

Research Lines of 3D Generation

Score Distillation Sampling (SDS)

- DreamFusion (ICLR'23), Magic3D (CVPR'23), Fantasia3d (ICCV'23)
- Slow optimization
- Multi-face Janus problem

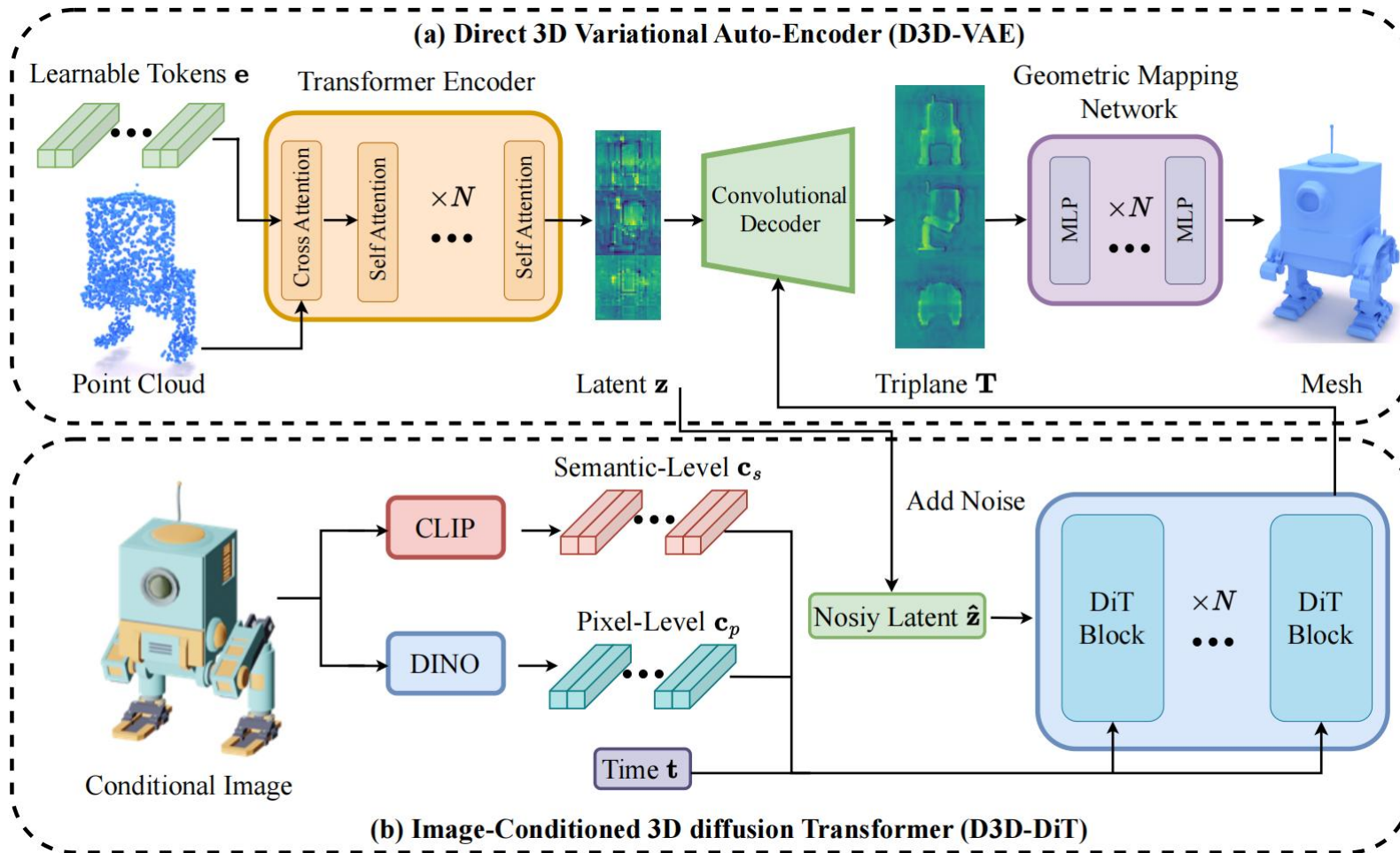
Multi-view Diffusion Model

- SyncDreamer (ICLR'24), Instant3D (ICLR'24), Wonder3D (CVPR'24)
- Low Quality

Native 3D Generation

