Magic3D: High-Resolution Text-to-3D Content Creation

Chen-Hsuan Lin*, Jun Gao*, Luming Tang*, Towaki Takikawa*, Xiaohui Zeng*,
Xun Huang, Karsten Kreis, Sanja Fidler†, Ming-Yu Liu†, Tsung-Yi Lin*
NVIDIA Corporation, Published in CVPR 2023

https://arxiv.org/abs/2211.10440

목차

1. 논문선정이유

● **주제 : 3D컨텐츠생성모델**

*** 3D 콘텐츠 제작의 중요성** .
3D 콘텐츠는 사실감과 상호작용성을 높여
사용 경험을 향상하며 영화제작 등
다양한 분야에서 필수적

*** 3D 데이터 부족 문제**
- 정교한 3D 콘텐츠 제작에
많은 시간과 비용 필요.
- 이미지 및 비디오 콘텐츠에 비해
3D 데이터 접근성 낮음.

“text to 3D generation”

- 고품질 3D 콘텐츠 생성의 혁신적 접근
- 텍스트 프롬프트를 기반 고품질 3D 모델 생성.
- 초보자도 쉽게 3D 콘텐츠 제작 가능

● **강의주제**

“Image Generative AI”
- **Image Generative Model**
- AutoEncoder
- Gan (DCGAN)
- **Diffusion Model**

* Magic3D 논문

Magic3D: High-Resolution Text-to-3D Content Creation

Chen-Hsuan Lin* Jun Gao* Luming Tang* Towaki Takikawa* Xiaohui Zeng*
Xun Huang Karsten Kreis Sanja Fidler† Ming-Yu Liu† Tsung-Yi Lin

NVIDIA Corporation

https://research.nvidia.com/labs/dir/magic3d

Abstract

DreamFusion [31] has recently demonstrated the utility of a pre-trained text-to-image diffusion model to optimize Neural Radiance Fields (NeRF) [23], achieving remarkable text-to-3D synthesis results. However, the method has two inherent limitations: (a) extremely slow optimization of NeRF and (b) low-resolution image space supervision on NeRF, leading to low-quality 3D models with a long processing time. In this paper, we address these limitations by utilizing a two-stage optimization framework. First, we obtain a coarse model using a low-resolution diffusion prior and accelerate with a sparse 3D hash grid structure. Using the coarse representation as the initialization, we further optimize a textured 3D mesh model with an efficient differentiable renderer interacting with a high-resolution latent diffusion model. Our method, dubbed Magic3D, can create high quality 3D mesh models in 40 minutes, which is 2× faster than DreamFusion (reportedly taking 1.5 hours on average), while also achieving higher resolution. User studies show 61.7% raters to prefer our approach over DreamFusion. Together with the image-conditioned generation capabilities, we provide users with new ways to control 3D synthesis, opening up new avenues to various creative applications.

Image content creation from text prompts [2, 28, 33, 36] has seen significant progress with the advances of diffusion models [13, 41, 42] for generative modeling of images. The key enablers are large-scale datasets comprising billions of samples (images with text) scrapped from the Internet and massive amounts of compute. In contrast, 3D content generation has progressed at a much slower pace. Existing 3D object generation models [4, 9, 47] are mostly categorical. A trained model can only be used to synthesize objects for a single class, with early signs of scaling to multiple classes shown recently by Zeng *et al.* [47]. Therefore, what a user can do with these models is extremely limited and not yet ready for artistic creation. This limitation is largely due to the lack of diverse large-scale 3D datasets — compared to image and video content, 3D content is much less accessible on the Internet. This naturally raises the question of whether 3D generation capability can be achieved by leveraging powerful text-to-image generative models.

Recently, DreamFusion [31] demonstrated its remarkable ability for text-conditioned 3D content generation by utilizing a pre-trained text-to-image diffusion model [36] that generates images as a strong image prior. The diffusion model acts as a critic to optimize the underlying 3D representation. The optimization process ensures that rendered

CVPR 2023.06

975회 인용 (2024.12.03 기준)

NVIDIA corporation에서 발표,
텍스트 기반 고해상도 3D 콘텐츠 생성방법 제안

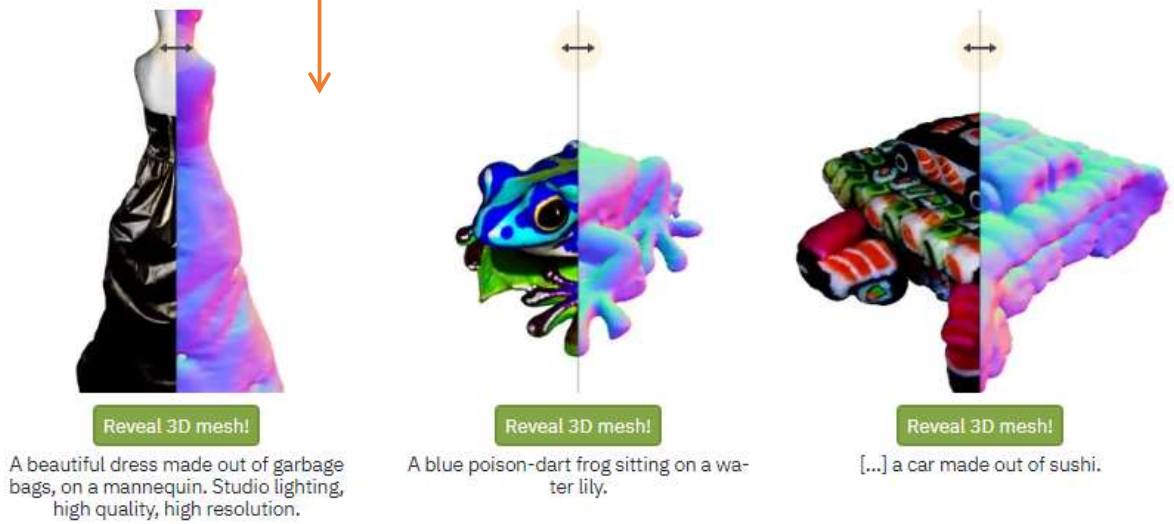
a blue poison-dart frog sitting on a water lily

