2D GANs Meet Unsupervised Single-view 3D Reconstruction — Supplementary Material

Feng Liu, Xiaoming Liu

Michigan State University, Computer Science & Engineering
{liufeng6,liuxm}@msu.edu

In this supplementary material, we provide:

i) Network structures.

ii) Additional visual results, including quantitative analysis of generated Images, additional qualitative single-view 3D reconstruction, multi-view generator comparisons, predicted uncertainty maps and failure cases.

1 Network Structures

Image Encoder $\mathcal{E}$ As shown in Fig. 1, we use an image encoder $\mathcal{E}$, which is a modified ResNet-18, as a hypernetwork to predict the network parameters θ of the geometry MLP $\mathcal{F}$. The network takes an image with the resolution 128×128 as input.

Uncertainty Prediction Module We leverage the intermediate features from the first three convolutional blocks of the image encoder, resulting in a set of feature maps, to predict the uncertainty maps. Since the features in layers are smaller in dimension than the original image, we reshape them to the original size with bilinear interpolation (Fig. 1). The uncertainty prediction module is implemented using two convolution layers with sizes of $1 \times 1 \times 128$ and $1 \times 1 \times 1$.

Hyperparameter Prediction Module and the Geometry MLP $\mathcal{F}$ Both the hyperparameter prediction module and the geometry MLP $\mathcal{F}$ are composed of 4 fully-connected (FC) blocks/layers. Each block maps a code vector $\mathbb{R}^{256}$ to the parameters θ_*, which consists the weights $\mathbf{W} = \mathbf{g}\frac{\mathbf{V}}{||\mathbf{V}||}$, where $\mathbf{V} \in \mathbb{R}^{l_m \times l_n}$, $\mathbf{g} \in \mathbb{R}^{l_n}$, and biases $\mathbf{b} \in \mathbb{R}^{l_n}$ of the corresponding FC layer. The detailed architecture is depicted in Fig. 2a. $\mathcal{F}$ takes a 3D point with positional encodering [1] as input and outputs the singed distance value s and local geometry feature $\mathbf{f} \in \mathbb{R}^{256}$. There are skip connection from the input concatenated vector every hidden layers. Following [2], each hidden layer is applied with $softplus$ activation.

The Texture MLP $\mathcal{G}$ Similarly, the texture MLP architecture is composed of 5 FC layers (Fig. 2b). The inputs to the network include the surface point $\hat{\mathbf{x}} \in \mathbb{R}^3$, the surface normal $\hat{\mathbf{n}} \in \mathbb{R}^3$, the local geometry feature $\mathbf{f}$, and the viewing direction $\mathbf{v} \in \mathbb{R}^3$. The output is a 3-channel RGB value $\mathbf{c}$.

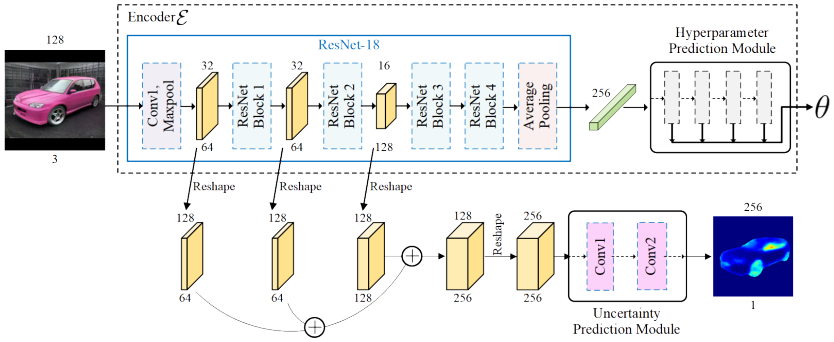
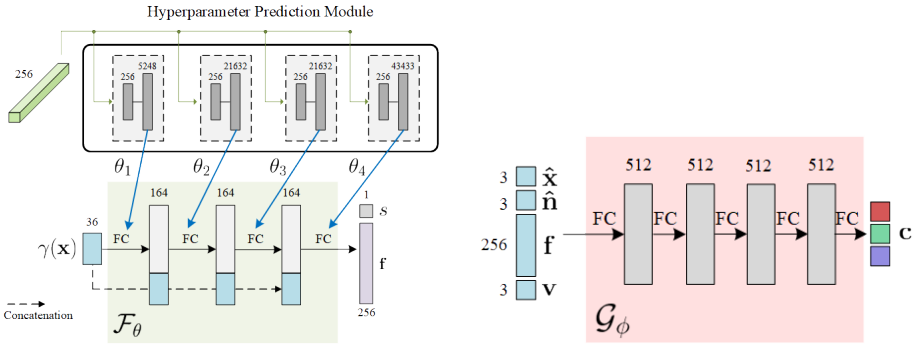


