

HyperNeRFGAN: Hypernetwork approach to 3D NeRF GAN

Adam Kania¹ Artur Kasymov¹ Maciej Zieba² Przemysław Spurek¹

Abstract

Recently, generative models for 3D objects are gaining much popularity in VR and augmented reality applications. Training such models using standard 3D representations, like voxels or point clouds, is challenging and requires complex tools for proper color rendering. In order to overcome this limitation, Neural Radiance Fields (NeRFs) offer a state-of-the-art quality in synthesizing novel views of complex 3D scenes from a small subset of 2D images.

In the paper, we propose a generative model called HyperNeRFGAN, which uses hypernetworks paradigm to produce 3D objects represented by NeRF. Our GAN architecture leverages a hypernetwork paradigm to transfer gaussian noise into weights of NeRF model. The model is further used to render 2D novel views, and a classical 2D discriminator is utilized for training the entire GAN-based structure. Our architecture produces 2D images, but we use 3D-aware NeRF representation, which forces the model to produce correct 3D objects. The advantage of the model over existing approaches is that it produces a dedicated NeRF representation for the object without sharing some global parameters of the rendering component. We show the superiority of our approach compared to reference baselines on three challenging datasets from various domains.

1. Introduction

Generative Adversarial Nets (GANs) (Goodfellow et al., 2014) allow us to generate high-quality 2D images (Yu et al., 2017; Karras et al., 2017; 2019; 2020; Struski et al., 2022). On the other hand, maintaining similar quality for

^{*}Equal contribution ¹Faculty of Mathematics and Computer Science, Jagiellonian University 6 Lojasiewicza Street, 30-348 Kraków, Poland ²Department of Artificial Intelligence, University of Science and Technology Wyspiańskiego 27, 50-370, Wrocław, Poland. Correspondence to: Przemysław Spurek <przemyslaw.spurek@uj.edu.pl>.

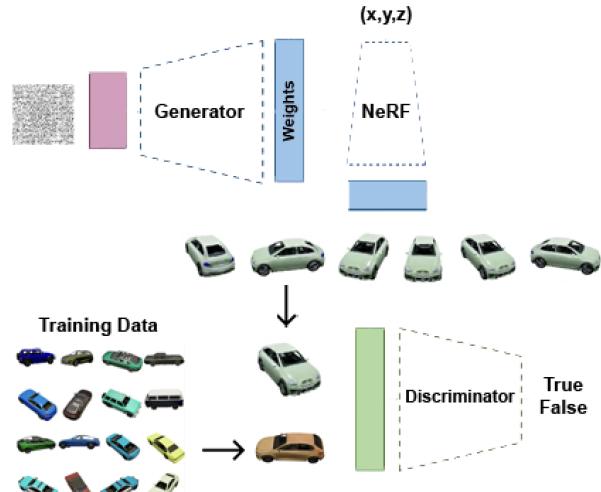


Figure 1. HyperNeRFGAN architecture leverages a hypernetwork paradigm to transfer gaussian noise into weights of NeRF model. After that, we render 2D novel views by NeRF and use a classical 2D discriminator. Our architecture produces 2D images, but we use 3D-aware NeRF representation, which forces the model to produce correct 3D objects.

3D objects is challenging. It is mainly caused by using 3D representations like voxels and point clouds that require massive deep architectures and have problems with proper color rendering.

We can solve this problem by operating directly on 2D image space. We expect our approach to extract information from unlabeled 2D views to obtain 3D shapes. To obtain such effects, we can use Neural Radiance Fields (NeRFs) (Mildenhall et al., 2021), which allow synthesizing novel views of complex 3D scenes from a small subset of 2D images. Based on the relations between those base images and computer graphics principles, such as ray tracing, this neural network model can render high-quality images of 3D objects from previously unseen viewpoints.

Unfortunately, it is not trivial how to use NeRF representation with GAN-type architecture. The most challenging problem is connected with the conditioning mechanism (Rebain et al., 2022) dedicated to NeRF. Therefore, most models



Figure 2. Comparison of HyperNeRFGAN and HoloGAN, GRAF, π -GAN on CARLA. We obtain similar results to π -GAN, but we have a better value of FID score, see Tab 2.

use SIREN (Sitzmann et al., 2020) instead of NeRF, where we can naturally add conditioning. But the quality of the 3D object is slice worst than in NeRF. In GRAF (Schwarz et al., 2020) and π -GAN (Chan et al., 2021), authors propose a model which uses SIREN and a conditioning mechanism to produce implicit representation. Such solutions give promising results, but it is not trivial how to use NeRF instead of SIREN in such solutions. In Fig. 2 we present a qualitative comparison between our model, GRAF (Schwarz et al., 2020) and π -GAN (Chan et al., 2021). As we can see, our model can model the transparency of glass.

In the paper, we propose a generative model called HyperNeRFGAN¹, which combines the hypernetworks paradigm and NeRF representation. Hypernetworks, introduced in (Ha et al., 2016), are defined as neural models that generate weights for a separate target network solving a specific task. Our GAN-based model leverages a hypernetwork paradigm to transfer gaussian noise into weights of NeRF (see Fig. 1). After that, we render 2D novel views by NeRF and use a classical 2D discriminator to train the entire GAN-based structure in implicit form. Our architecture produces 2D images, but we use 3D-aware NeRF representation, which forces the model to produce correct 3D objects.

Our contributions to this paper include the following:

- We introduce the NeRF-based implicit GAN architecture - the first GAN model for generating high-quality 3D NeRF representations.
- We show that utilizing the hypernetwork paradigm for NeRFs leads to a better quality of 3D representations

¹The source code is available at: <https://github.com/gmum/HyperNeRFGAN>

than SIREN-based architectures.

- Our model allows 3D-aware image synthesis from unsupervised 2D images.

2. Related Work

Neural representations and rendering 3D objects can be represented by using many different approaches, including voxel grids (Choy et al., 2016), octrees (Häne et al., 2017), multi-view images (Arsalan Soltani et al., 2017; Liu et al., 2022), point clouds (Achlioptas et al., 2018; Shu et al., 2022; Yang et al., 2022), geometry images (Sinha et al., 2016), deformable meshes (Girdhar et al., 2016), and part-based structural graphs (Li et al., 2017).