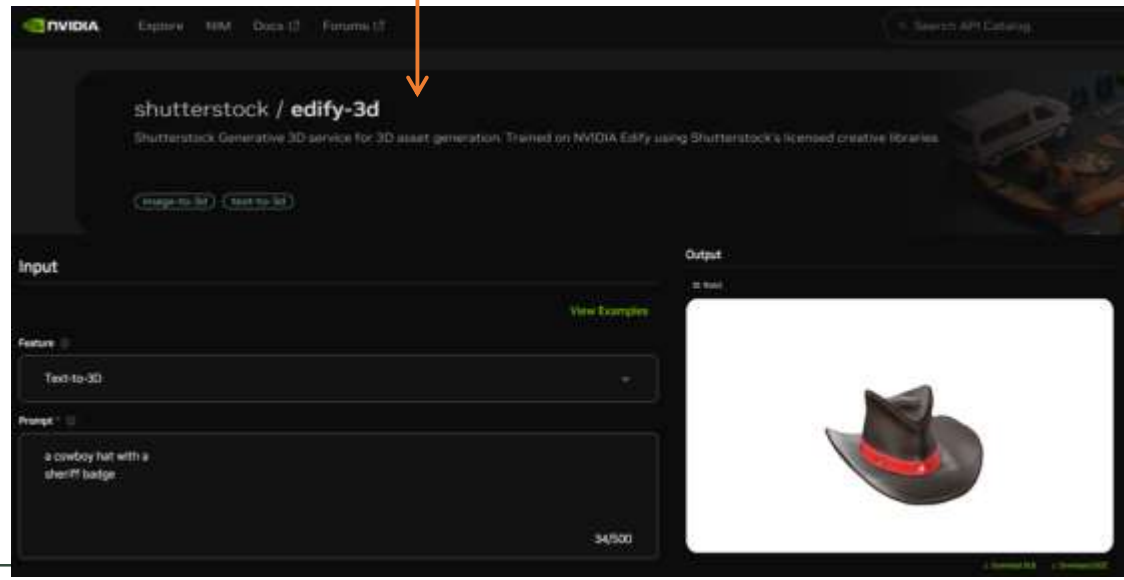
neuschwanstein castle, aerial view

2. Introduction

* Text to 3D 모델의 발전

	발표시기 /개발사	주요기술 /특징	해상도 /생성시간	논문	웹사이트
DreamFusion	2022/ Googleresearch	2D Diffusion +NeRF +SDS손실함수 /저해상도모델로 디테일 부족, 느린 최적화 속도	낮은해상도/ 1.5시간	https://arxiv.org/abs/2209.14988	https://dreamfusion3d.github.io/gallery.html
Magic3D	2022~2023/ 엔비디아	2단계 최적화프레임워크 /**고해상도 텍스트-3D 생성 모델** /해시 그리드 NeRF 사용	높은해상도/ 40분	https://arxiv.org/abs/2211.10440	**https://research.nvidia.com/labs/dir/magic3d/**
Edify3D	2024.11월/ 엔비디아	멀티뷰 Diffusion모델, 재구성모델로 더 세밀한 형상과 텍스처 생성 가능 /각 뷰의 일관성을 보장 ControlNet사용 /참조 이미지를 활용가능	매우높은해상도/ 2분	https://arxiv.org/abs/2411.07135	https://build.nvidia.com/shutterstock/edify-3d





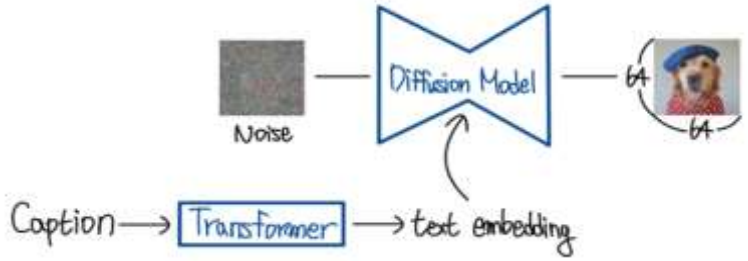
3. Method

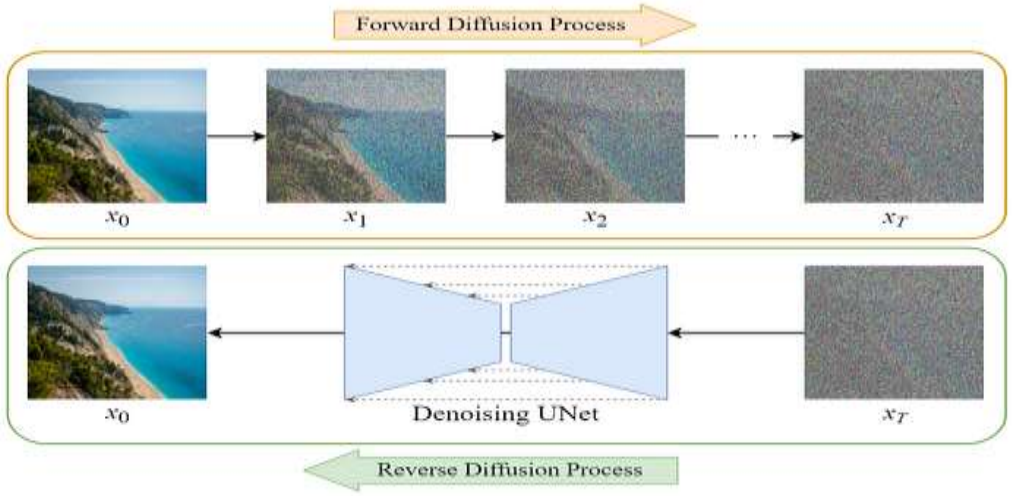
◆ Diffusion 모델

텍스트 입력에 따라 이미지를 생성하는 모델로, 노이즈에서 데이터로 전환을 거침

*** Forward 과정: 점진적으로 노이즈를 추가**

*** Reverse 과정: 노이즈를 제거하며 데이터 복원**





◆ NeRF (Neural Radiance Fields)

3D 장면을 표현을위해 밀도,색상 학습하는 볼륨 렌더링 기술

- **n개 시점에서의 2D 이미지를 합쳐 3D로 나타냄**

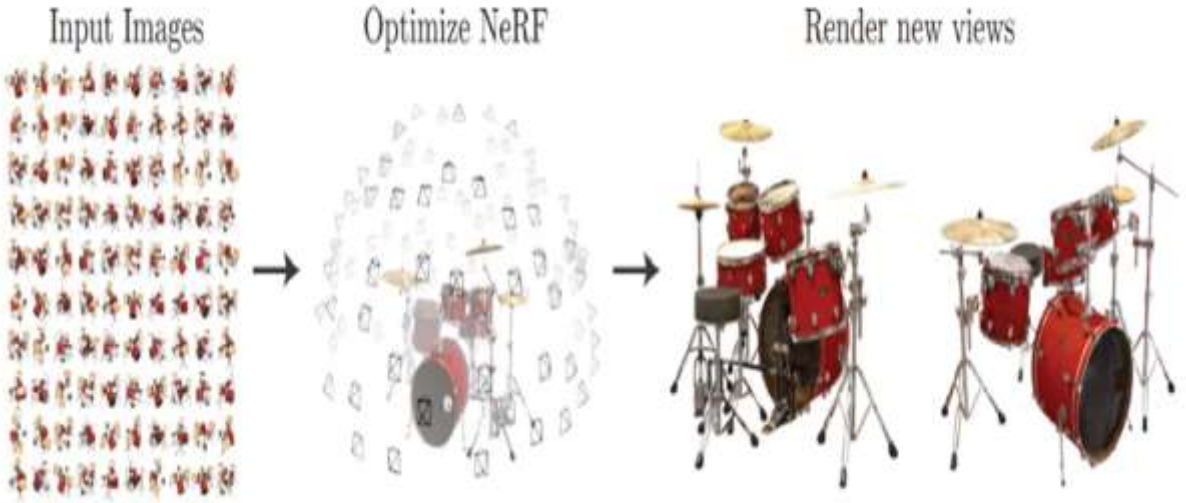
$$\hat{C}_c(\mathbf{r}) = \sum_{i=1}^{N_c} w_i c_i$$