Fig. 1: Image encoder network (modified from ResNet-18) and the uncertainty prediction module structures.



(a) (b)

Fig. 2: (a) The detailed structure of the hyperparameter prediction module and the geometry MLP $\mathcal{F}_\theta$. The hyperparameter prediction module includes 4 fully-connected blocks. Each block is composed of 2 fully connected layers, denotes as "FC". The activation function in the hidden layer of each blocks is ReLU. The geometry MLP $\mathcal{F}_\theta$ is composed of 4 FC layers. It takes a 3D point with positional encoding as input and outputs the signed distance value s and a local geometry feature $\mathbf{f}$. (b) The texture MLP $\mathcal{G}_\phi$ is composed of 5 fully connected layers. Specifically, it takes the surface points $\hat{\mathbf{x}}$, along with its surface normals $\hat{\mathbf{n}}$, local geometry feature $\mathbf{f}$, viewing direction $\mathbf{v}$, and outputs the RGB color values $\mathbf{c}$. *ReLU* activation is applied to the first 3 FC layers while the final value is obtained with *Tanh*.

2 Additional Visual Results

Quantitative Analysis of Generated Images To quantitative analysis of our multi-view generation, we collected 41 multi-view images one car instance for the novel view synthesis evaluation. For each image, we first compute its

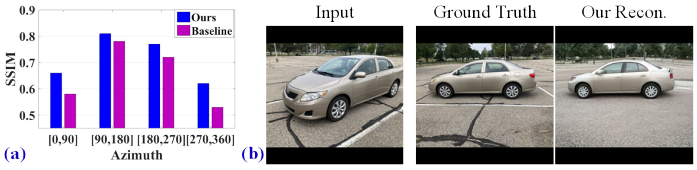


Fig. 3: Test-set novel view synthesis of our multi-view generator.

styleGAN latent code $\mathcal{W}^+$ via the GAN inversion technique. Then, we randomly select one image as the source, and the other 40 samples as the **ground truth** reference images. Given the source latent code as input, we then apply our multi-view generation to synthesize images with the reference viewpoint and measure the similarity between the generated image and the ground truth via SSIM. As shown in Fig. 3 (a), our method outperforms baseline [62] in all azimuth groups. Fig. 3 (b) shows one example.

Additional Qualitative Single-view 3D Reconstruction We provide more reconstruction results on airplanes, cars, birds, horses, motorbikes and potted plants in Fig. 4 and 5 (please also refer to the supplementary video). In Fig. 6a, we show that our approach can not only recover 3D shapes, but also predict their plausible texture from a single image.

Multi-view Generator Comparisons Fig. 6b shows multi-view generator comparisons on more categories. Compared to StyleGANRender [3] (Baseline), our generated images show more consistency in object shapes across views.

Predicted Uncertainty Maps In Fig. 7, we provide predicted uncertainty maps of more categories. As can be observed, the uncertainty maps are able to visualize the unreliable/inconsistent areas in the GAN-generated multi-view pseudo images.

