DIRECT-3D: Learning Direct Text-to-3D Generation on Massive Noisy 3D Data

Qihao Liu[1], Yi Zhang[1], Song Bai[2], Adam Kortylewski[3,4], Alan Yuille[1]

[1] Johns Hopkins University [2] ByteDance [3] Max Planck Institute for Informatics [4] University of Freiburg

Motivation

- Generating diverse and high-quality 3D objects is an important task.
- Challenging due to the lack of 3D data:

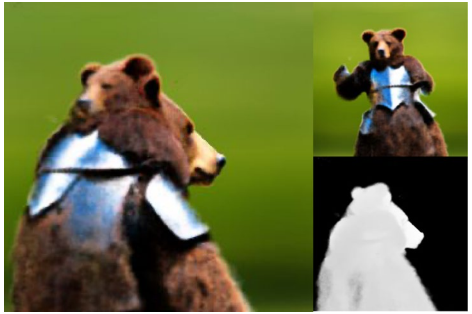
Dataset name		size	annotation	feature
2D dataset	LAION	5B	Image-text pair	Filtered with CLIP
3D dataset	ShapeNet	51K	Class name	Clean and aligned, can be directly used for training
	Objaverse-XL	10M	No annotation	Noisy, not aligned

Random samples of “dog” from the Objaverse dataset.

Related work

- 2D-lifting methods:
    - Pro: using 2D image diffusion models, no needs for 3D data
    - Cons:
        - Slow: optimization process
        - Janus problem (multi face problem)

- Directly train 3D generative model on clean data (proprietary data):
    - Pros:
        - Fast in 3D generation
        - More accurate 3D geometry consistency
    - Cons:
        - Single/few class generation
        - Lack of diversity
        - Hard to scale up (needs considerable efforts to collect and clean data)



DreamFusion suffers from the Janus problem

Shap-E from OpenAI is trained on extensive proprietary data: is costly to obtain, requires huge efforts to further enhance data quality

Method

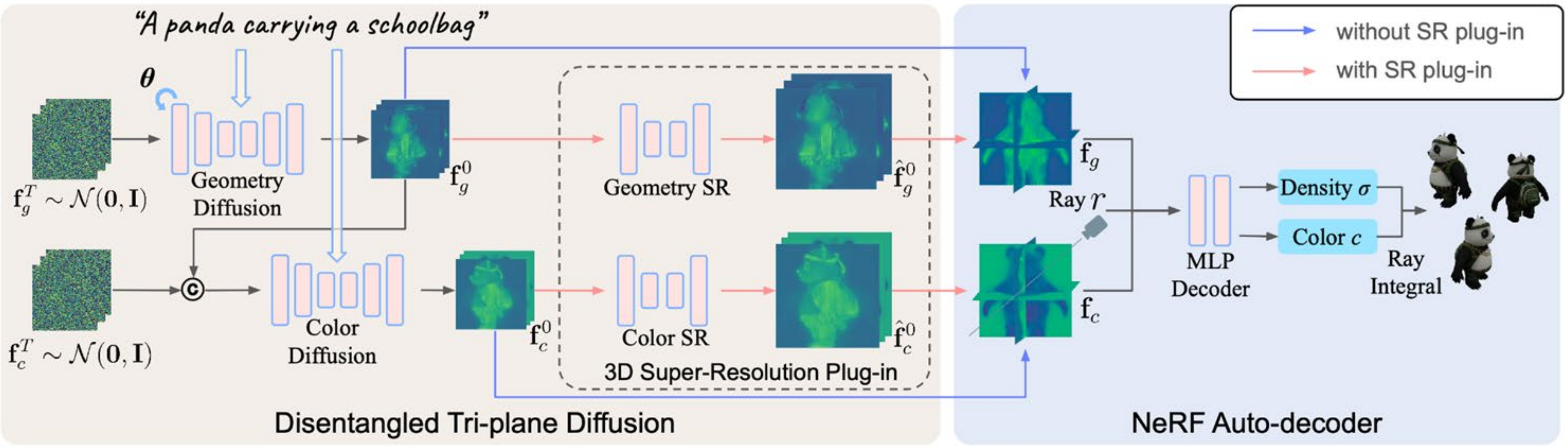
- Can we directly train a 3D generative model on massive noisy and not-aligned ‘in-the-wild’ 3D data such as Objaverse-XL?