*** 볼륨렌더링공식**

색상c와 가중치w를곱해 최종픽셀색상 C

$$w_i = \alpha_i \prod_{j<i}(1-\alpha_j), \qquad \alpha_i = 1 - \exp\left(-\tau_i \left\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_{i+1}\right\|\right)$$

불투명도 α_i와 투명도 $(1-\alpha_j)$를 곱하여 가중치 w_i를 계산.



◆ SDS를 통한 최적화

- 텍스트 → 2D Diffusion 이미지생성 → SDS를 통한 3D모델 학습 (NeRF모델 파라미터 최적화)
- SDS는 2D Diffusion모델이 생성한 이미지와 NeRF모델이 랜더링한 이미지와 비교하고
- 두 차이를 SDS loss로 구해서 미분한뒤 NeRF를 재학습한다

2D Diffusion모델 이미지생성 → **NeRF모델 렌더링 이미지**

*** 두 이미지간 차이를 줄이는 방향으로 최적화**

*** SDS loss의 그래디언트 계산식**

$$\nabla_\theta \mathcal{L}_{\text{SDS}}(\phi, \mathbf{x} = g(\theta)) \triangleq \mathbb{E}_{t,\epsilon}\left[w(t) \left(\hat{\epsilon}_\phi(\mathbf{z}_t; y, t) - \epsilon\right) \frac{\partial \mathbf{x}}{\partial \theta} \right]$$

시간에 대한 노이즈 기대값 ($\mathbb{E}_{t,\epsilon}$):

가중치 $w(t)$: 오차 ($\hat{\epsilon}_\phi(z_t; y, t) - \epsilon$): 파라미터 변화량 ($\frac{\partial x}{\partial \theta}$):

#training

iteration:

1. random camera position에서 NeRF-like model을 사용하여 rendering
2. 1번의 rendering값과 text-embedding을 통해 나온 diffusion 값을 SDS loss 적용 (loss는 추후서술)
3. NeRF weight update

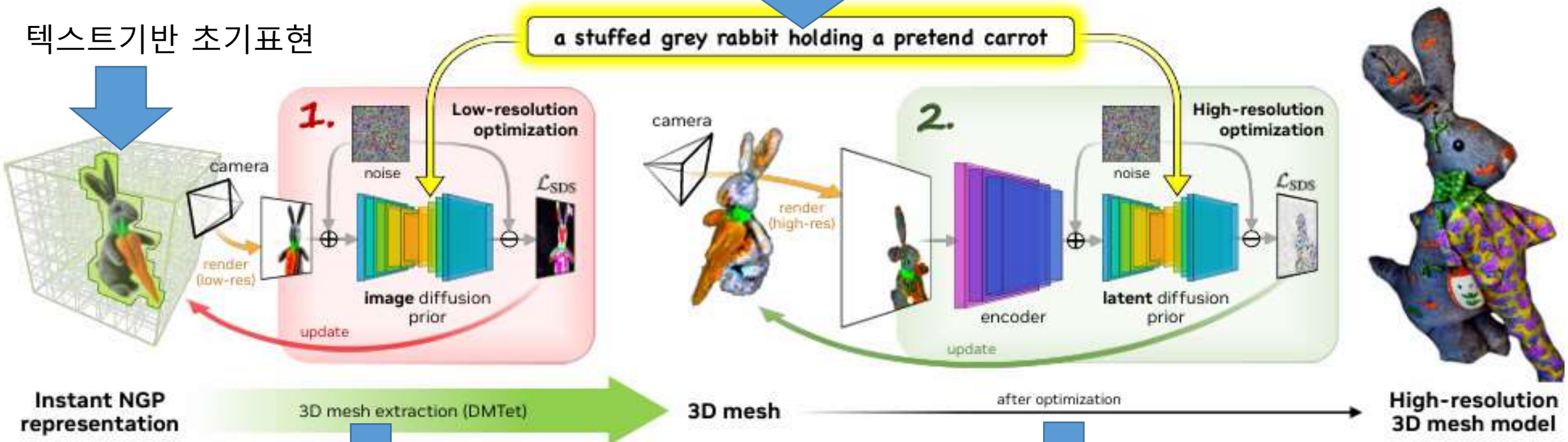
#inference(test)

train된 NeRF모델로 특정 camera position, direction에서 image 생성

◆ Magic3D : Two-stage diffusion 사용

텍스트 입력

텍스트기반 초기표현



최종결과

NeRF에서 초기3D메쉬 추출
(저해상도)

디테일추가된 3D메쉬 모델 생성
(고해상도 최적화)

◆ Magic3D : Two-stage diffusion 사용

< Coarse Stage >

1) eDiff-I 확산모델

- 텍스트 조건에 따라 장면을 학습하여 3D모델 방향성을 제공
- 저해상도 이미지(64×64) 에서 그래디언트를 계산해 NeRF의 초기 형상을 생성

2) 해시 그리드 기반 NeRF 모델사용

- eDiff-I의 그래디언트를 적용하여 3D 구조와 텍스처를 업데이트
- 저해상도 이미지(64×64) 생성한 뒤, SDS를 사용해 해시그리드 NeRF 최적화.

