NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

ByteDance

TAPS3D: Text-Guided 3D Textured Shape Generation from Pseudo Supervision

Jiacheng Wei*, Hao Wang*, Jiashi Feng, Guosheng Lin, Kim-Hui Yap

THU-AM-031

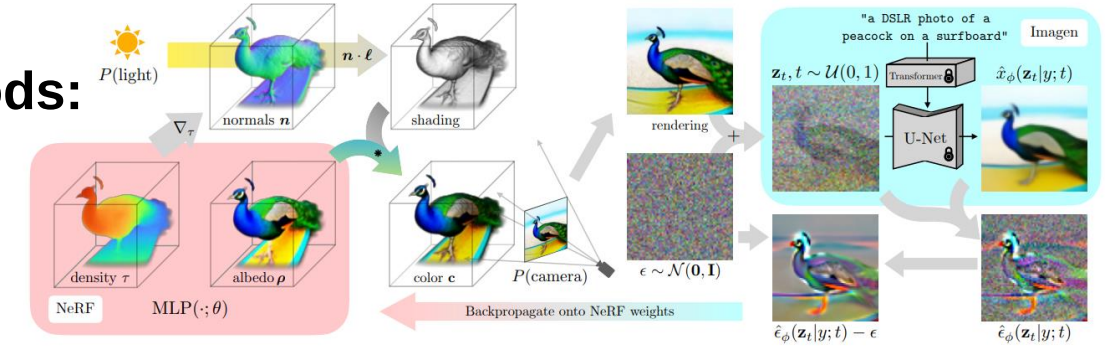
Motivations

Existing text-to-3D object generation methods:

(a) Optimization-based methods:

Pros: High fidelity.
Zero shot generation.

Cons: Slow and computationally expensive.
Poor geometry.

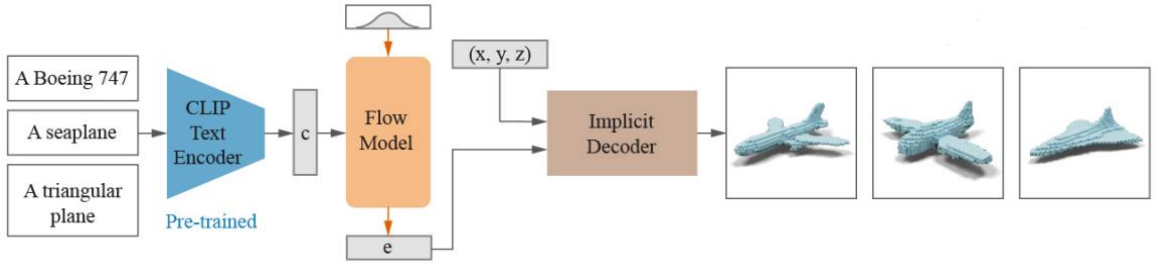


(a) Optimization-based methods

(b) Feed-forward methods:

Pros: Fast generation speed.

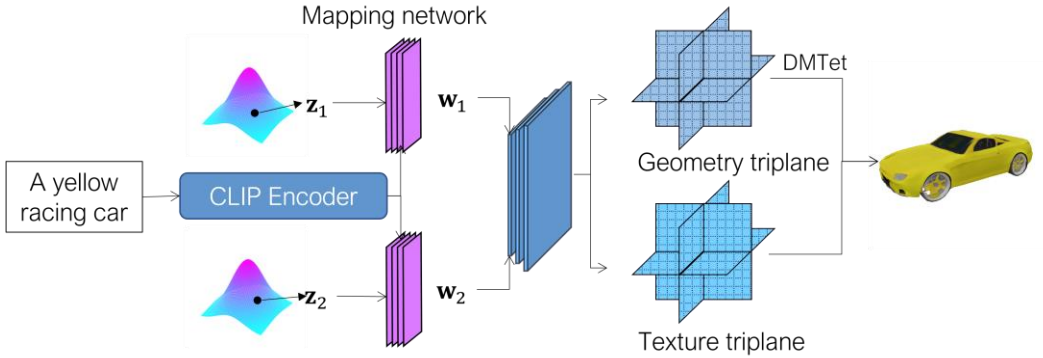
Cons: Low resolution voxels.
Paired text-3D training data.



(b) Existing feed-forward methods

Our method:

(1) Use only 2D image without paired text captions.
(2) Feed-forward, no test-time optimization.
(3) High quality and fidelity generation.

