HyperNeRFGAN: Hypernetwork approach to 3D NeRF GAN

Adam Kania [1] **Artur Kasymov** [1] **Jakub Kościukiewicz** [1] **Artur Górak** [1] **Marcin Mazur** [1] **Maciej Zieba** [2]
Przemysław Spurek [1]

Abstract

The recent surge in popularity of deep generative models for 3D objects has highlighted the need for more efficient training methods, particularly given the difficulties associated with training with conventional 3D representations, such as voxels or point clouds. Neural Radiance Fields (NeRFs), which provide the current benchmark in terms of quality for the generation of novel views of complex 3D scenes from a limited set of 2D images, represent a promising solution to this challenge. However, the training of these models requires the knowledge of the respective camera positions from which the images were viewed. In this paper, we overcome this limitation by introducing HyperNeRFGAN, a Generative Adversarial Network (GAN) architecture employing a hypernetwork paradigm to transform a Gaussian noise into the weights of a NeRF architecture that does not utilize viewing directions in its training phase. Consequently, as evidenced by the findings of our experimental study, the proposed model, despite its notable simplicity in comparison to existing state-of-the-art alternatives, demonstrates superior performance on a diverse range of image datasets where camera position estimation is challenging, particularly in the context of medical data.

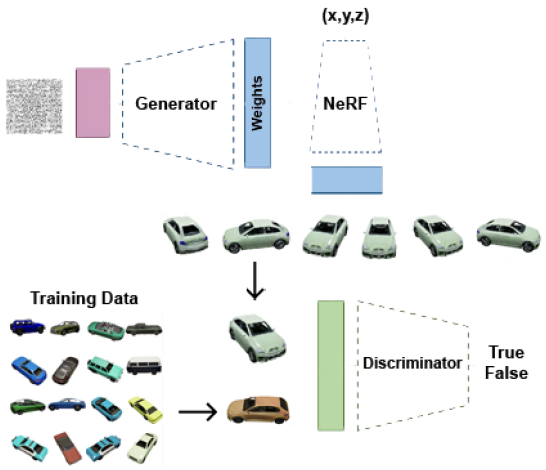

Figure 1. Our HyperNeRFGAN model employs a hypernetwork to convert a Gaussian noise into the weights of the simplified NeRF architecture (not requiring information about camera positions), subsequently used for the generation of novel 2D views. During the training phase, a standard GAN-based framework (incorporating a typical 2D discriminator) is employed. Despite the generation of 2D images, our model utilizes a 3D-aware NeRF representation, thereby facilitating precise 3D object generation.

1. Introduction

Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) facilitate the generation of high-quality 2D images (Yu et al., 2017; Karras et al., 2017; 2019; 2020; Struski et al., 2022). However, achieving a comparable level of quality for 3D objects remains a challenge. The primary difficulty arises from the necessity of extensive deep architectures in conjunction with the use of 3D representations, including voxels and point clouds, which present challenges in an accurate color rendering. One potential solution to this issue is the direct working within the 2D image domain, as exemplified by Neural Radiance Fields (NeRFs) (Mildenhall et al., 2021), which permit the generation of unseen views of intricate 3D scenes based on a limited set of 2D images. In essence, by leveraging the relationships between these base images and computer graphics techniques such as ray tracing, these neural models are capable of producing high-quality renderings of 3D objects from novel viewpoints.

It must be acknowledged that the incorporation of NeRF representation with a GAN architecture is not a straightforward process due to the inherent complexity of the NeRF conditioning mechanism (Rebain et al., 2022). Consequently, the majority of models tend to favour the use of the SIREN (Sitz-

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

Figure 2. Qualitative comparison of HyperNeRFGAN (our) with HoloGAN (Nguyen-Phuoc et al., 2019), GRAF (Schwarz et al., 2020), and π-GAN (Chan et al., 2021) trained on the CARLA dataset (Dosovitskiy et al., 2017). It is noteworthy that our model has been shown to produce outcomes that are comparable to those of the most successful competitor, namely π-GAN.

mann et al., 2020) architecture instead, as this allows for natural conditioning. However, the quality of rendered 3D objects is somewhat inferior when NeRF is replaced with SIREN, as exemplified by models such as GRAF (Schwarz et al., 2020) and π-GAN (Chan et al., 2021) which employ SIREN along with a conditioning mechanism to generate implicit representations. While these approaches yield promising results, the substitution of SIREN with NeRF in such frameworks presents an intriguing yet complex issue.