Failure Cases Our approach is unable to learn a chair model since the StyleGAN cannot converge to satisfying results on chair category, resulting in distortions in the generated chairs (Fig. 8).

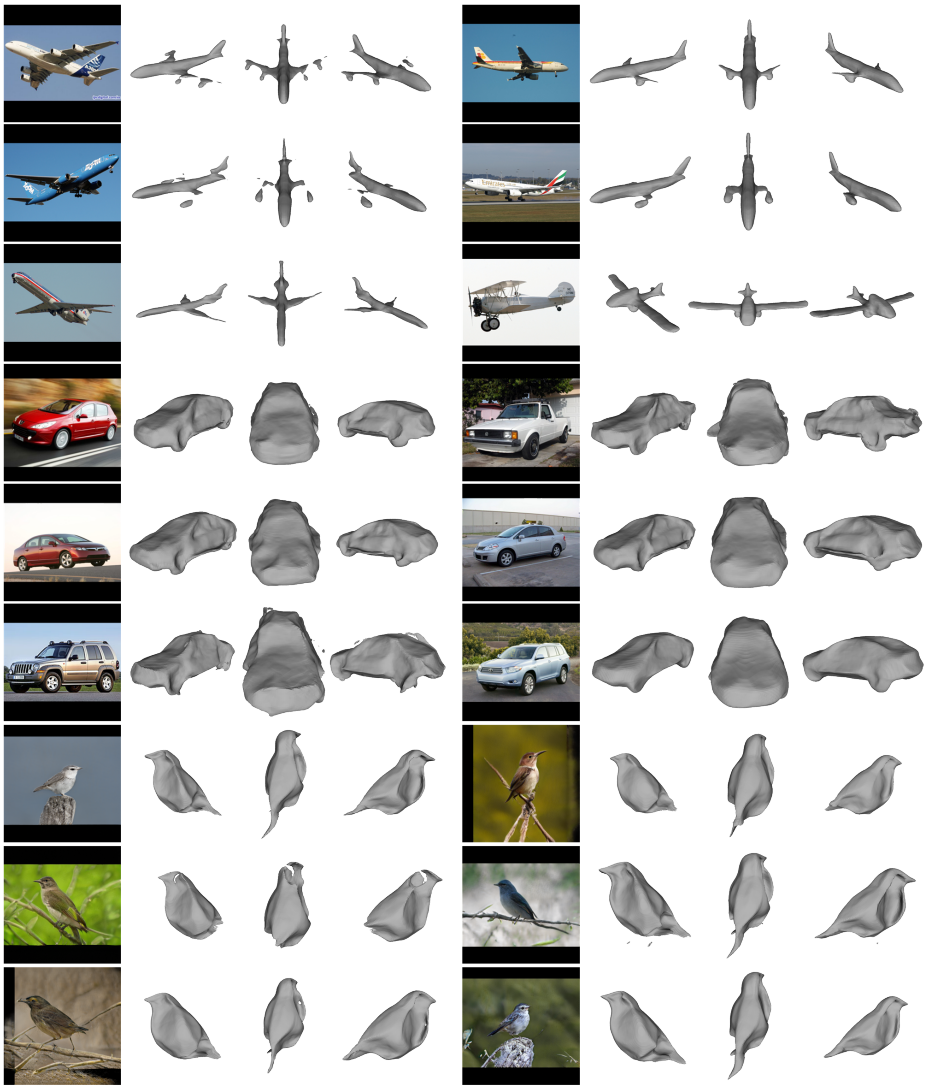


Fig. 4: Qualitative reconstruction results of airplanes, cars and birds.

Fig. 5: Qualitative reconstruction results of horses, motorbikes and potted plants.