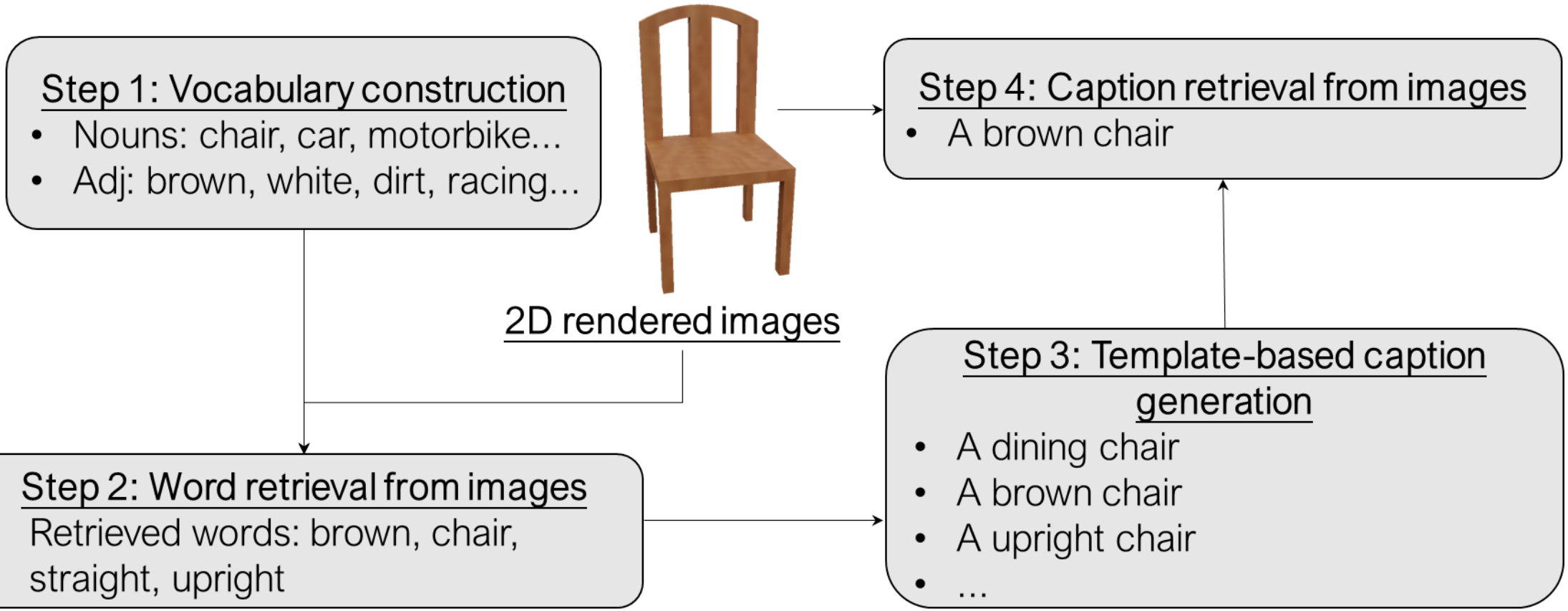


(c) Our method

Introduction

"a red hatchback car" *"a gray armless chair"* *"a wooden office table"* *"a black motorbike"*

