within a mere 5 minutes. Additionally, the angular composition of diffusion priors empowers our framework to reveal the diversity of unseen views while maintaining 3D consistency. Our approach surpasses previous methods in terms of realism and detail, striking an optimal balance between generation time and the quality of the textured mesh. We hope our contributions will inspire future advancements and foster future exploration into the 3D darkside of single images.

Acknowledgement

This project is supported by the Ministry of Education, Singapore, under its Academic Research Fund Tier 2 (Award Number: MOE-T2EP20122-0006), and the National Research Foundation, Singapore, under its Medium Sized Center for Advanced Robotics Technology Innovation.