In order to address the aforementioned challenge, this paper incorporates NeRF and GAN architectures by means of the hypernetwork paradigm (Ha et al., 2016). Specifically, we introduce the HyperNeRFGAN model, which employs a hypernetwork designed to convert a Gaussian noise into the weights of a simplified NeRF target architecture that no longer utilizes viewing directions to produce the output for given 3D positions. Subsequently, the novel 2D views rendered by NeRF are evaluated by a traditional 2D discriminator. Consequently, the entire model is implicitly trained in accordance with the principles of a GAN-based framework. A diagrammatic representation of the proposed architectural design is presented in Figure 1. It should be noted that although HyperNeRFGAN generates 2D images, it employs a 3D-aware NeRF representation, thereby ensuring that the model produces accurate 3D objects. Furthermore, as the model does not employ viewing directions during training, it can be successfully applied to a variety of datasets where camera position estimation may be challenging or impossible. To illustrate this, we apply HyperNeRFGAN to three different datasets: CARLA (Dosovitskiy et al., 2017) and CelebA (Liu et al., 2015), which are both real-world datasets, as well as a medical dataset consisting of digitally reconstructed radiographs (DRR) of chests and knees (Corona-Figueroa et al., 2022). The results of the experiments conducted demonstrate that our solution, despite its notable simplicity, outperforms (or is at least on a similar level to) existing state-of-the-art methods, including HoloGAN (Nguyen-Phuoc et al., 2019), GRAF (Schwarz et al., 2020), and π-GAN (Chan et al., 2021) for CARLA and CelebA, and GRAF (Schwarz et al., 2020), pixelNeRF (Yu et al., 2021), MedNeRF (Corona-Figueroa et al., 2022), and UMedNeRF (Hu et al., 2023) for the medical dataset.

In conclusion, our contribution can be summarized as follows:

- we introduce HyperNeRFGAN, an implicit GAN-based model that generates a simplified 3D-aware NeRF representation (not requiring information about camera positions) directly through the use of a hypernetwork,
- the derived NeRF architecture is employed for the generation of novel 2D views, subsequently evaluated through a conventional 2D discriminator (thus, the complete model is trained in accordance with the tenets of a 2D adversarial methodology),
- the experiments conducted demonstrate that HyperNeRFGAN, despite its notable simplicity in comparison to existing state-of-the-art alternatives, exhibits superior performance on a diverse range of image datasets where camera position estimation is challenging, particularly in the context of medical data.

Figure 3. Sample 2D images generated by the HyperNeRFGAN model (our) trained on the ShapeNet-based dataset proposed in (Zimny et al., 2022), consisting of 50 images of each object from the car, plane, and chair classes.

2. Related work

Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) permit 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 dimensional images and thus fail to account for the 3D dimensional nature of the physical world.

The initial approaches for 3D-aware image synthesis were Visual Object Network (Zhu et al., 2018) and PrGAN (Gadelha et al., 2017). These methods generate voxelized 3D shapes using a 3D-GAN (Wu et al., 2016) and then project them into 2D. An alternative approach is taken by HoloGAN (Nguyen-Phuoc et al., 2019) and BlockGAN (Nguyen-Phuoc et al., 2020), which employ a similar fusion but utilize implicit 3D representation to model a 3D representation of the world. The use of an explicit volume representation has, however, constrained the resolution of the models, as discussed in (Lunz et al., 2020). Furthermore, the authors of (Szabó et al., 2019) suggest the use of meshes to represent 3D geometry. Conversely, in (Liao et al., 2020), collections of primitives for image synthesis are employed.

GRAF (Schwarz et al., 2020) and π-GAN (Chan et al., 2021) employ implicit neural radiance fields for the generation of 3D-aware images and geometry. These models also utilize the SIREN algorithm in conjunction with a conditioning mechanism. As an alternative option, ShadeGAN (Pan et al., 2021) employs a shading-guided pipeline, whereas GOF (Xu et al., 2021) utilizes a gradual reduction in the sampling region of each camera ray. In the GIRAFFE approach (Niemeyer & Geiger, 2021), the initial step is to generate low-resolution feature maps. In the second step, the representation is passed to a 2D convolutional neural network (CNN) to generate outputs at a higher resolution. In contrast, StyleSDF (Or-El et al., 2022) integrates an SDF-based 3D representation with a StyleGAN2 for image generation, whereas in (Chan et al., 2022), the authors employ a StyleGAN2 generator and a tri-plane representation of 3D objects. These approaches yield superior results in terms of generated object quality but are exceedingly challenging to train.

A distinct family of models has been developed for use in medical applications. One such model is MedNeRF (Corona-Figueroa et al., 2022), which employs a GRAF-based approach to generate precise 3D projections from a single X-ray view. By combining GRAF with DAG (Tran et al., 2021) and employing multiple discriminator heads, the authors of (Corona-Figueroa et al., 2022) achieved markedly superior results compared to standard GRAF on the medical dataset. Another related model, UMedNeRF (Hu et al., 2023), is constructed upon MedNeRF. It employs automated calculation of weight parameters for discriminator losses, thereby facilitating more precise adaptation to the specific requirements of each task. This results in enhanced image clarity and the discernment of finer details within bone structures, when compared to its underlying MedNeRF architecture.