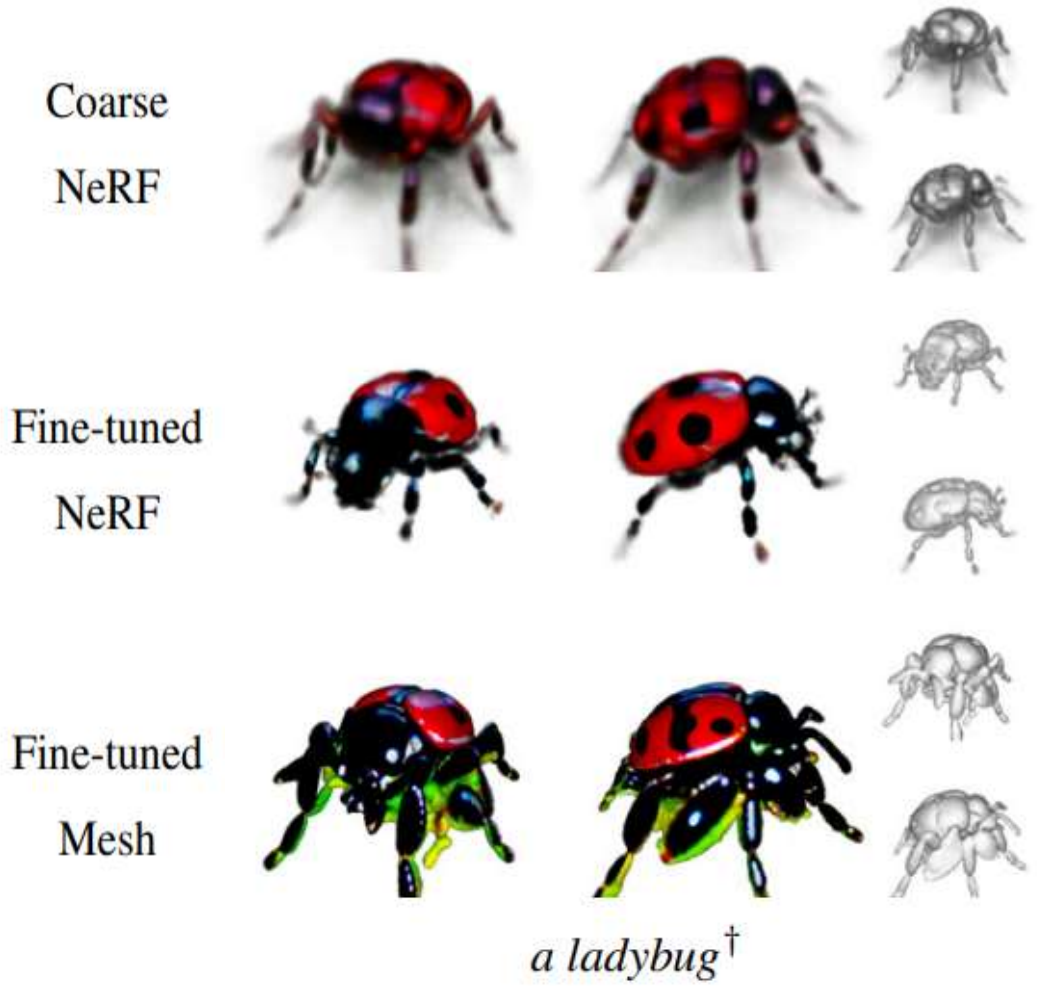
< Fine Stage: 고해상도 최적화 >

1) LDM 기반 Stable Diffusion모델

- Coarse Stage에서 전달된 결과로 고해상도 이미지(512×512)를 생성

2) 텍스처 메시 모델:

- LDM이 생성한 고해상도 정보를 3D 메시로 변환
- 3D 모델의 표면과 디테일을 효율적으로 표현



◆ 최적화 방법

* DreamFusion보다 빠르고 효율적인 학습 가능.

1) NeRF 최적화

- Mip-NeRF 360 대신 **해시그리드 기반 NeRF** 사용

: 해시테이블 이용 3D 공간좌표 정보 인코딩하여 계산속도 높임

- **옥트리 기반 레이 샘플링**사용

: 3D 공간을 밀도가 높은 영역만 집중 샘플링

2) 실시간 렌더링

- 고해상도 3D 모델 생성 시, 미분 가능한 래스터라이저를 사용하여 **실시간으로 2D 뷰를 렌더링**

이 과정에서 카메라의 위치와 각도를 조정하며, 특정 뷰에 초점을 맞춰 세부적인 3D 정보를 복원

- 특정 뷰에서의 고주파 세부 정보를 강조하여 디테일을 보완

클로즈업 카메라 뷰를 활용해 중요한 영역에 학습 자원을 집중

- 사용자가 자신의 작업을 빠르고 정확하게 시각화할 수 있으며 고품질의 3D 콘텐츠를 글자를 통해 생성, 수정, 렌더링, 공유 가능

4. Experiments

Magic3D 실험

DreamFusion 웹사이트에서 제공한 **397개의 텍스트 프롬프트**에 대해 Magic3D를 훈련

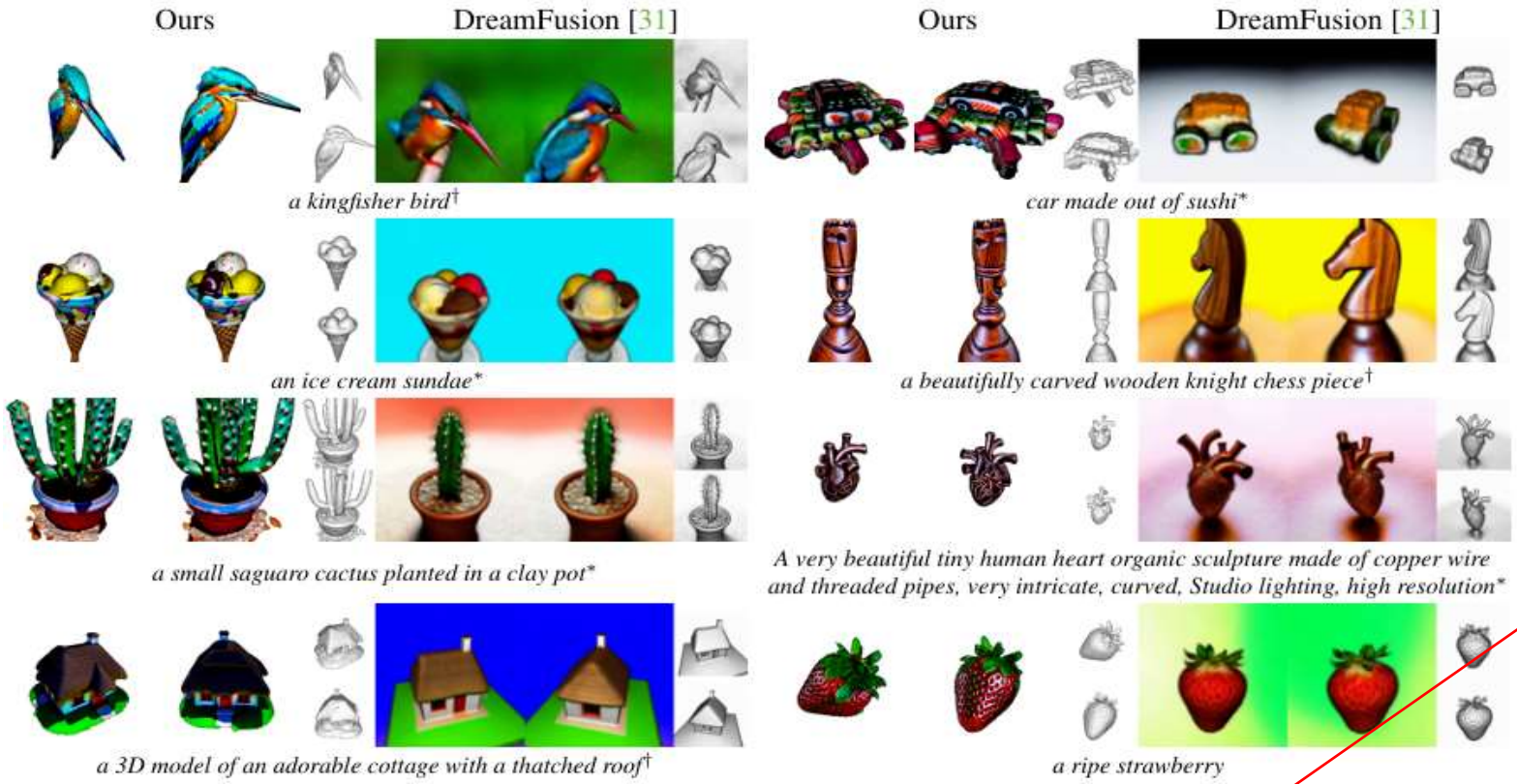
1. 실험 설정
모델 학습:
1) Coarse Stage : 5000번반복, 15분소요
2) Fine Stage : 3000번반복, 25분소요
3) 총 학습시간 : 약40분 (8개의 NVIDIA A100 GPU사용)