The above representations are discreet, which causes some problems in real-life applications. Contrary, we can represent 3D objects as a continuous function (Dupont et al., 2022). In practice implicit occupancy (Chen & Zhang, 2019; Mescheder et al., 2019; Peng et al., 2020), distance field (Michalkiewicz et al., 2019; Park et al., 2019) and surface parametrization (Yang et al., 2019; Spurek et al., 2020; 2022; Cai et al., 2020) models use a neural network to parameterize a 3D object. In such a case, we do not have a fixed number of voxels, points, or vertices, but we represent shapes as a continuous function.

These models are limited by their requirement of access to ground truth 3D geometry. Implicit neural representations (NIR) have been proposed to solve such a problem. Such architectures can reconstruct 3D structures from multi-view 2D images (Mildenhall et al., 2021; Niemeyer et al., 2020; Tewari et al., 2020).

The two most important methods are NeRF (Mildenhall

et al., 2021) and SIREN (Sitzmann et al., 2020). NeRF uses volume rendering (Kajiya & Von Herzen, 1984) for reconstructing a 3D scene using neural radiance and density fields to synthesize novel views. SIREN replaced the popular ReLU activation function with sine functions with modulated frequencies. Most NeRF and SIREN-based methods focus on a single 3D object or scene. In practice, we overfit individual objects or scenes. In our paper, we focus on generating 3D models represented by NeRF.

Single-View Supervised 3D-Aware GANs Generative Adversarial Nets (GANs) (Goodfellow et al., 2014) allow for the generation of high-quality images (Yu et al., 2017; Karras et al., 2017; 2019; 2020; Struski et al., 2022). However, GANs operate on 2D images and ignore the 3D nature of our physical world. Therefore, it is important to use 3D structures of objects to generate images and 3D objects.

The first approach for 3D-aware image syntheses like Visual Object Networks (Zhu et al., 2018) and PrGANs (Gadelha et al., 2017) first generating a voxelized 3D shape using a 3D-GAN (Wu et al., 2016) and then projecting it into 2D.

HoloGAN (Nguyen-Phuoc et al., 2019) and BlockGAN (Nguyen-Phuoc et al., 2020) work in a similar fusion but use implicit 3D representation for modeling 3D representation of the world. Unfortunately, using an explicit volume representation has constrained their resolution (Lunz et al., 2020). In (Szabó et al., 2019), authors propose using meshes to represent 3D geometry. On the other hand, in (Liao et al., 2020) uses collections of primitives for image synthesis.

In GRAF (Schwarz et al., 2020) and π -GAN (Chan et al., 2021), authors use implicit neural radiance fields for 3D-aware image and geometry generation. In our work, we use NeRF instead of SIREN and hypernetwork paradigm instead of a conditioning procedure.

Authors use a shading-guided pipeline in ShadeGAN (Pan et al., 2021), and in GOF (Xu et al., 2021), they gradually shrink the sampling region of each camera ray. GI-RAFFE (Niemeyer & Geiger, 2021) we first generate low-resolution feature maps. In the second step, we passed representation to a 2D CNN to generate outputs at a higher resolution.

In StyleSDF (Or-El et al., 2022), authors merge an SDF-based 3D representation with a StyleGAN2 for image generation. In (Chan et al., 2022), authors use StyleGAN2 generator and tri-plane representation of 3D objects. Such methods outperform other methods in the quality of generated objects but are extremely hard to train.

Hypernetworks + generative modeling Combining hypernetworks and generative models is not new. In (Ratzlaff & Fuxin, 2019; Henning et al., 2018) authors built GANs

to generate parameters of a neural network dedicated to regression or classification tasks. HyperVAE (Nguyen et al., 2020) is designated to encode any target distribution by producing generative model parameters given distribution samples. HCNAF (Oechsle et al., 2019) is a hypernetwork that produces parameters for a conditional autoregressive flow model (Kingma et al., 2016; Oord et al., 2018; Huang et al., 2018). In (Skorokhodov et al., 2021) authors proposed INR-GAN (Skorokhodov et al., 2020) uses a hypernetwork to produce a continuous representation of images. The hypernetwork can modify the shared weights by the low-cost mechanism of factorized multiplicative modulation.

3. HyperNeRFGAN: Hypernetwork for Generating NeRF representations

In this section, we present HyperNeRFGAN - a novel generative model for 3D objects. The main idea of the proposed approach is that generator serves as a hypernetwork (Ha et al., 2016) and transforms the noise vector sampled from the known distribution to the weights of the target model. Compared to previous works (Skorokhodov et al., 2020), the target model is represented by NeRF (Mildenhall et al., 2021) 3D representation of the object. Consequently, it is possible to generate many images of the object from various perspectives in a controllable manner. Moreover, thanks to the NeRF-based image rendering, the discriminator operates on 2D images generated from multiple perspectives, compared to GAN-based models fed by complex 3D structures. In this section, we first briefly discuss the basic concepts used in our approach, and further, we focus on the architecture and training details.

Hypernetwork Hypernetworks, introduced in (Ha et al., 2016), are defined as neural models that predict weights for a different target network designed to solve a specific task. This approach reduces the number of trainable parameters compared to standard methods that inject additional information into the target model using a single embedding. A significant reduction of the size of the target model can be achieved since it is not sharing the global weights, but they are returned by the hypernetwork. Making an analogy between Hypernetworks and generative models, the authors of (Sheikh et al., 2017), use this mechanism to generate a diverse set of target networks approximating the same function.