“*A statue of a black dog*”

Challenges:

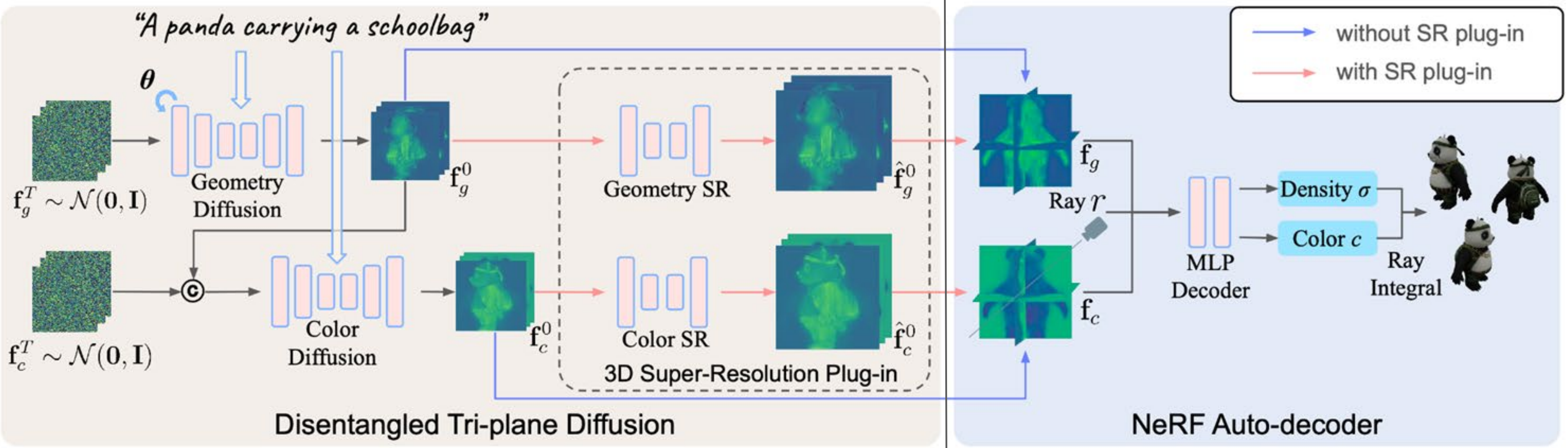
- Training directly on non-aligned data is challenging and may result in non-convergence.
- There is no consensus on 3D data representation or network architecture that can efficiently handle high-dimensional 3D data.

Method



- NeRF is used to represent 3D objects.
- Tri-plane features enable the use of a 2D diffusion architecture.
- An iterative optimization process in the diffusion step explicitly estimates the pose and quality of the 3D data based on the conditional density.
- We disentangle the 3D geometry and 2D color of the object, modeling them hierarchically with two separate diffusion models.

Method



Disentangled tri-plane generation:

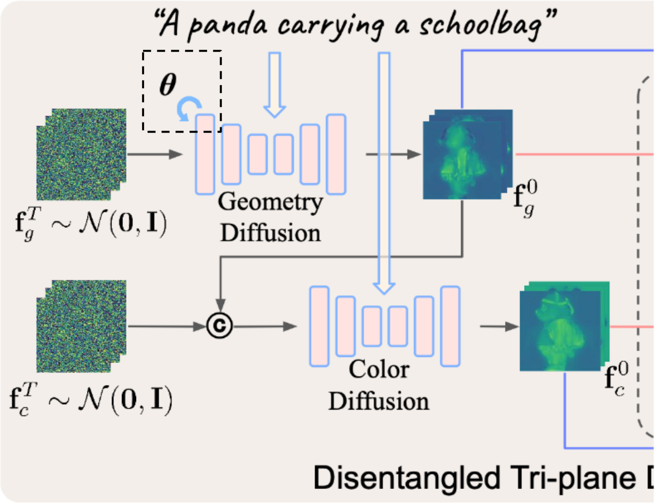
$$\mathcal{L}_{geo}(\phi) = \mathbb{E}_{\mathbf{f}_g^0,\epsilon,p,t}[||\epsilon - \epsilon_\phi(\mathbf{f}_g^t, t, \tau(p))||_2^2]$$

$$\mathcal{L}_{col}(\psi) = \mathbb{E}_{\mathbf{f}_c^0,\epsilon,p,\mathbf{f}_g,t}[||\epsilon - \epsilon_\psi(\mathbf{f}_c^t, t, \tau(p), \mathbf{f}_g)||_2^2]$$

NeRF generation from disentangled tri-plane representation:

$$\mathcal{L}_{rad}(\mathbf{f}_g, \mathbf{f}_c, \omega) = \sum_i ||\hat{y}_i - \mathcal{R}(\mathcal{D}_\omega(\mathbf{f}_g, \mathbf{f}_c, r_i))||_2^2$$

Method



Training with noisy and unaligned data:

We explicitly model the 3D rotation angle of an object as $\theta = \{\theta\mu, \theta\sigma\}$

$$\mathcal{L}_{geo}(\phi) = \mathbb{E}_{\mathbf{f}_g^0,\epsilon,p,t}[||\epsilon - \epsilon_\phi(\mathbf{f}_g^t, t, \tau(p))||_2^2]$$

$$\Downarrow$$

$$\mathcal{L}_{geo}(\phi, \theta) = \mathbb{E}_{\mathbf{f}_g^0;\theta,\epsilon,p,t}[||\epsilon - \epsilon_\phi(\mathbf{f}_g^t; \theta, t, \tau(p))||_2^2]$$

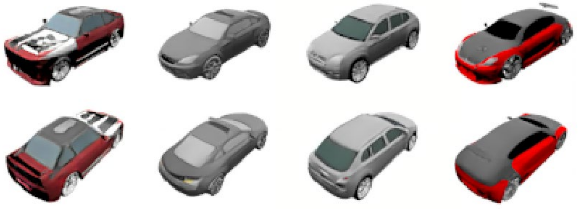
We consider θ as a hidden variable and estimate it based on the based on the conditional density.

Experiments

- Single-class 3D generation

Method	Car		Chair		Table	
	FID (↓)	KID (↓)	FID (↓)	KID (↓)	FID (↓)	KID (↓)
π-GAN [7]	36.7	-	52.71	13.64	41.67	13.82
EG3D [8]	10.46	4.90	16.54	8.41	31.18	11.67
DiffRF [43]	-	-	15.95	7.94	27.06	10.3
SSDNeRF [10]	11.08	3.47	-	-	14.27	4.08
Ours	**6.90**	**1.84**	**7.01**	**2.12**	**7.26**	**1.89**

Table 1. **Single-class generation on SRN Cars, PS Chairs, and ABO Tables.** Baseline results are reported by DiffRF and SSDNeRF. We train our model from scratch using exactly the same rendered images as the baselines. KID is multiplied by 10^3.