결과1) 학습후 생성된 3D모델의 사용자 선호도 : 61%의 사용자가 Magic3D선택

Comparison	Preference
Magic3D *vs.* DreamFusion [31]	
• More realistic	58.3%
• More detailed	66.0%
• More realistic & detailed	61.7%
Magic3D *vs.* Magic3D (coarse only)	87.7%

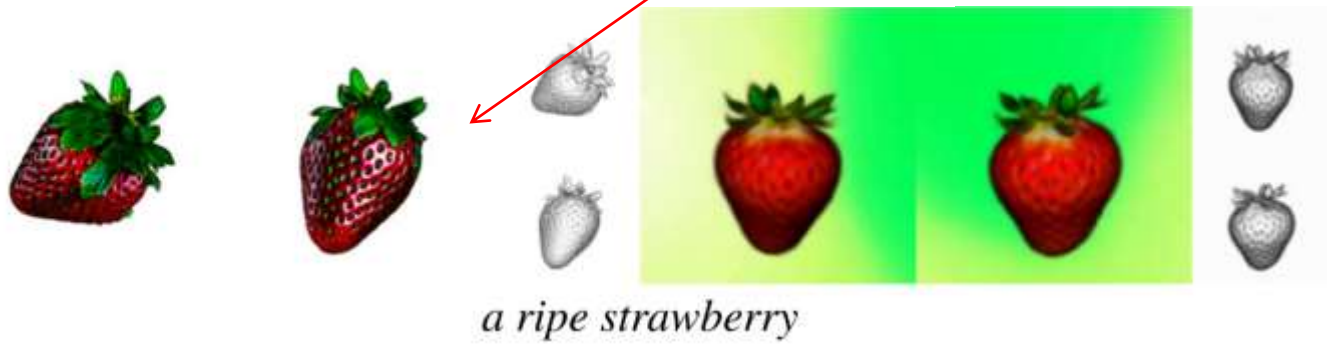
87.7%는 Magic3D에서 coarse only 모델보다 Magic3D 모델을 선호

결과2) 학습결과 생성된 3D모델 비교



고해상도 디테일
기하학적구조 명확
텍스쳐, 색상 풍부

Ex)
* 딸기의 표면 씨앗 명확
* 색상그라데이션 명확



결과3) CLIP R-Precision를 통한 정량적 평가

CLIP 모델을 사용해 텍스트와 렌더링된 이미지간 컬러 및 기하학적 일치도 측정
평가데이터: MS-COCO 데이터셋에 포함된 실제 이미지와 비교.

* CLIP R-Precision:
렌더링된 이미지가 주어졌을 때, CLIP 모델이 올바른 텍스트 프롬프트를 정확하게 식별하는 능력을 측정하는 지표
-> 생성된 3D 모델이 주어진 텍스트 프롬프트와 얼마나 일치하는지 평가

Method	R-Precision ↑					
	CLIP B/32		CLIP B/16		CLIP L/14	
	Color	Geo	Color	Geo	Color	Geo
GT Images	77.1	–	79.1	–	–	–
Dream Fields	68.3	–	74.2	–	–	–
(reimpl.)	**78.6**	1.3	(99.9)	(0.8)	**82.9**	1.4
CLIP-Mesh	67.8	–	75.8	–	74.5†	–
DreamFusion	75.1	**42.5**	**77.5**	46.6	79.7	**58.5**

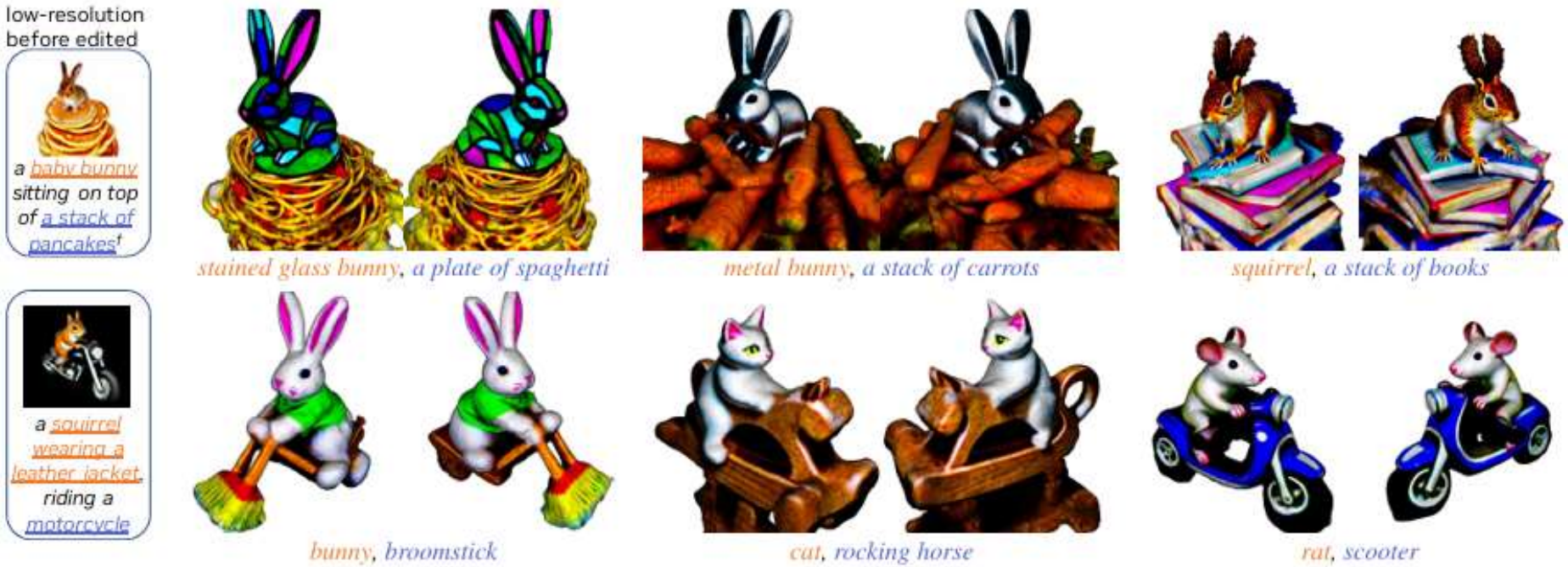
프롬프트 가능한 3D생성

파인튜닝:
고양이 11장, 개 4장의 이미지를 특정주제 반영한 학습 데이터로 사용.
미세 조정 과정:

eDiff-I: 학습률 1×10^{-5}, 1500회 반복.