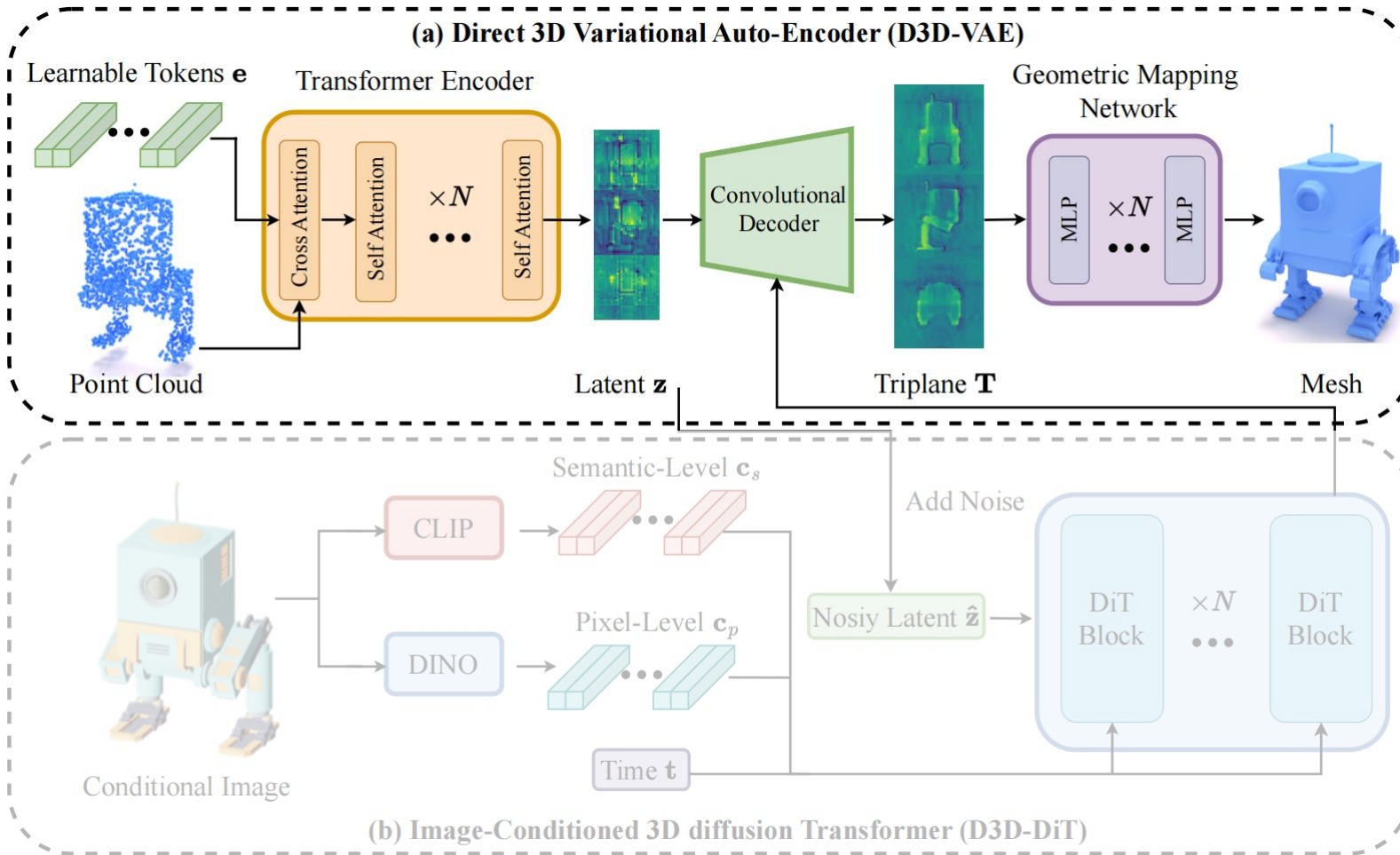
- 3DILG (NeurIPS'22), Michelangelo (NeurIPS'23), 3Dshape2VecSet (ToG'23)
- Smooth and fine geometry
- Unscalable

The Pipeline of Direct3D



- We introduce D3D-VAE to encode 3D shape into an explicit triplane latent space.
- We train the image-conditioned 3D diffusion transformer in the 3D latent space obtained by D3D-VAE.