Hypernetworks are widely used in many domains, including few-shot problems (Sendera et al., 2023) or probabilistic regression scenarios (Zieba et al., 2020). Various methods also use them to produce a continuous representation of 3D objects (Spurek et al., 2020; 2022). For instance, HyperCloud (Spurek et al., 2020) represents a 3D point cloud as a classical MLP that serves as a target model and transforms

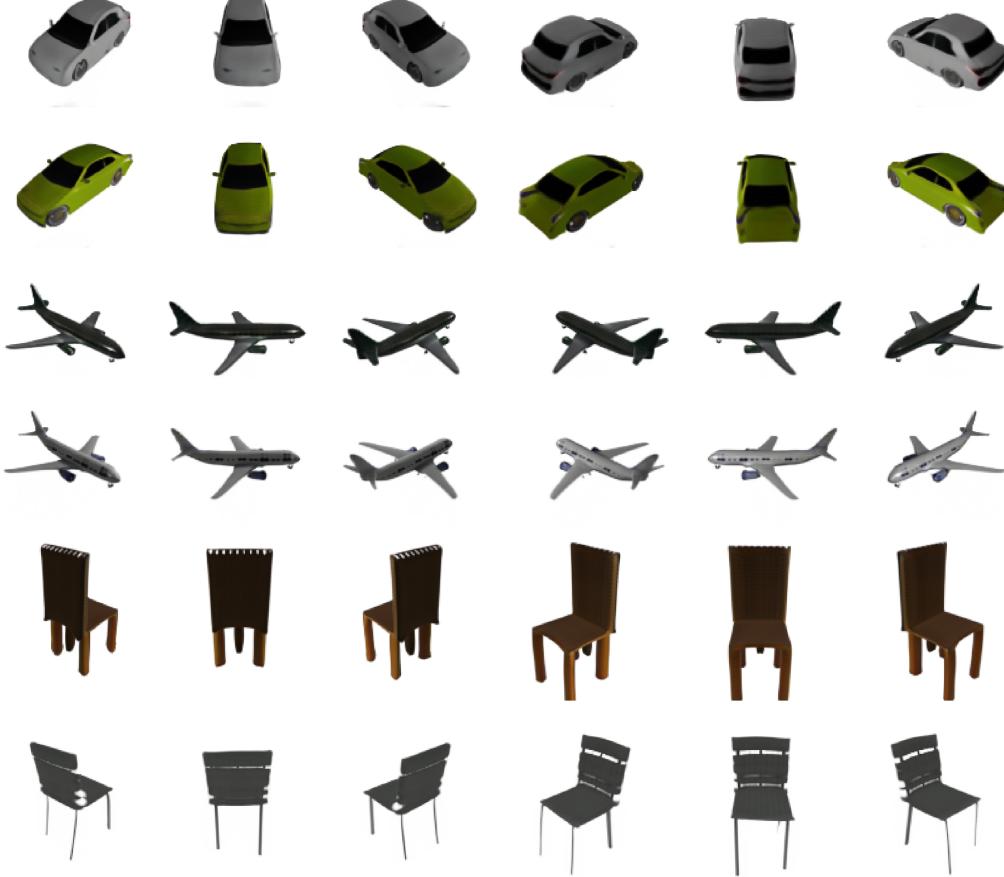


Figure 3. Elements generated by model train on three classes of ShapeNet (car, plane, chairs).

points from a uniform distribution on the gaussian ball to the point clouds that represent the desired shape. , In (Spurek et al., 2022), the target model is represented by a Continuous Normalizing Flow (Grathwohl et al., 2018), the generative model that creates the point cloud from the assumed base distribution in 3D space.

GAN is a framework for training deep generative models using a minimax game. The goal is to learn a generator distribution $P_G(x)$ that matches the real data distribution $P_{data}(x)$. GAN learns a generator network \mathcal{G} that produces samples from the generator distribution P_G by transforming a noise variable $z \sim P_{noise}(z)$ (usually Gaussian noise $N(0, I)$) into a sample $\mathcal{G}(z)$. The generator learns by playing against an adversarial discriminator network \mathcal{D} aiming to distinguish between samples from the true data distribution P_{data} and the generator's distribution P_G . More formally, the minimax game is given by the following expression:

$$\min_{\mathcal{G}} \max_{\mathcal{D}} [V(\mathcal{D}, \mathcal{G}) =$$

$$\mathbb{E}_{x \sim P_{data}} [\log \mathcal{D}(x)] + \mathbb{E}_{z \sim P_{noise}} [\log(1 - \mathcal{D}(\mathcal{G}(z)))]].$$

The main advantage over other models is producing sharp images indistinguishable from real ones. GANs are impressive regarding the visual quality of images sampled from the model, but the training process is often challenging and unstable. This phenomenon is caused by direct optimization of the training objective is intractable, and the model is usually trained by optimizing the parameters of the discriminator and generator in alternating steps.

In recent years, many researchers focused on modifying the vanilla GAN procedure to improve the stability of the training process. Some modifications were based on changing the objective function to Wasserstein distance (WGAN) (Arjovsky et al., 2017), restrictions on the gradient penalties (Gulrajani et al., 2017; Kodali et al., 2017), Spectral Normalization (Miyato et al., 2018), or imbalanced learning rate for generator and discriminator(Gulrajani et al., 2017;