LDM: 학습률 1×10^{-6}, 800회 반복.

결과: 텍스트 지시에 맞추면서도 입력 이미지의 고유한 특징을 유지한 3D 모델 생성 가능.



5. Conclusion

*** 3D 생성 모델 연구의 새로운 패러다임 제시**
3D 데이터를 필요로 하지 않는 텍스트-3D 생성 방식

*** 사용자 제어 강화**:
프롬프트 기반 편집으로 콘텐츠를 세밀히 조정하며 일반사용자에게 창작 가능성 확대.

*** 해상도한계 극복**
Diffusion 모델(Imagen)의 해상도한계(64x64)가 3D 품질에 영향을미침.
Magic3D는 **DreamFusion**에 비해 고해상도 3D 모델 생성에서 더 높은 품질과 효율성을 제공

*** 실시간렌더링**
작업을 빠르게 시각화하여 프로세스의 효율성 향상, 빠른확인가능

*** 효율적 3D자산 생성**
3D자산 효율적생산으로 Magic3D는 그래픽, 게임, AR/VR, 창작도구 전반에 걸쳐 중요

Appendix –Edify3D

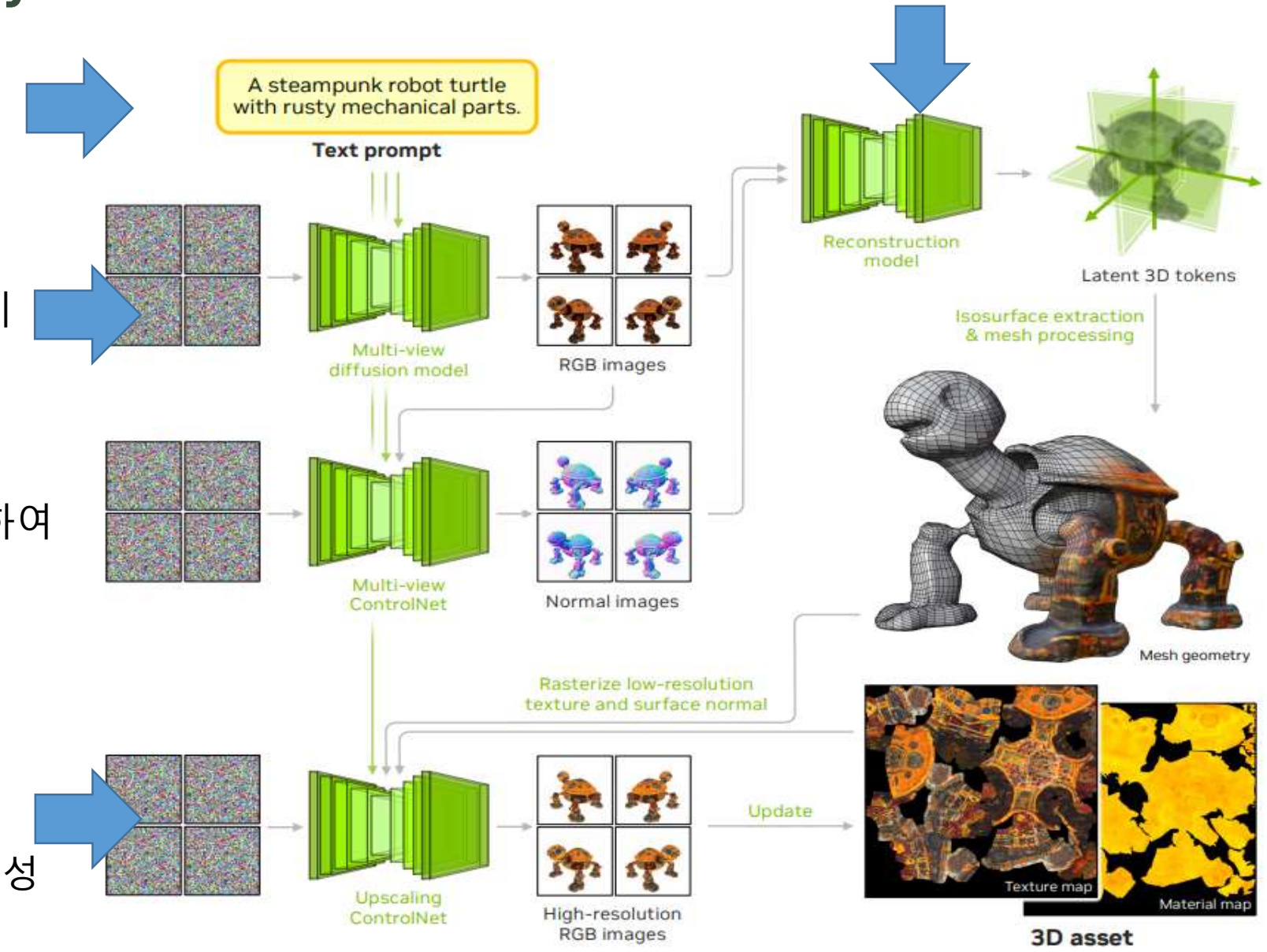
1. 텍스트입력

