Supplementary Materials for "Lift3D: Synthesize 3D Training Data by Lifting 2D GAN to 3D Generative Radiance Field"

Leheng Li[1*] Qing Lian[2] Luozhou Wang[1] Ningning Ma[3] Ying-Cong Chen[1,2†]
[1]HKUST(GZ) [2]HKUST [3]NIO Autonomous Driving

The supplementary material is organized as follows. Sec. A describes the detailed configurations used in StyleGAN2 interpretation. Sec. B provides the network structure of NeRF used in Lift3D. Sec. C introduces the sampling parameters used in composition. Sec. D verifies the multi-view consistency of our 3D generation framework.

A. Interpretation of StyleGAN2

We derive disentanglement of latent space in StyleGAN2 [6] from GANSpace [5]. We disentangle StyleGAN2 and identify the first eight layers latents as the latents that control the object pose and the other eight layers latents as the latents that control the attributes except object shape. In Lift3D, our goal is to annotate these latents with pose labels for lifting process. We first use Blender EEVEE engine [3] to render a ShapeNet [2] model under 200 different views $\mathbf{P}$, ranging from $0-360°$ in azimuth, and $0-20°$ in elevation. The rendered images thus naturally contain accurate ground truth pose labels.

With a fixed pretrained StyleGAN2 [6], we initialize 200 different latents $\mathbf{z} \in \mathbb{R}^{512}$ from Gaussian distribution $\mathbf{Z} \in \mathcal{N}(0, \mathbf{I})$. The latents $\mathbf{z}$ are mapped to $\mathbf{w} \in \mathbb{R}^{16\times512}$ by the mapping network in StyleGAN2. We optimize the latents $\mathbf{w}$ using Adam [7] optimizer with learning rate of 1e-3 for 5000 iterations. The loss function is a simple $L1$ loss. After optimization, the first eight layers of latents $\mathbf{w}$ are annotated with "pseudo" pose labels, as the disentanglement and interpretation process is non-perfect.

B. Conditional NeRF

Our conditional NeRF mainly builds upon EG3D [1]. Fig. 1 depicts the detailed structure of NeRF in Lift3D. The overall network composes two parts: the mapping network and the synthesis network. The mapping network contains 8-layer MLPs and 16 affine transformations that maps the randomly sampled noise latents $z \in \mathbb{R}^{512}$ to $w \in \mathbb{R}^{16\times512}$. The latents are then modulate the synthesis network to generate orthogonal tri-planes that form the axis-aligned feature grid. The feature planes are of size $N \times N \times (C \times 3)$, where $N = 256$ denotes the spatial resolution and $C = 32$ the feature dimension. Any sampled 3D point $x \in \mathbb{R}^3$ of volumetirc rendering is projected onto the three feature planes to retrieve its interpolated feature vector. The final SIREN-based [13] MLP that condition on the mean of w then convert the feature vector to the RGB and density value.

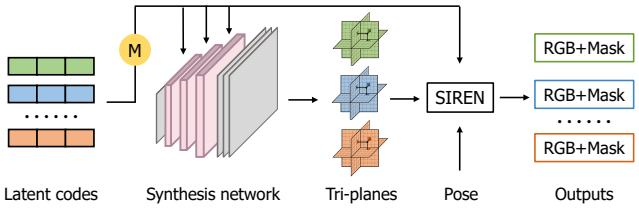

Figure 1. Network structure of NeRF in **Lift3D**, which converts latent codes to images and masks. $\mathbf{M}$ denotes the mapping network that maps latents $\mathbf{z} \in \mathbb{R}^{512}$ to $\mathbf{w} \in \mathbb{R}^{16\times512}$. $\mathbf{w}$ are fed in AdaIN [4] to modulate the synthesis network to map the constant input to the tri-planes. SIREN is a single-layer MLP converts feature vector to RGB and density value.

A normalized 3D bounding box [14] is utilized to filter out the background sampling points. During ray casting, we utilize an AABB-ray intersection algorithm [8] to find the nearest and furthest hitting points of 3D bounding box. The same parametrization can also be found in [9, 11]. The normalized sampling points lied in $[-1, 1]$ are projected to exactly cover the content of the tri-plane for tight parametrization.

We further compare the lifting results of our shared NeRF with isolated NeRF in Fig. 2. Given the images generated from StyleGAN2, we ablate two lifting processes: isolating training and joint training. The isolated NeRF is trained by optimizing the randomly sampled tri-planes and a single-layer SIREN to fit multi-view images. The shared NeRF is our proposed lifting process. We jointly optimize mapping network, synthesis network, SIREN, and a set of latents in the same time. The learned mapping network successfully maps randomly sampled latents $\mathbf{z}$ to object prior latent space $\mathbf{w}$ that can decoded by synthesis network to output meaning shape and appearance.

*Work done during an internship at NIO Autonomous Driving.
†Corresponding author.