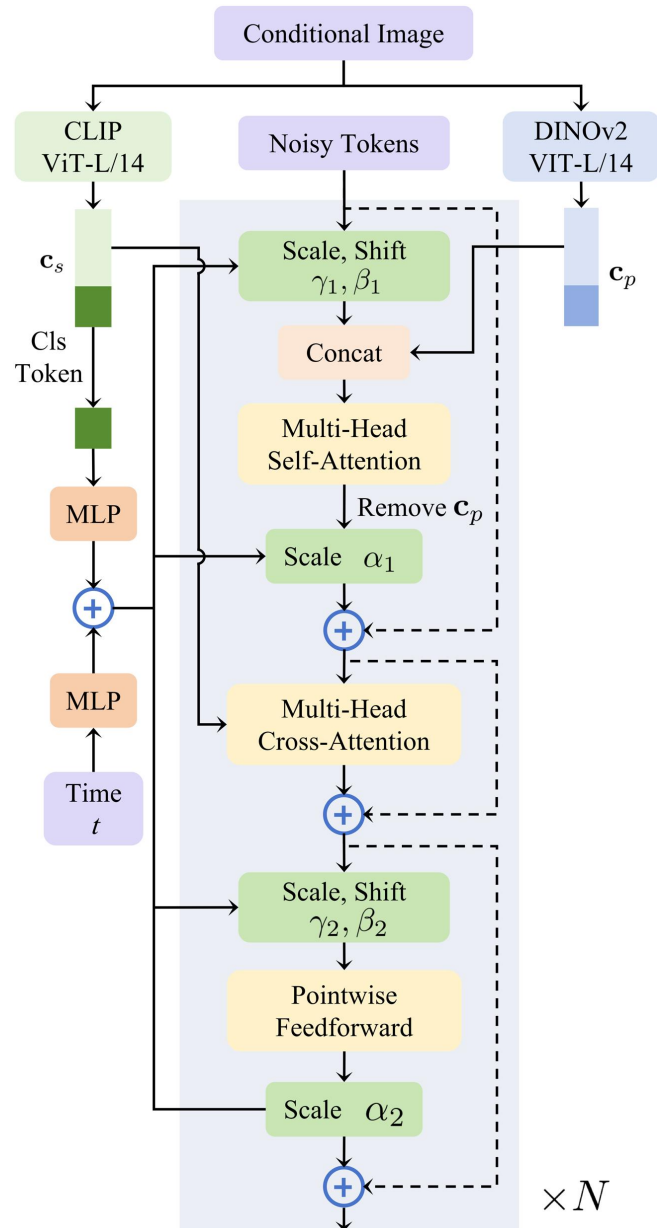
Direct 3D Variational Auto-Encoder



- We utilize transformer to encode point clouds into an explicit triplane latent space.
- A CNN-based decoder is employed to up-sample the latent representations into triplane feature maps.
- We supervise the decoded geometry directly using a semi-continuous surface sampling strategy:

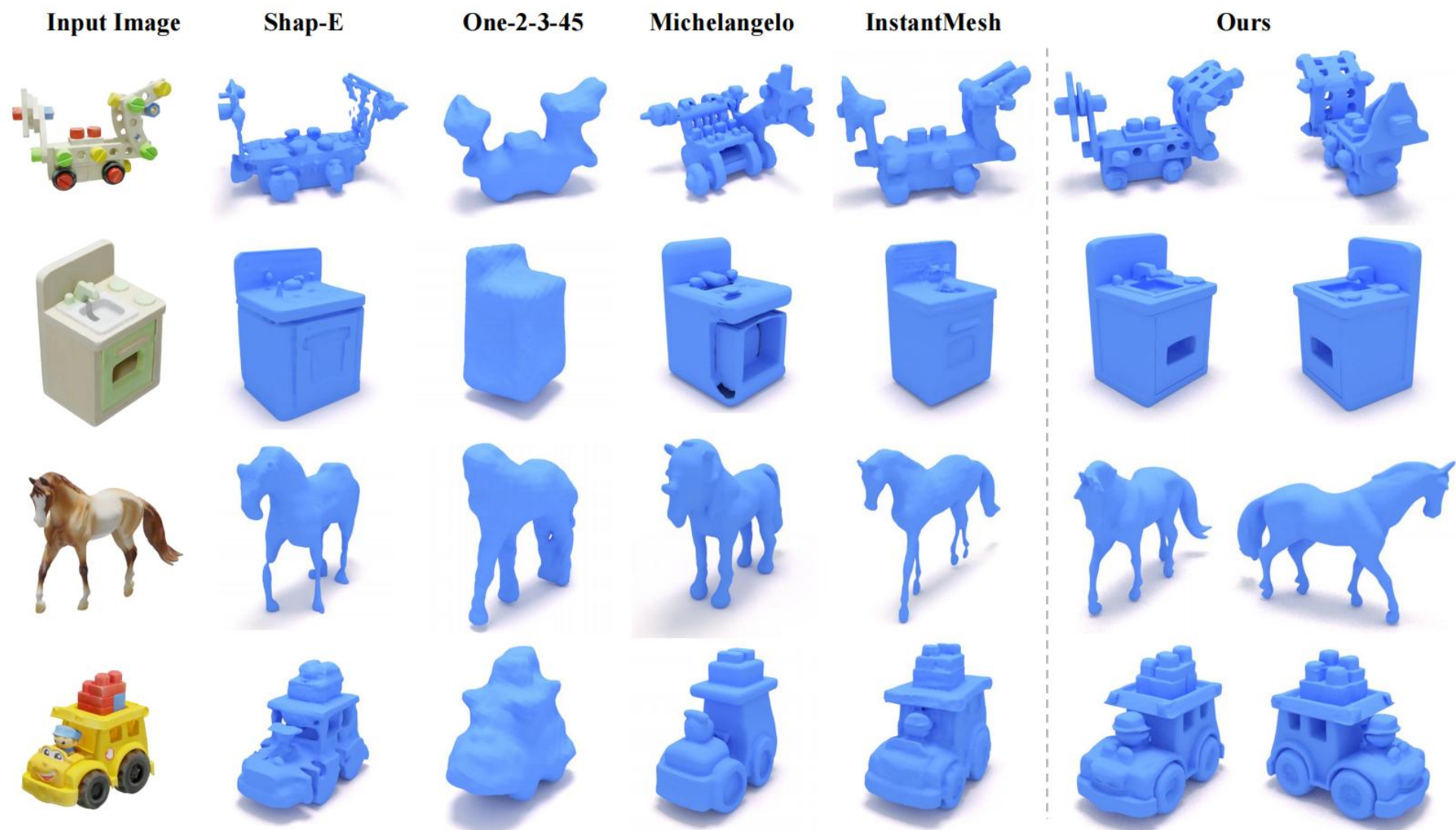
$$o(\mathbf{x}) = \begin{cases} 1, & \text{if } sdf(\mathbf{x}) < -s \\ 0.5 - \frac{0.5 \cdot sdf(\mathbf{x})}{s}, & \text{if } -s \leq sdf(\mathbf{x}) \leq s \\ 0, & \text{if } sdf(\mathbf{x}) > s \end{cases}$$

Image-Conditioned 3D Diffusion Transformer






































- The pre-trained CLIP is employed to incorporate semantic-level information from the conditioned images into the noisy latent tokens by cross-attention layers.
- The classification token extracted by CLIP is added to the time embedding to enhance semantic features.
- The pre-trained DINO-v2 is utilized to inject pixel-level information from the conditioned images into the noisy latent tokens by self-attention layers.

Image-to-3D Results



Text-to-3D Results

Input	Shap-E	One-2-3-45	Michelangelo	InstantMesh	Ours
					 
Spider-man, photorealistic					
					 
Flat flying dragon					
					 
A 3D model of an adorable cottage with a thatched roof					
					 
An astronaut is riding a horse					
					 
A pair of sunglasses					



Direct3D: Scalable Image-to-3D Generation via 3D Latent Diffusion Transformer



Shuang Wu^{1,2*} Youtian Lin^{2*} Feihu Zhang¹ Yifei Zeng^{1,2} Jingxi Xu¹ Philip Torr³ Xun Cao² Yao Yao²✉

Thanks!



<https://www.neural4d.com/research/direct3d>