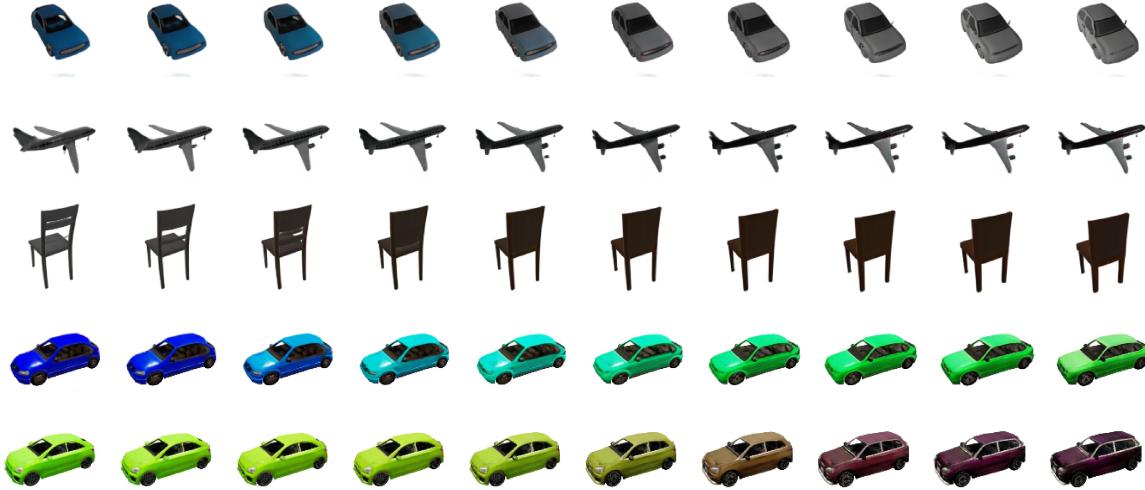


Figure 4. Linear interpolation examples generated with models trained on images of cars, planes, and chairs from ShapeNet (three first lines) and CARL data set (last two rows).

Miyato et al., 2018). The model architectures were also more deeply explored by utilizing self-attention mechanisms SAGAN (Zhang et al., 2018), and progressively growing ProGAN (Karras et al., 2017) and style-gan architectures StyleGAN (Karras et al., 2019).

INR-GAN Implicit Neural Representation GAN (Srokohodov et al., 2020) is a variant of the GAN-based model that utilizes hypernetworks to generate parameters for the target model instead of direct image generation. The target model, represented by simple MLP, returns the color in RGB format for a given pixel location. The model is very close architecturally to StyleGAN2 (Karras et al., 2020) and has clear advantages over the direct approach, mainly because using INR-GAN enables generating images without assuming the arbitrarily given resolution.

NeRF representation of 3D objects NeRFs (Mildenhall et al., 2021) represent a scene using a fully-connected architecture. As the input, NeRF takes a 5D coordinate (spatial location $\mathbf{x} = (x, y, z)$ and viewing direction $\mathbf{d} = (\theta, \psi)$) and it outputs an emitted color $\mathbf{c} = (r, g, b)$ and volume density σ .

A vanilla NeRF uses a set of images for training. In such a scenario, we produce many rays passing through the image and a 3D object represented by a neural network. NeRF approximates this 3D object with a MLP network:

$$F_\Theta : (\mathbf{x}, \mathbf{d}) \rightarrow (\mathbf{c}, \sigma),$$

and optimizes its weights Θ to map each input 5D coordinate to its corresponding volume density and directional emitted color.

The loss of NeRF is inspired by classical volume rendering (Kajiya & Von Herzen, 1984). We render the color of all rays passing through the scene. The volume density $\sigma(\mathbf{x})$ can be interpreted as the differential probability of a ray. The expected color $C(\mathbf{r})$ of camera ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ (where \mathbf{o} is ray origin and \mathbf{d} is direction) can be computed with an integral.

In practice, this continuous integral is numerically estimated using a quadrature. We use a stratified sampling approach where we partition our ray $[t_n, t_f]$ into N evenly-spaced bins and then draw one sample uniformly at random from within each bin:

$$t_i \sim \mathcal{U}[t_n + \frac{i-1}{N}(t_f - t_n), t_n + \frac{i}{N}(t_f - t_n)].$$

We use these samples to estimate $C(\mathbf{r})$ with the quadrature rule discussed in the volume rendering review by Max (Max, 1995):

$$\hat{C}(\mathbf{r}) = \sum_{i=1}^N T_i(1 - \exp(-\sigma_i \delta_i)) \mathbf{c}_i,$$

where $T(t) = \exp\left(-\sum_{j=1}^{i-1} \sigma_j \delta_j\right),$

where $\delta_i = t_{i+1} - t_i$ is the distance between adjacent samples. This function for calculating $\hat{C}(\mathbf{r})$ from the set of (\mathbf{c}_i, σ_i) values is trivially differentiable.

We then use the volume rendering procedure to render the color of each ray from both sets of samples. Contrary to the baseline NeRF (Mildenhall et al., 2021), where two "coarse"

and "fine" models were simultaneously trained, we use only the "coarse" architecture.

3.1. HyperNeRFGAN

In this work, we propose a novel GAN architecture, HyperNeRFGAN, for generating 3D representations. The proposed approach utilizes INR-GAN, the implicit approach for generating samples. We postulate using the NeRF model as a target network compared to standard INR-GAN architecture, which uses the MLP model to create the output image. Thanks to that approach, the generator creates a specific 3D representation of the scene or object by delivering the specific NeRF parameters.

The architecture of our model is provided in Fig. 1. The generator \mathcal{G} takes the sample from the assumed base distribution (Gaussian) and returns the set of parameters Θ . These parameters are further used inside the NeRF model F_Θ to transform the spatial location $\mathbf{x} = (x, y, z)$ to emitted color $\mathbf{c} = (r, g, b)$ and volume density σ . Instead of standard linear architecture, F_Θ uses *factorized multiplicative modulation* (FMM) layers.

The FMM layer with input of size n_{in} and output of size n_{out} can be defined as:

$$y = W \odot (A \times B) \cdot x_{in} + b = \tilde{W} \cdot x_{in} + b,$$

where W and b are matrices that share the parameters across 3D representations, and A, B are two modulation matrices of shapes $n_{out} \times k, k \times n_{in}$ respectively, created by the generator. The parameter k controls the rank of $A \times B$. Higher values of k increase the expressiveness of the FMM layer but also increase the amount of memory required by the hypernetwork. We always use $k = 10$.

The INR model F_Θ is a simplified version of the baseline NeRF. To make training less computationally expensive, we do not optimize two networks as the original NeRF. We reject the bigger "fine" network and only employ the smaller "coarse" network. Additionally, we reduce the size of the "coarse" network by decreasing the number of channels in each hidden layer from 256 to 128. In some experiments, we also decrease the number of layers from 8 to 4.

We differ from the baseline NeRF in one more aspect, as we don't use the viewing direction. That's because the images used for training don't have view-dependent features like reflections. Even though the viewing direction is not used in our architecture, there is no reason why it couldn't be enabled for datasets that would benefit from it.