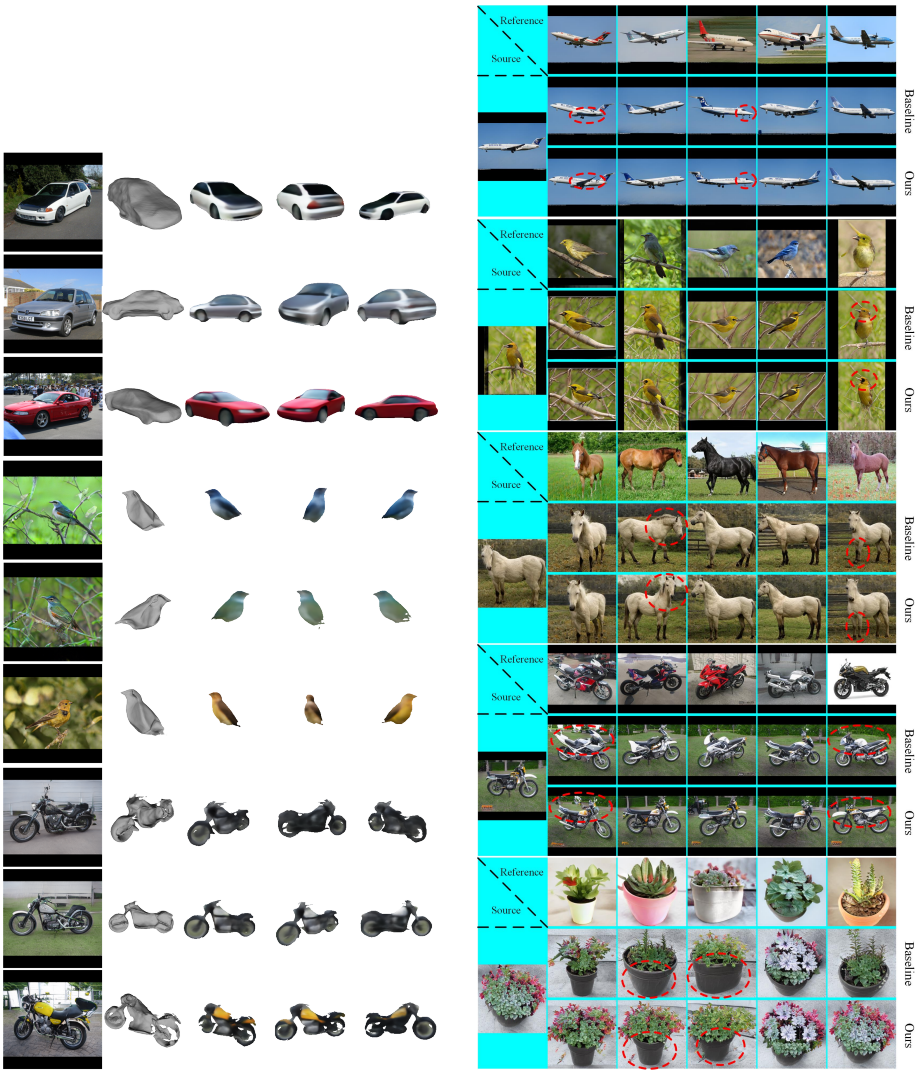


Fig. 6: (a) Qualitative texture predictions. For each test image on the left, we show the predicted mesh and the textured mesh from multiple views. It can be observed that our approach can not only estimate 3D shapes, but also produce plausible 3D textures. (b) Additional multi-view generator comparisons between StyleGANRender [3] (Baseline) and our approach (Ours). It can be observed, besides the viewpoint, the baseline perceives shape cues from the references (red circle, best view in zoom in). While our generated images show more consistency in object shape across views, which benefits shape learning.