C. Sample Parametrization

The final sampling pose $\mathbf{P}'$ can be written as $(x, y, z, l, w, h, \theta)$, where x, y, z is position of 3D bounding box, l, w, h represent length, width, height of bounding box, θ is rotation along y axis. We detail the parametrization of $\mathbf{P}'$ in Tab. 1.

Pose	Distribution	Parameters
x	Uniform	$[-20m, 20m]$
y	Gaussian	$\mu = height, \sigma = 0.2$
z	Uniform	$[5m, 45m]$
l	Gaussian	$\mu = l_{mean}, \sigma = 0.5$
w	Gaussian	$\mu = w_{mean}, \sigma = 0.5$
h	Gaussian	$\mu = h_{mean}, \sigma = 0.5$
θ	Gaussian	$\mu = \pm\pi/2, \sigma = \pi/2$

Table 1. Detailed sampling parameters during composition. l_{mean}, w_{mean} and h_{mean} is the mean value of length, width, height of 3D box obtained from the statistic of datasets.

D. Multi-view Consistency

We additionally compare multi-view consistency of our 3D generation framework with 3D generative model composed of a 2D upsampler. We use Reprojection Error (RE) proposed in [12] to evaluate the consistency of generated images. We randomly choose two adjacent views, then render the images and corresponding depth map of the same object. We use the predicted depth to warp the image from one view to the other. The error is calculated between the predicted image and the warped image on 10K pairs.

Method	Reprojection Error
GIRAFFE [10]	0.225
GIRAFFE HD [15]	0.207
Ours	**0.079**

Table 2. Comparison of multi-view consistency measured by Reprojection Error.

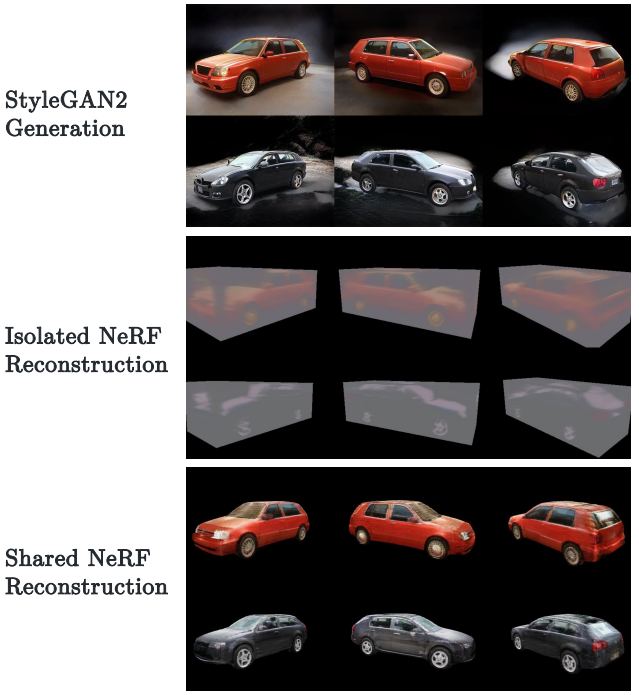

Figure 2. Qualitative comparison of our lifting process. Compared with training a large number of individual NeRFs, our method learns the object prior in the mapping network and synthesis network, which allows to use a single-layer MLP to generate diverse object radiance field.

References

[1] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *CVPR*, 2022. 1

[2] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 1

[3] Blender Online Community. Blender—a 3d modelling and rendering package. *Blender Foundation*, 2018. 1

[4] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 1

[5] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. In *NeurIPS*, 2020. 1

[6] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *CVPR*, 2020. 1

[7] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[8] Alexander Majercik, Cyril Crassin, Peter Shirley, and Morgan McGuire. A ray-box intersection algorithm and efficient dynamic voxel rendering. *Journal of Computer Graphics Techniques Vol*, 7(3):66–81, 2018. 1

[9] Norman Müller, Andrea Simonelli, Lorenzo Porzi, Samuel Rota Bulò, Matthias Nießner, and Peter Kontschieder. Autorf: Learning 3d object radiance fields from single view observations. In *CVPR*, 2022. 1

[10] Michael Niemeyer and Andreas Geiger. GIRAFFE: Representing scenes as compositional generative neural feature fields. In *CVPR*, 2021. 2

[11] Julian Ost, Fahim Mannan, Nils Thuerey, Julian Knodt, and Felix Heide. Neural scene graphs for dynamic scenes. In

Figure 3. Novel view synthesis result of our generated objects.

CVPR, 2021. 1

[12] Zifan Shi, Yinghao Xu, Yujun Shen, Deli Zhao, Qifeng Chen, and Dit-Yan Yeung. Improving 3d-aware image synthesis with a geometry-aware discriminator. *arXiv preprint arXiv:2209.15637*, 2022. 2

[13] Vincent Sitzmann, Julien N.P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *NeurIPS*, 2020. 1

[14] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J. Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *CVPR*, 2019. 1

[15] Yang Xue, Yuheng Li, Krishna Kumar Singh, and Yong Jae Lee. Giraffe hd: A high-resolution 3d-aware generative model. In *CVPR*, 2022. 2