Our NeRF is a single MLP, which takes only the spatial location as input:

$$F_\Theta : \mathbf{x} \rightarrow (\mathbf{c}, \sigma),$$

In this work, we utilize the StyleGAN2 architecture, follow-

ing the design patterns of INR-GAN. The entire model is trained using the StyleGAN2 objective in a similar way as in INR-GAN. In each training iteration, the noise vector is sampled and transformed using generator \mathcal{G} to obtain the weights of the target NeRF model F_Θ . The target model is further used to render 2D images from various angles. The generated 2D images further serve as fake images for the discriminator \mathcal{D} . The role of the generator \mathcal{G} is to create the 3D representation that enables to render 2D images that will fool the discriminator. The discriminator aims to distinguish between fake renders and authentic 2D images from the data distribution.

4. Experiments

In this section, we first evaluate the quality of generating 3D objects by HyperNeRFGAN. We use a data set containing 2D images of 3D objects obtained from ShapeNet (Zimny et al., 2022). The data set contains 50 images of each element from the plane, chair, and car classes. It is the most suitable data set for our purpose since each object has a few images of each element. Then we use CARLA (Dosovitskiy et al., 2017), which contains images of cars. In such a case, we have only one image per object, but still, we have photos of objects from all sides. We can produce full 3D objects, which can be used in VR or augmented reality. In the end, we use classical CelebA (Liu et al., 2015) dataset, which contains faces. From a 3D generation point of view, it is challenging since we only have fronts of faces. In practice, 3D based generative model can be used to 3D-aware image synthesis (Chan et al., 2022).

4.1. 3D object generation from ShapeNet

In our first experiments, we use a ShapeNet base data set containing 50 images of each element from the plane, chair, and car classes. Such representation is perfect for training 3D models since each element has been seen from many views. The data was taken from (Zimny et al., 2022), where authors train an autoencoder-based generative model. In Fig. 3, we present objects generated from our model. In Fig. 4, we also present linear interpolation of objects. As we can see, objects are of very good quality, see Tab 1.

ShapeNet	cars	planes	chairs
Points2NeRF	82.1	239	129.3
HyperNeRFGAN	29.6	33.4	22.0

Table 1. Competition of HyperNeRFGAN and autoencoder based model by using FID. Competition between GAN and autoencoder and GAN is difficult. But we can obtain a better FID score.



Figure 5. Examples from a model trained on CARLA.



Figure 6. Meshes generated with a model trained on CARLA dataset and with models trained on planes and chairs from ShapeNet.

4.2. 3D object generation from CARLA data set

In the second experiment, we compare our model on CARLA dataset with other GAN-based models: HoloGAN (Nguyen-Phuoc et al., 2019), GRAF (Schwarz et al., 2020) and π -GAN (Chan et al., 2021). CARLA (Dosovitskiy et al., 2017) contains images of cars. We have only one image per object, but still, we have photos of objects from all sides. Consequently, full 3D objects can be used in VR or augmented reality. The visual comparison we present in Fig. 2. As illustrated, we can effectively model the transparency of glass in cars, see Fig. 5. In Tab. 2 we present a numerical comparison of the Frechet Inception Distance (FID), Kernel Inception Distance (KID), and In-

ception Score (IS). As can be seen, we obtain better results than the π -GAN model.

In the case of NeRF representation, we can produce meshes, see Fig. 6.

CARL	FID	KID	IS
HoloGAN	67.5	3.95	3.52
GRAF	41.7	2.43	3.60
π -GAN	29.2	1.36	4.27
HyperNeRFGAN	20.5	0.78	4.20

Table 2. FID, KID mean \times 100, and IS for CARLA dataset.

4.3. 3D-aware image synthesis from CelebA

In our third experiment, we further compare the same models as in the second experiment by changing the setup to face generation. For this task, we utilize the CelebA (Liu et al., 2015) dataset, which contains 200,000 high-resolution face images of 10,000 different celebrities. We crop the images from the top of the hair to the bottom of the chin and resize them to 128x128 resolution as π -GAN authors did. We present quantitative results in Tab. 3. As you can notice, HyperNeRFGAN and π -GAN achieve similar performance, which can also be seen in Fig. 7.

5. Conclusions

In this work, we presented a novel approach to generating NeRF representation from 2D images. Our model leverages

CelebA	FID	KID	IS
HoloGAN	39.7	2.91	1.89
GRAF	41.1	2.29	2.34
π -GAN	14.7	0.39	2.62
HyperNeRFGAN	15.04	0.66	2.63

Table 3. FID, KID mean \times 100, and IS for CelebA dataset.Figure 7. Comparison between HyperNeRFGAN (first 3 columns) and π -GAN uncurated generated faces

a hypernetwork paradigm and NeRF representation of the 3D scene. HyperNeRFGAN take a Gaussian noise and return the weights of a NeRF network that reconstructs 3D objects from 2D images. In training, we use only unlabeled images and a StyleGan2 discriminator. Such representation gives several advantages over the existing approaches. First of all, we can use NeRF instead of SIREN representation in the GAN type algorithm. Secondly, our model is simple and can be effectively trained on 3D objects. Finally, our model directly produces NeRF objects without sharing some global parameters of the rendering component.

Limitations The main limitation of HyperNeRFGAN is the fact that we use only 2D images instead any knowledge about 3D object representation. In future work, we plan to add some information about the structure of 3D meshes.

References

Achlioptas, P., Diamanti, O., Mitliagkas, I., and Guibas, L. Learning representations and generative models for 3d point clouds. In *International conference on machine learning*, pp. 40–49. PMLR, 2018.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein gen-

erative adversarial networks. In *International conference on machine learning*, pp. 214–223, 2017.

Arsalan Soltani, A., Huang, H., Wu, J., Kulkarni, T. D., and Tenenbaum, J. B. Synthesizing 3d shapes via modeling multi-view depth maps and silhouettes with deep generative networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1511–1519, 2017.