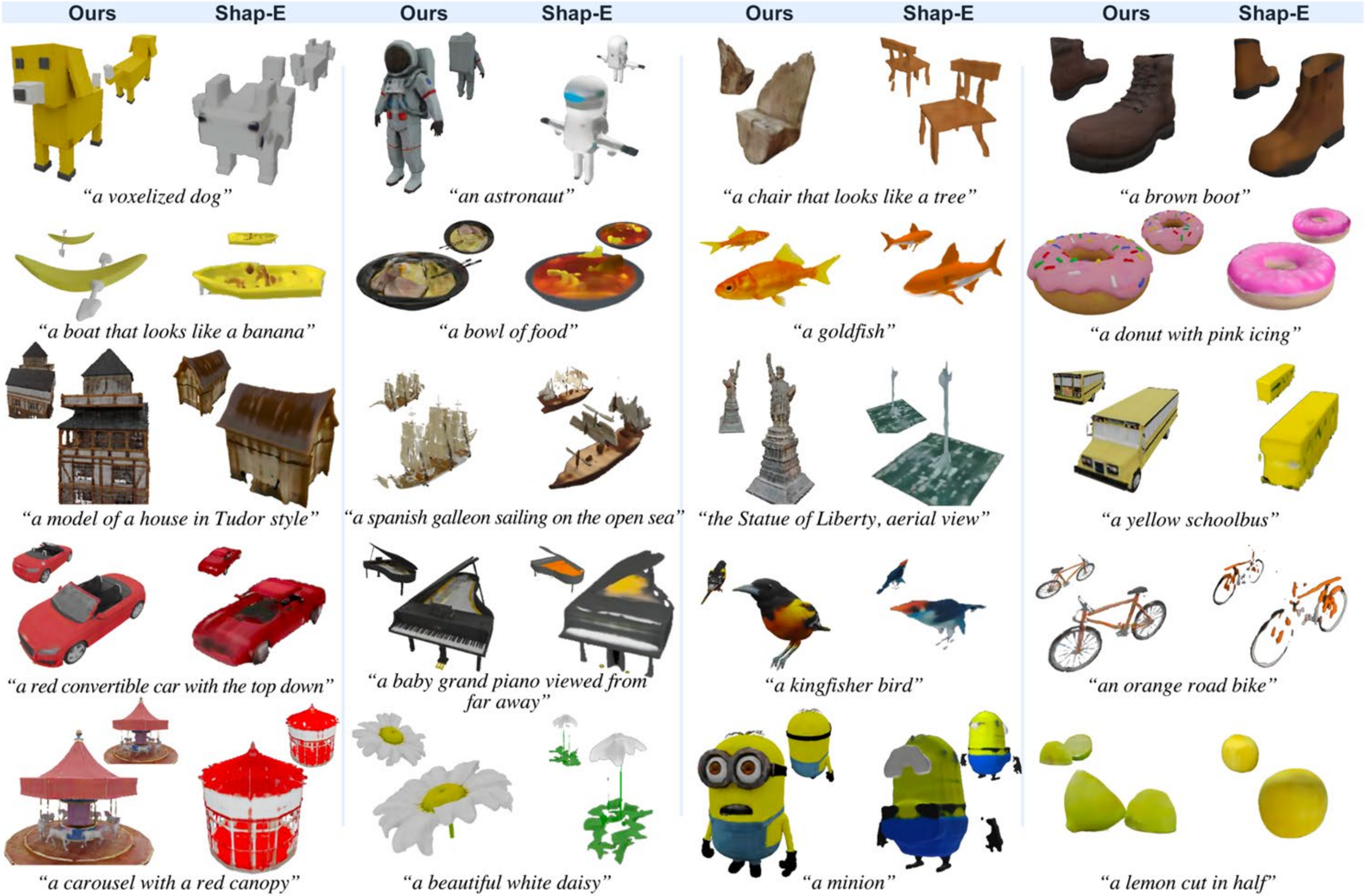
EG3D

SSDNeRF

Ours

Experiments

- Direct text-to-3D generation



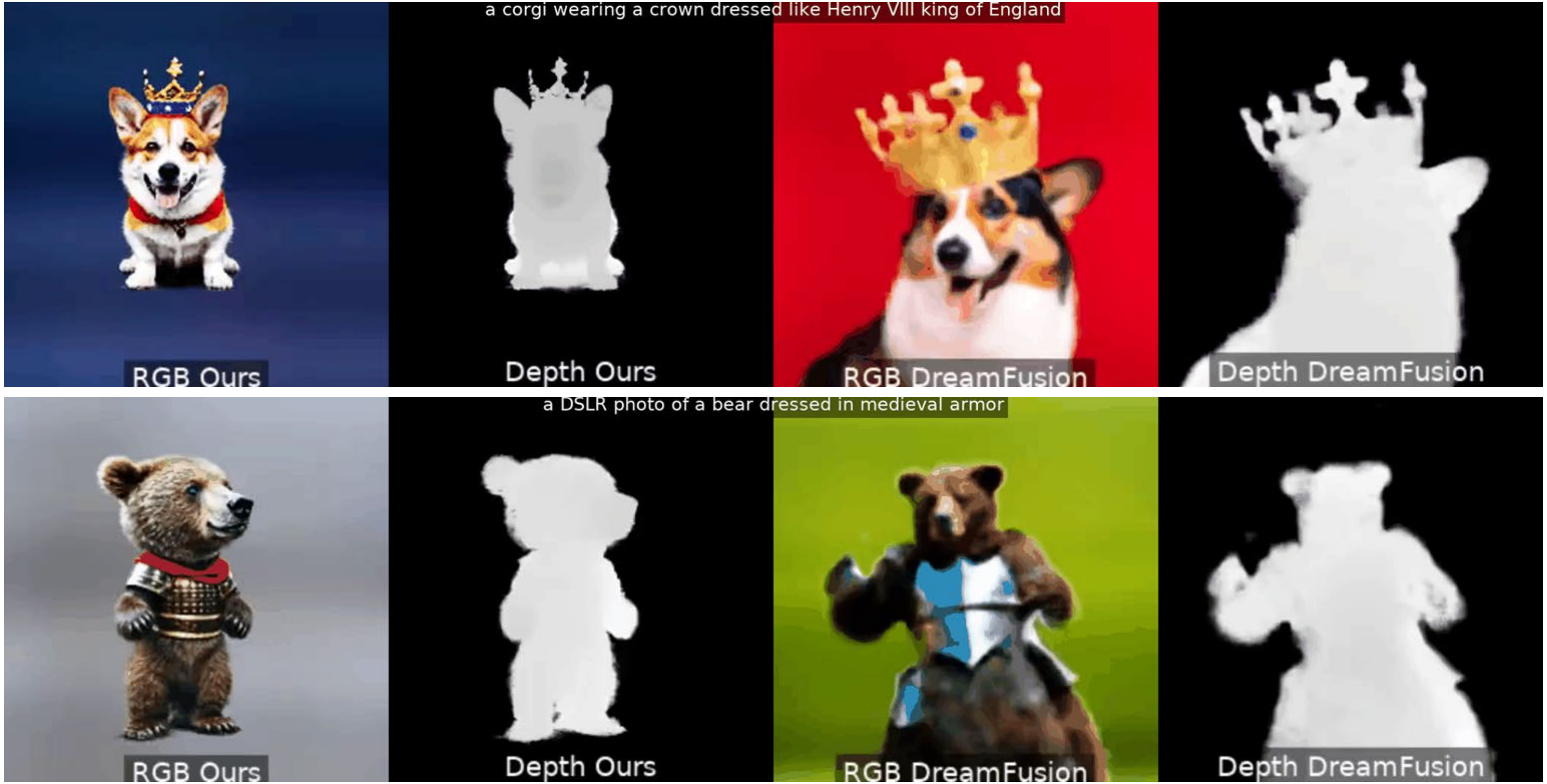
Prompts from Shap-E

Prompts from DreamFusion

complex objects

Experiments

- Using Ours as 3D prior to improve 2D-lifting optimization-based methods (to solve multi-face problem)



Experiments

- Using Ours as 3D prior to improve 2D-lifting optimization-based methods (to solve multi-face problem)

Experiments

- Quantitative comparison on text-to-3D

	More realistic	More detailed	Overall preference
Shap-E [29]	28.4%	22.9%	26.1%
Ours	71.6%	77.1%	73.9%

Table 2. **User preference studies.** We conduct user studies on 475 prompts, including all prompts from Shap-E and 162 prompts from DreamFusion. 73.9% of users prefer ours over Shape-E.