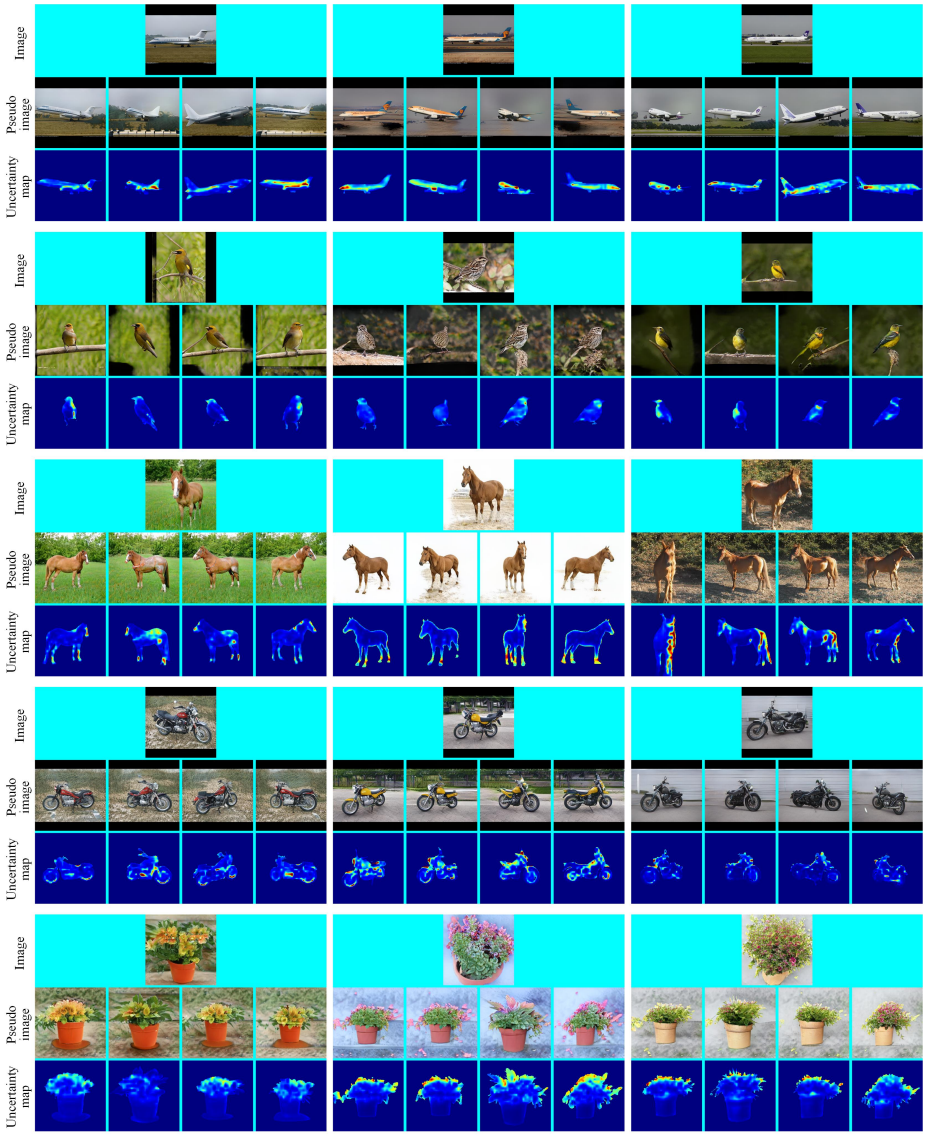


Fig. 7: Additional predicted uncertainty maps of airplanes, birds, horses, motorbike, and potted plants.

Fig. 8: Chair images produced by a pre-trained StyleGAN model.

References

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017) 1
2. Yariv, L., Kasten, Y., Moran, D., Galun, M., Atzmon, M., Basri, R., Lipman, Y.: Multiview neural surface reconstruction by disentangling geometry and appearance. In: NeurIPS (2020) 1
3. Zhang, Y., Chen, W., Ling, H., Gao, J., Zhang, Y., Torralba, A., Fidler, S.: Image GANs meet differentiable rendering for inverse graphics and interpretable 3D neural rendering. In: ICLR (2021) 3, 6