Cai, R., Yang, G., Averbuch-Elor, H., Hao, Z., Belongie, S., Snavely, N., and Hariharan, B. Learning gradient fields for shape generation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pp. 364–381. Springer, 2020.

Chan, E. R., Monteiro, M., Kellnhofer, P., Wu, J., and Wetzstein, G. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5799–5809, 2021.

Chan, E. R., Lin, C. Z., Chan, M. A., Nagano, K., Pan, B., De Mello, S., Gallo, O., Guibas, L. J., Tremblay, J., Khamis, S., et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16123–16133, 2022.

Chen, Z. and Zhang, H. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5939–5948, 2019.

Choy, C. B., Xu, D., Gwak, J., Chen, K., and Savarese, S. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *European conference on computer vision*, pp. 628–644. Springer, 2016.

Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., and Koltun, V. Carla: An open urban driving simulator. In *Conference on robot learning*, pp. 1–16. PMLR, 2017.

Dupont, E., Teh, Y. W., and Doucet, A. Generative models as distributions of functions. In *International Conference on Artificial Intelligence and Statistics*, pp. 2989–3015. PMLR, 2022.

Gadelha, M., Maji, S., and Wang, R. 3d shape induction from 2d views of multiple objects. In *2017 International Conference on 3D Vision (3DV)*, pp. 402–411. IEEE, 2017.

Girdhar, R., Fouhey, D. F., Rodriguez, M., and Gupta, A. Learning a predictable and generative vector representation for objects. In *European Conference on Computer Vision*, pp. 484–499. Springer, 2016.

- Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., and Bengio, Y. Generative adversarial nets. In *NIPS*, 2014.
- Grathwohl, W., Chen, R. T., Bettencourt, J., Sutskever, I., and Duvenaud, D. Ffjord: Free-form continuous dynamics for scalable reversible generative models. In *International Conference on Learning Representations*, 2018.
- Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of wasserstein gans. In *Advances in neural information processing systems*, pp. 5767–5777, 2017.
- Ha, D., Dai, A. M., and Le, Q. V. Hypernetworks. 2016.
- Häne, C., Tulsiani, S., and Malik, J. Hierarchical surface prediction for 3d object reconstruction. In *2017 International Conference on 3D Vision (3DV)*, pp. 412–420. IEEE, 2017.
- Henning, C., von Oswald, J., Sacramento, J., Surace, S. C., Pfister, J.-P., and Grewe, B. F. Approximating the predictive distribution via adversarially-trained hypernetworks. 2018.
- Huang, C.-W., Krueger, D., Lacoste, A., and Courville, A. Neural autoregressive flows. In *International Conference on Machine Learning*, pp. 2078–2087. PMLR, 2018.
- Kajiya, J. T. and Von Herzen, B. P. Ray tracing volume densities. *ACM SIGGRAPH computer graphics*, 18(3): 165–174, 1984.
- Karras, T., Aila, T., Laine, S., and Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.
- Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.
- Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8110–8119, 2020.
- Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. Improved variational inference with inverse autoregressive flow. *Advances in neural information processing systems*, 29, 2016.
- Kodali, N., Abernethy, J., Hays, J., and Kira, Z. On convergence and stability of gans. *arXiv preprint arXiv:1705.07215*, 2017.
- Li, J., Xu, K., Chaudhuri, S., Yumer, E., Zhang, H., and Guibas, L. Grass: Generative recursive autoencoders for shape structures. *ACM Transactions on Graphics (TOG)*, 36(4):1–14, 2017.
- Liao, Y., Schwarz, K., Mescheder, L., and Geiger, A. Towards unsupervised learning of generative models for 3d controllable image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5871–5880, 2020.
- Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.
- Liu, Z., Zhang, Y., Gao, J., and Wang, S. Vfmvac: View-filtering-based multi-view aggregating convolution for 3d shape recognition and retrieval. *Pattern Recognition*, 129: 108774, 2022. ISSN 0031-3203.
- Lunz, S., Li, Y., Fitzgibbon, A., and Kushman, N. Inverse graphics gan: Learning to generate 3d shapes from unstructured 2d data. *arXiv preprint arXiv:2002.12674*, 2020.
- Max, N. Optical models for direct volume rendering. *IEEE Transactions on Visualization and Computer Graphics*, 1 (2):99–108, 1995.
- Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., and Geiger, A. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4460–4470, 2019.
- Michalkiewicz, M., Pontes, J. K., Jack, D., Baktashmotlagh, M., and Eriksson, A. Implicit surface representations as layers in neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4743–4752, 2019.
- Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., and Ng, R. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.
- Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.
- Nguyen, P., Tran, T., Gupta, S., Rana, S., Dam, H.-C., and Venkatesh, S. Hypervae: A minimum description length variational hyper-encoding network. *arXiv preprint arXiv:2005.08482*, 2020.