	Succ. Rate	Geo. Consist.	Tex. Consist.
DreamFusion-SD [47]	12%	16%	30%
DreamFusion-IF [47]	10%	10%	72%
DreamFusion-SD + Ours	**84%**	**84%**	**98%**

Table 3. **Improving 2D-lifting text-to-3D generation.** DIRECT-3D provides a useful 3D geometry prior, enhancing the geometry consistency and increasing the generation success rate.

Experiments

- Ablation of Automatic Alignment and Cleaning (AAC)

Figure 5. **Tri-plane feature learned with/without Automatic Alignment and Cleaning (AAC) on Objaverse.** It roughly aligns the objects to get clear tri-plane features. Unaligned objects can be captured by tri-plane representation, but the inadequate axis disentanglement makes it challenging for the diffusion model to learn.

Figure 6. **Model learned with/without AAC on Objaverse.** AAC enables direct and more efficient training on noisy, unaligned data.

Experiments

- Ablation of disentanglement

	Car		Table		Car + Chair + Table	
	FID (↓)	KID (↓)	FID (↓)	KID (↓)	FID (↓)	KID (↓)
Not Disentangled	9.98	2.96	12.86	3.87	17.74	8.15
Disentangled	6.90	1.84	7.26	1.89	10.06	3.44

Table 5. **Improvement of Disentanglement.**

Figure 7. **Disentangling geometry and color provides a proper 3D geometrical prior, while improving the high-fidelity texture from 2D image diffusion models.**

Experiments

- Ablation of prompt enrichment

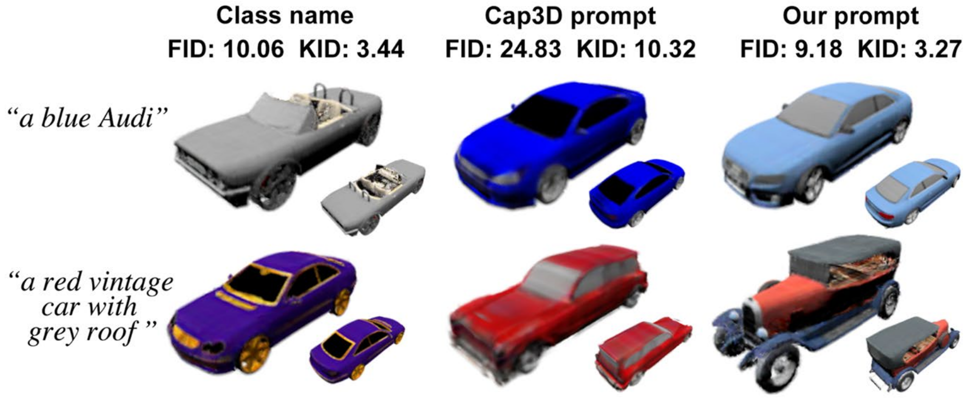


Figure 8. **Prompt Enrichment.** FID and KID are computed on the entire test set. We provide captions with varying granularities: Coarse captions enhance object-category connections, simplifying the training, while fine-gained captions enable a better understanding of detailed features such as color and part-level information.

Experiments

- More results

"a Wall-E"

"an astronaut wearing a colorful spacesuit"

"a Transformed Bumblebee robot with intricate body details"

"an french throne chair"

"a voxelized cupcake made with LEGO"

"a biplane with yellow wings"

"a red convertible car with the top down"

"a batman mask"