Generate Pseudo Captions



Generate Pseudo Captions

a white automobile

a low race car

a ural motorbike

a brown motorcycle

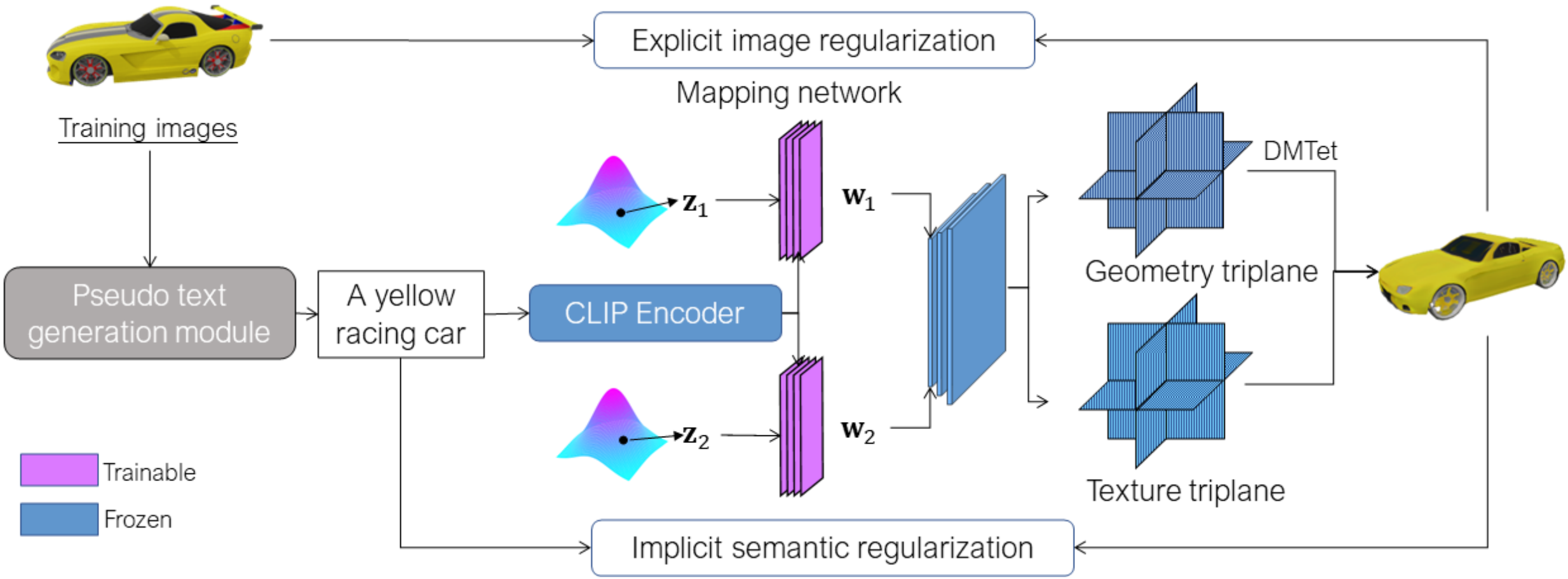
a stationary chair

a white ladder-back chair

a brown rectangular table

a structural table

Framework


Training images
Explicit image regularization
Mapping network
Pseudo text generation module
A yellow racing car
CLIP Encoder
z_1
z_2
w_1
w_2
DMTet
Geometry triplane
Texture triplane
Implicit semantic regularization
Trainable
Frozen