- Nguyen-Phuoc, T., Li, C., Theis, L., Richardt, C., and Yang, Y.-L. Hologan: Unsupervised learning of 3d representations from natural images. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 2037–2040. IEEE, 2019.
- Nguyen-Phuoc, T. H., Richardt, C., Mai, L., Yang, Y., and Mitra, N. Blockgan: Learning 3d object-aware scene representations from unlabelled images. *Advances in Neural Information Processing Systems*, 33:6767–6778, 2020.
- Niemeyer, M. and Geiger, A. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11453–11464, 2021.
- Niemeyer, M., Mescheder, L., Oechsle, M., and Geiger, A. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3504–3515, 2020.
- Oechsle, M., Mescheder, L., Niemeyer, M., Strauss, T., and Geiger, A. Texture fields: Learning texture representations in function space. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4530–4539. IEEE Computer Society, 2019.
- Oord, A., Li, Y., Babuschkin, I., Simonyan, K., Vinyals, O., Kavukcuoglu, K., Driessche, G., Lockhart, E., Cobo, L., Stimberg, F., et al. Parallel wavenet: Fast high-fidelity speech synthesis. In *International conference on machine learning*, pp. 3918–3926. PMLR, 2018.
- Or-El, R., Luo, X., Shan, M., Shechtman, E., Park, J. J., and Kemelmacher-Shlizerman, I. Stylesdf: High-resolution 3d-consistent image and geometry generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13503–13513, 2022.
- Pan, X., Xu, X., Loy, C. C., Theobalt, C., and Dai, B. A shading-guided generative implicit model for shape-accurate 3d-aware image synthesis. *Advances in Neural Information Processing Systems*, 34:20002–20013, 2021.
- Park, J. J., Florence, P., Straub, J., Newcombe, R., and Lovegrove, S. Deep sdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 165–174, 2019.
- Peng, S., Niemeyer, M., Mescheder, L., Pollefeys, M., and Geiger, A. Convolutional occupancy networks. In *European Conference on Computer Vision*, pp. 523–540. Springer, 2020.
- Ratzlaff, N. and Fuxin, L. Hypergan: A generative model for diverse, performant neural networks. In *International Conference on Machine Learning*, pp. 5361–5369. PMLR, 2019.
- Rebain, D., Matthews, M. J., Yi, K. M., Sharma, G., Lagun, D., and Tagliasacchi, A. Attention beats concatenation for conditioning neural fields. *arXiv preprint arXiv:2209.10684*, 2022.
- Schwarz, K., Liao, Y., Niemeyer, M., and Geiger, A. Graf: Generative radiance fields for 3d-aware image synthesis. *Advances in Neural Information Processing Systems*, 33:20154–20166, 2020.
- Sendera, M., Przewiezlikowski, M., Karanowski, K., Zieba, M., Tabor, J., and Spurek, P. Hypershot: Few-shot learning by kernel hypernetworks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2469–2478, 2023.
- Sheikh, A.-S., Rasul, K., Merentitis, A., and Bergmann, U. Stochastic maximum likelihood optimization via hypernetworks. *arXiv preprint arXiv:1712.01141*, 2017.
- Shu, D. W., Park, S. W., and Kwon, J. Wasserstein distributional harvesting for highly dense 3d point clouds. *Pattern Recognition*, 132:108978, 2022.
- Sinha, A., Bai, J., and Ramani, K. Deep learning 3d shape surfaces using geometry images. In *European Conference on Computer Vision*, pp. 223–240. Springer, 2016.
- Sitzmann, V., Martel, J., Bergman, A., Lindell, D., and Wetzstein, G. Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems*, 33:7462–7473, 2020.
- Skorokhodov, I., Ignatyev, S., and Elhoseiny, M. Adversarial Generation of Continuous Images. *arXiv*, November 2020. doi: 10.48550/arXiv.2011.12026.
- Skorokhodov, I., Ignatyev, S., and Elhoseiny, M. Adversarial generation of continuous images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10753–10764, 2021.
- Spurek, P., Winczowski, S., Tabor, J., Zamorski, M., Zieba, M., and Trzcinski, T. Hypernetwork approach to generating point clouds. In *International Conference on Machine Learning*, pp. 9099–9108. PMLR, 2020.
- Spurek, P., Zieba, M., Tabor, J., and Trzcinski, T. General hypernetwork framework for creating 3d point clouds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9995–10008, 2022.

Struski, L., Knop, S., Spurek, P., Daniec, W., and Tabor, J. Locogan—locally convolutional gan. *Computer Vision and Image Understanding*, 221:103462, 2022.

Szabó, A., Meishvili, G., and Favaro, P. Unsupervised generative 3d shape learning from natural images. *arXiv preprint arXiv:1910.00287*, 2019.

Tewari, A., Elgharib, M., Bernard, F., Seidel, H.-P., Pérez, P., Zollhöfer, M., and Theobalt, C. Pie: Portrait image embedding for semantic control. *ACM Transactions on Graphics (TOG)*, 39(6):1–14, 2020.

Wu, J., Zhang, C., Xue, T., Freeman, B., and Tenenbaum, J. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. *Advances in neural information processing systems*, 29, 2016.

Xu, X., Pan, X., Lin, D., and Dai, B. Generative occupancy fields for 3d surface-aware image synthesis. *Advances in Neural Information Processing Systems*, 34:20683–20695, 2021.

Yang, F., Davoine, F., Wang, H., and Jin, Z. Continuous conditional random field convolution for point cloud segmentation. *Pattern Recognition*, 122:108357, 2022.

Yang, G., Huang, X., Hao, Z., Liu, M.-Y., Belongie, S., and Hariharan, B. Pointflow: 3d point cloud generation with continuous normalizing flows. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4541–4550, 2019.

Yu, Y., Gong, Z., Zhong, P., and Shan, J. Unsupervised representation learning with deep convolutional neural network for remote sensing images. In *International conference on image and graphics*, pp. 97–108. Springer, 2017.

Zhang, H., Goodfellow, I., Metaxas, D., and Odena, A. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018.

Zhu, J.-Y., Zhang, Z., Zhang, C., Wu, J., Torralba, A., Tenenbaum, J., and Freeman, B. Visual object networks: Image generation with disentangled 3d representations. *Advances in neural information processing systems*, 31, 2018.

Zieba, M., Przewiezkowski, M., Smieja, M., Tabor, J., Trzcinski, T., and Spurek, P. Regflow: Probabilistic flow-based regression for future prediction. *arXiv preprint arXiv:2011.14620*, 2020.

Zimny, D., Trzcinski, T., and Spurek, P. Points2nerf: Generating neural radiance fields from 3d point cloud. *arXiv preprint arXiv:2206.01290*, 2022.

A. Additional qualitative results of HyperNeRFGAN.

Figure 8. Linear interpolation between latent codes with model trained on CARLA.