2. 멀티뷰 DiffusionModel이 여러시점에서 이미지생성

* ControlNet 사용하여 정확한결과를 얻음

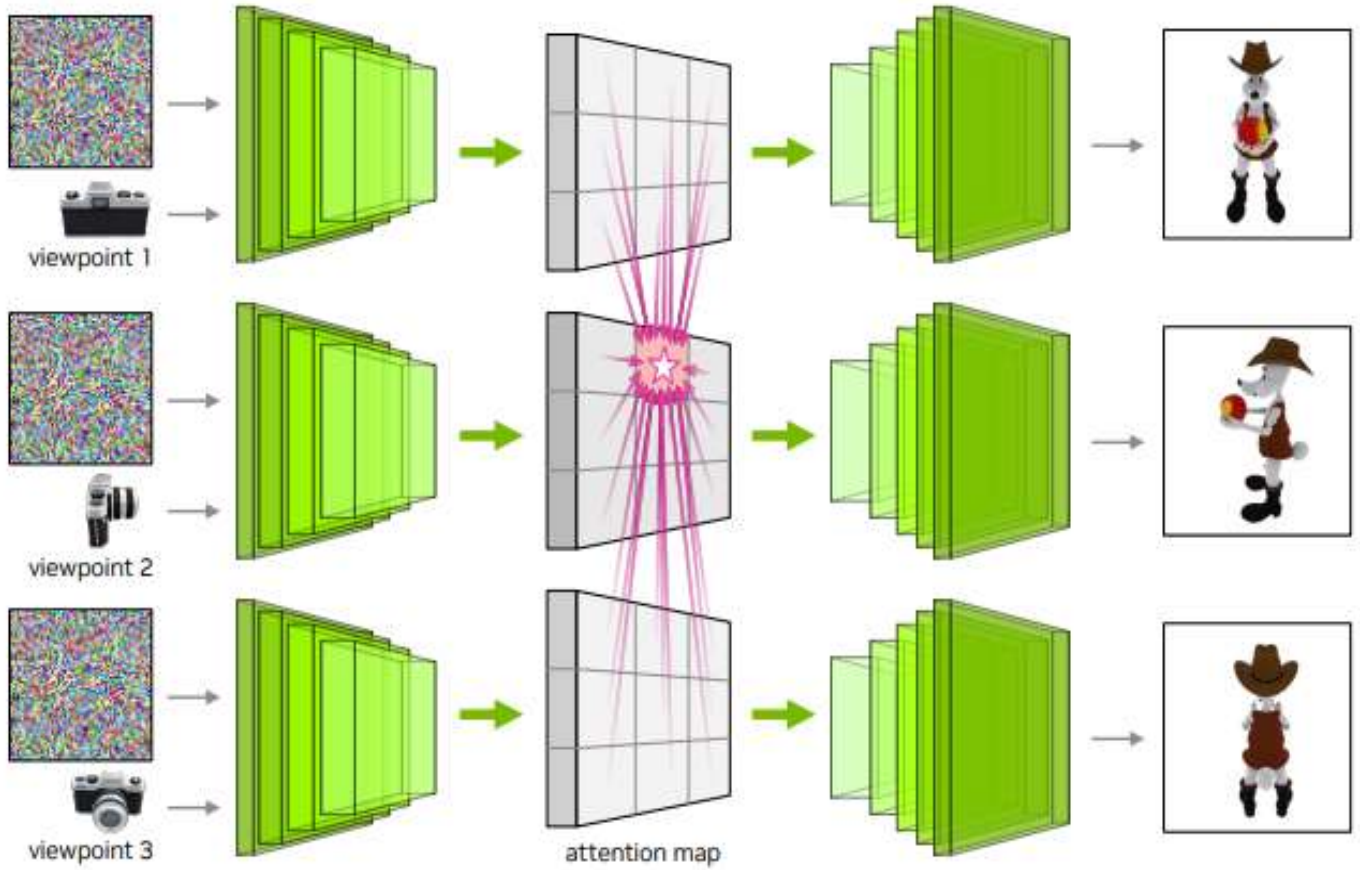
4. 업스케일링 고해상도텍스쳐 생성

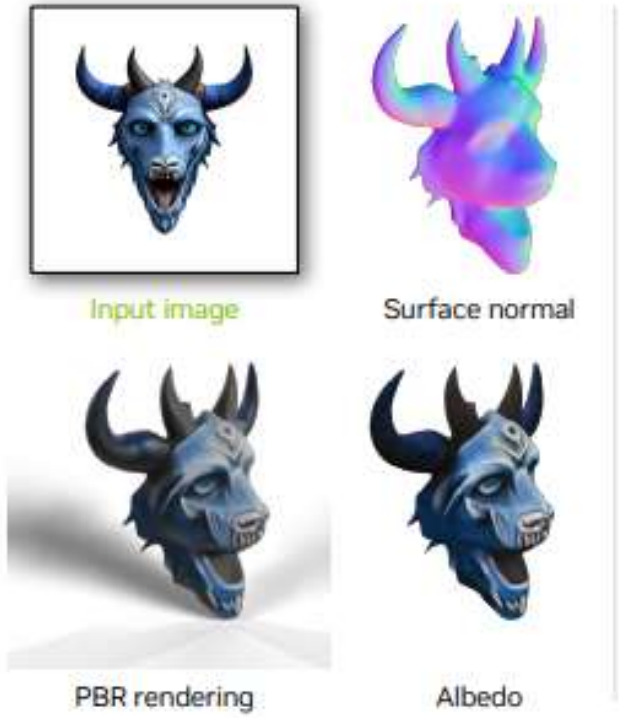
3. 재구성모델이 3D 텍스쳐, 메쉬생성



Appendix –Edify3D

- 각 입력뷰에서 특징맵 추출
- 중앙의 attention map에서 다중뷰데이터 통합
- 각 입력데이터의 연관성 학습, 3D객체 렌더링





Appendix –Edify3D

Albedo LPIPS error

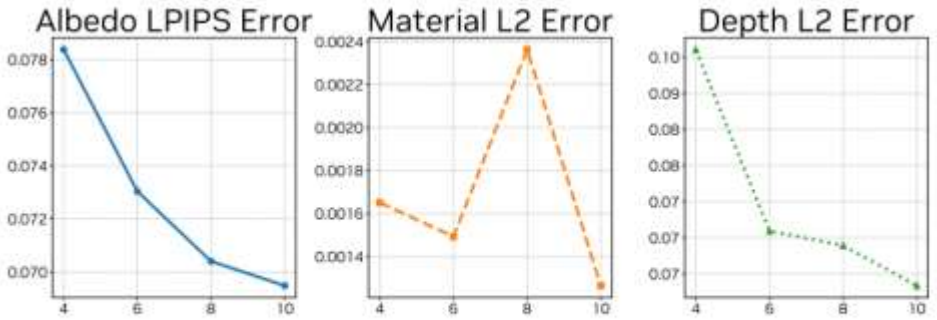
Input views	Validation views			
	4	4 (diag.)	8	16
4	0.0732	0.0791	0.0762	0.0768
4 (diag.)	0.0802	0.0756	0.0779	0.0783
8	0.0691	0.0698	0.0695	0.0699
16	0.0687	0.0689	0.0688	0.0687