Cross-modal learning constraints

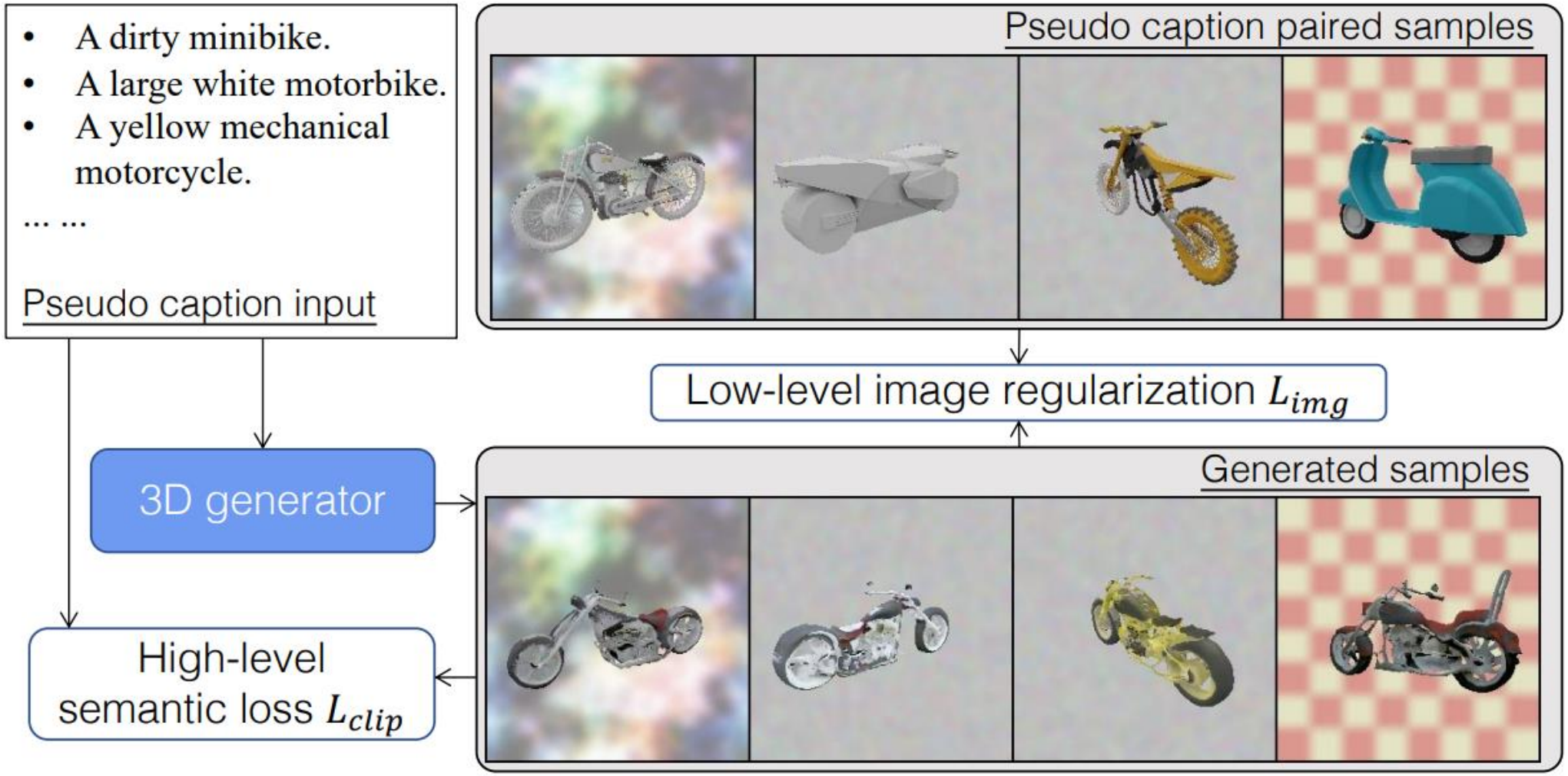
High-level semantic supervision

$$L_{clip} = 1 - \cos(E_i^{clip}(I_x), E_t^{clip}(t))$$

Low-level image regularization loss

$$L_{img} = 1 - \cos(E_i^{clip}(I_x), E_i^{clip}(I_x^{gt}))$$

Background augmentation



“a blue SUV”

“a yellow racing motorcycle”

“a red hatchback”

“a green flappy dirt bike”

“a yellow racing car”

“a red motorbike”

“a red dining table”

“a brown armchair”

“a wooden office desk”

“a gray armless chair”

“a silver metallic table”

“a green chair”

“a red {class}” *“a yellow {class}”* *“a green {class}”* *“a blue {class}”* *“a gray {class}”*

Interpolation

“an armed chair” *“an armless chair”*

“a red dining chair” *“a blue office chair”*

“a wooden office table” *“a metal dining table”*

“a red table” *“a blue table”*

“a yellow sports car” *“a blue SUV”*

“a yellow sports car” *“a yellow SUV”*

“a red car” *“a green car”*

“a red SUV” *“a blue SUV”*

“a yellow dirt bike” *“a red sportsbike”*

“a blue classic motorbike” *“a blue dual sport motorbike”*

Quantitative results

Table 1. Comparison with the existing work. We evaluate the rendered 2D images using Fréchet inception distance (FID). We downsample our result to the same resolution of CLIP-NeRF [50] for fair comparisons.

	Car		Chair	
	Resolution	FID	Resolution	FID
CLIP-NeRF [50]	256^2	67.8	128^2	48.4
Ours	256^2	**20.1**	128^2	**43.7**
Ours	1024^2	**21.7**	1024^2	**44.8**

Table 3. Comparison of 3D generation quality in FPD score.

Method	Chair	Table
TITG3SG [25]	1566.76	1639.68
CLIP-Forge [43]	825.96	3051.31
Ours	**342.23**	**1468.43**

Inference Speed

Method	Device	Output	Time
DreamFields [13]	TPU cores x8	Rendering	72 min
DreamFusion [34]	TPUv4 machine	Rendering	90 min
PureCLIPNeRF [18]	GTX 2080ti	Rendering	20 min
TITG3SG [25]	Telsa V100-32G	Voxel	2.21 sec
TITG3SG [25]	Telsa V100-32G	Mesh	24.44 sec
Ours	Telsa V100-32G	Rendering	0.05 sec
Ours	Telsa V100-32G	Mesh	7.09 sec