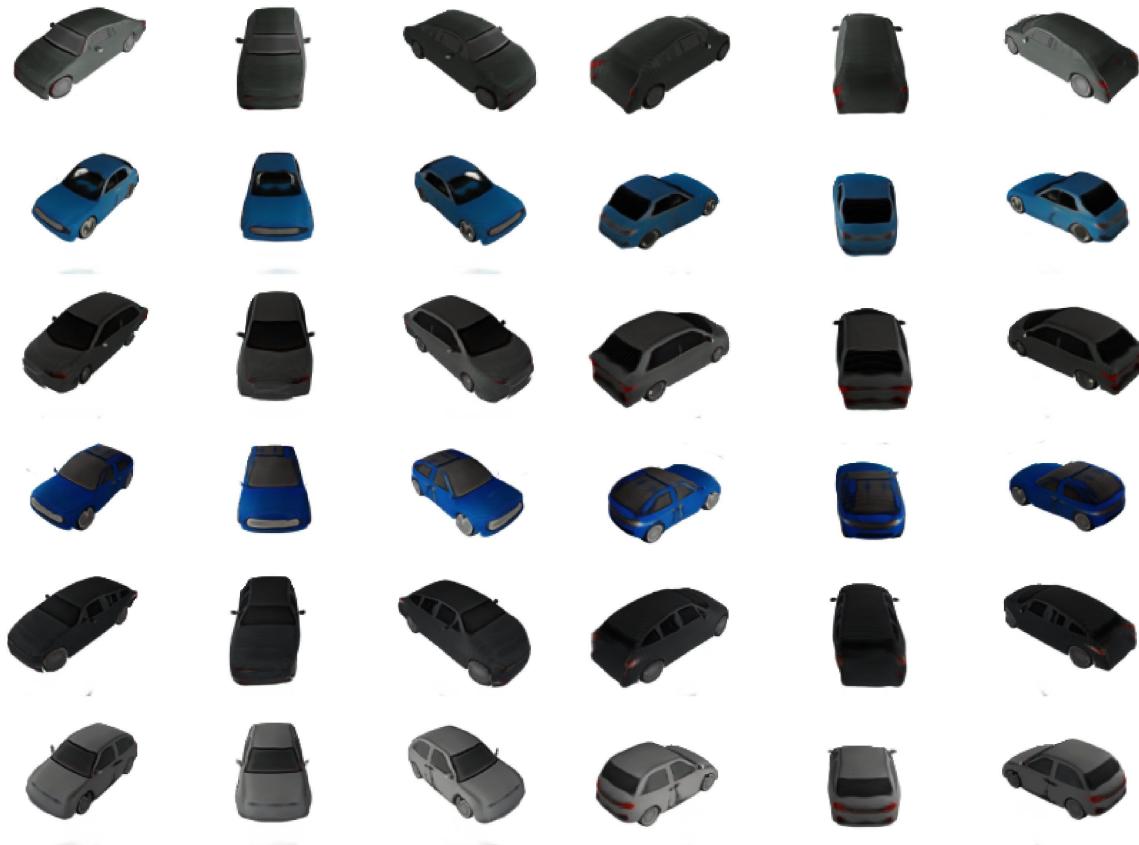


Figure 9. Elements generated by model trained on cars from ShapeNet.

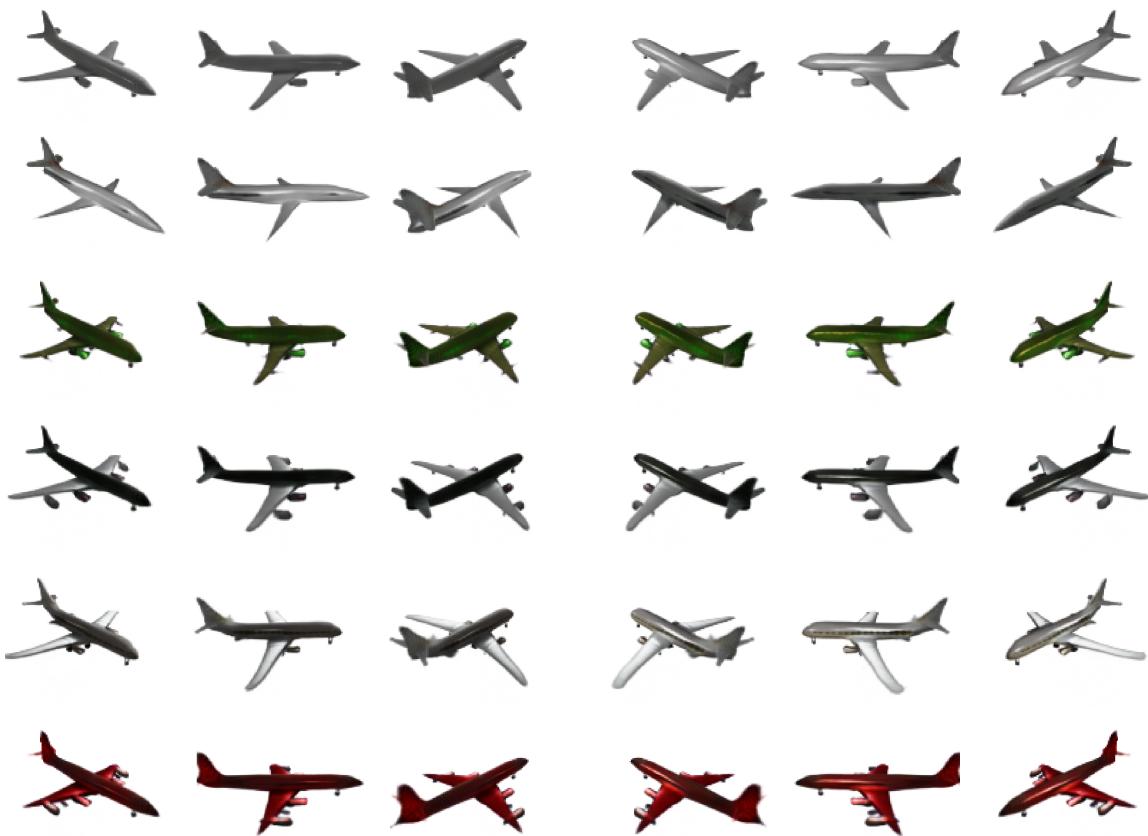


Figure 10. Elements generated by model trained on planes from ShapeNet.