Material L_2 error

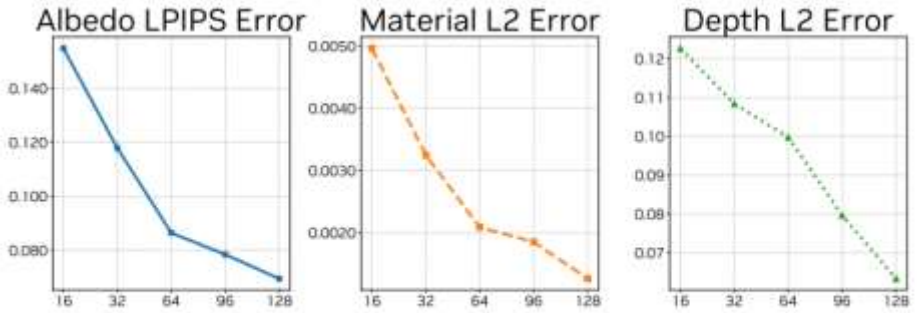
Input views	Validation views			
	4	4 (diag.)	8	16
4	0.0015	0.0020	0.0017	0.0018
4 (diag.)	0.0024	0.0019	0.0022	0.0022
8	0.0013	0.0012	0.0013	0.0013
16	0.0012	0.0013	0.0013	0.0013

Depth L_2 error

Input views	Validation views			
	4	4 (diag.)	8	16
4	0.0689	0.0751	0.0720	0.0722
4 (diag.)	0.0704	0.0683	0.0694	0.0696
8	0.0626	0.0641	0.0633	0.0633
16	0.0613	0.0626	0.0619	0.0616



(a) Number of Training Views



(b) Triplane Resolution

X축: 훈련에 사용된 뷰포인트 수 (4, 6, 8, 10).
Y축: 각 오류 지표(Albedo, Material, Depth)의 값.

* **Albedo LPIPS Error** (파란색):
뷰 수가 증가할수록 텍스처품질개선
* **Material L2L_2L2 Error** (주황색):
재질 품질도 입력 뷰 수가 많아질수록 향상
* **Depth L2L_2L2 Error** (초록색):
깊이(구조) 품질도 유사하게 입력 뷰 수 증가에 개선

X축: 트라이플레인 해상도 (16, 32, 64, 96, 128).
Y축: 각 오류 지표(Albedo, Material, Depth)의 값

* **Albedo LPIPS Error** (파란색):
트라이플레인 해상도가 높아질수록 텍스처 품질 향상
* **Material L2L_2L2 Error** (주황색):
높은 해상도에서 재질 정보의 품질개선
* **Depth L2L_2L2 Error** (초록색):
해상도가 증가할수록 깊이 예측 오류가 감소

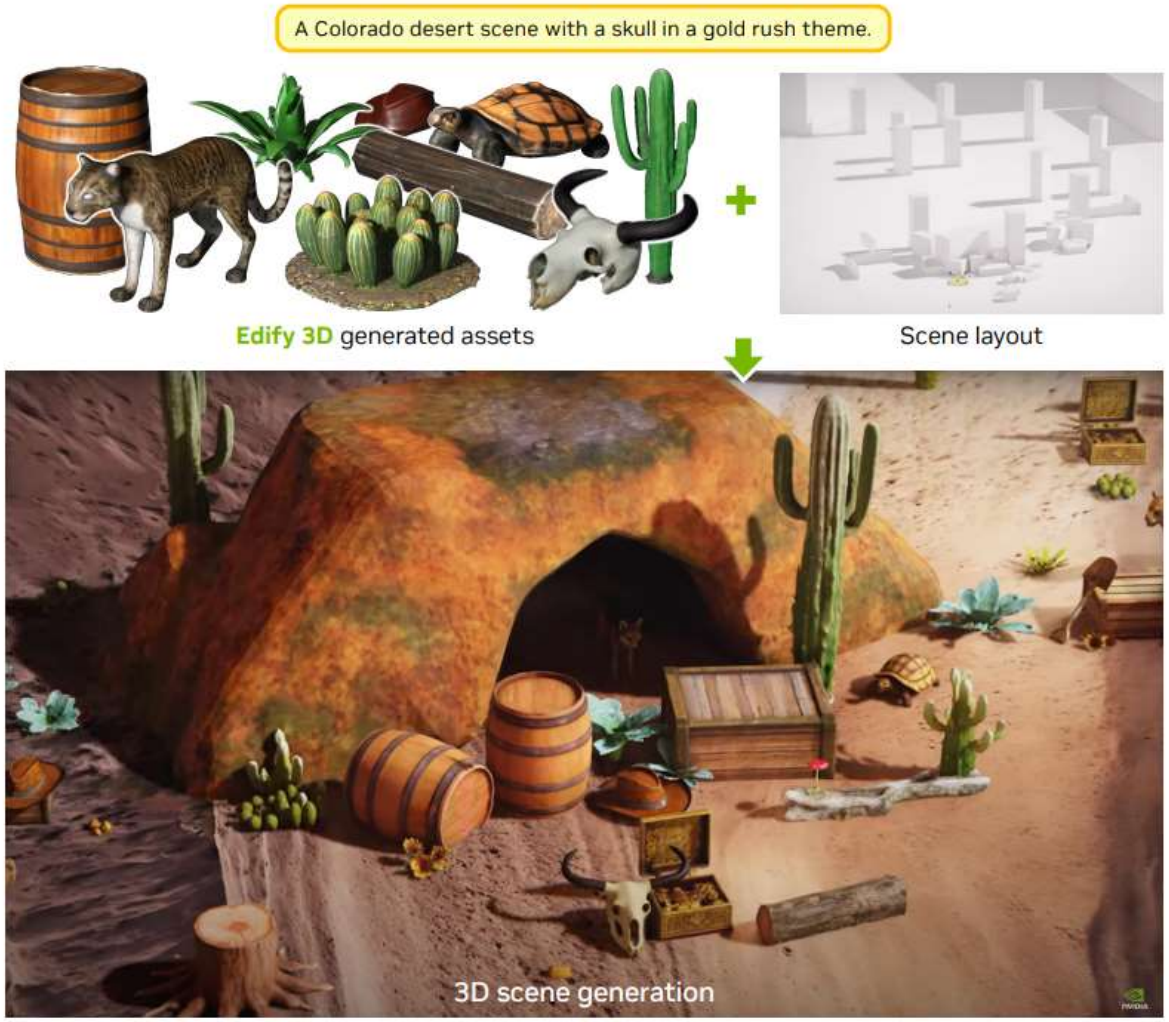
*** 트라이플레인 해상도**
: 3D 객체를 표현하기 위한 Triplane 데이터 밀도

Appendix –Edify3D

*** LLM과 Edify 3D를 활용한 3D장면 생성**

LLM : 텍스트프롬프트를 해석하여 장면에 포함될
객체의 위치,크기,배치 결정

Edify3D : LLM이 제공한 레이아웃 기반 고품질 3D에셋 생성하고,
이를 조합하여 3D장면 구성



Reference

https://arxiv.org/abs/2209.14988
https://dreamfusion3d.github.io/gallery.html

https://arxiv.org/abs/2211.10440
https://research.nvidia.com/labs/dir/magic3d/

https://arxiv.org/abs/2411.07135
https://build.nvidia.com/shutterstock/edify-3d