3. HyperNeRFGAN: hypernetwork for generating NeRF represention

This section presents the HyperNeRFGAN model. The fundamental premise of the proposed GAN-based methodology is that the generator serves as a hypernetwork, transforming a noise vector, sampled from a Gaussian distribution, into the weights of a NeRF representation of a given 3D object. Consequently, it is possible to generate a multitude of images of the object from a variety of perspectives in a manner that is fully controllable. Furthermore, the use of NeRF-based image rendering allows the discriminator to operate on generated 2D images, which is a notable simplification compared to existing state-of-the-art GAN-based models, such as HoloGAN (Nguyen-Phuoc et al., 2019),

Figure 4. Sample 2D images generated by the HyperNeRFGAN model (our) trained on the CARLA dataset (Dosovitskiy et al., 2017). It should be noted that our method permits the effective modeling of transparency in car windows.

GRAF (Schwarz et al., 2020), and π-GAN (Chan et al., 2021), which are fed by complex 3D structures. We begin by outlining the fundamental concepts that underpin our approach, after which we proceed to present the architectural and training details of the HyperNeRFGAN model.

Hypernetworks As defined in (Ha et al., 2016), hypernetworks are neural models that generate weights for another target network with the objective of solving a specific task. This approach results in a reduction of the number of trainable parameters in comparison to traditional methodologies that integrate supplementary information into the target model via a single embedding. A notable reduction in the size of the target model is achievable due to the fact that it does not share global weights. Instead, these weights are provided by the hypernetwork. Similarly, the authors of (Sheikh et al., 2017) employ this mechanism to generate a variety of target networks that approximate the same function, thereby establishing a parallel between hypernetworks and generative models.

Hypernetworks have a multitude of applications, including few-shot learning (Sendera et al., 2023) and probabilistic regression (Zieba et al., 2020). Additionally, they are utilized in numerous techniques to generate continuous 3D object representations (Spurek et al., 2020; 2022). For instance, HyperCloud (Spurek et al., 2020) employs a classical MLP as the target model to transform points from a uniform distribution on the unit sphere into point clouds that conform to the desired shape. Conversely, in (Spurek et al., 2022), the target model is a Continuous Normalizing Flow (Grathwohl et al., 2018), a generative model that constructs the point cloud from an assumed base distribution in 3D space.

NeRF representation of 3D objects A Neural Radiance Field (NeRF) is a scene modeling technique that uses a fully-connected network to represent the visual data (Mildenhall et al., 2021). The input to NeRF is a 5D coordinate, comprising a spatial position $\mathbf{x} = (x, y, z)$ and a viewing direction $\mathbf{d} = (\theta, \psi)$. The output is an emitted color $\mathbf{c} = (r, g, b)$ and a volume density σ.

A standard NeRF model is trained using a collection of images. In this context, numerous rays are generated that intersect both the image and a 3D object, which is modeled by a multi-layer perceptron (MLP) network

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

with parameters θ adjusted to map each 5D input coordinate to its respective directional color emission and volume density.

The NeRF loss function draws inspiration from traditional volume rendering as described in (Kajiya & Von Herzen, 1984). The color is calculated for each ray traversing the scene. The volume density $\sigma = \sigma(\mathbf{x})$ can be regarded as the differential probability of a ray. The anticipated color $\mathbf{c} = \mathbf{c}(\mathbf{x}, \sigma)$ is interpreted as the color of a corresponding camera ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ (where $\mathbf{o}$ is the ray's origin and $\mathbf{d}$ is its direction). It is determined by an integral and numerically approximated by a sum in practical applications.

Hypernetwork paradigm for generative modeling The concept of integrating hypernetworks and deep generative models is a relatively well-established area of research. In (Ratzlaff & Fuxin, 2019; Henning et al., 2018), the authors construct a GAN to generate the parameters of a neural network dedicated to regression or classification tasks. Other related examples include HyperVAE (Nguyen et al., 2020), which is designed to encode an arbitrary target distribution by generating parameters of a deep generative model given distribution samples, Point2NeRF (Zimny et al., 2022), which uses hypernetwork to output weights of an autoencoder based generative model (VAE) to create NeRF representation from a 3D point cloud, and HCNAF (Oechsle et al., 2019), which is a hypernetwork that produces parameters for a conditional auto-regressive flow model (Kingma et al., 2016; Oord et al., 2018; Huang et al., 2018). Alternatively, the authors of (Skorokhodov et al., 2021) propose INR-GAN, which employs a hypernetwork to generate a continuous representation of images. The hypernetwork is capable of modifying the shared weights through a low-cost mechanism known as factorized multiplicative modulation.

HyperNeRFGAN The proposed HyperNeRFGAN utilizes a hypenetwork to produce weights for the NeRF target network. The model adheres to the design patterns of INR-GAN, employing the StyleGAN2 backbone architecture. Consequently, it is trained using the StyleGAN2 objective in a manner analogous to that employed in INR-GAN. Specifically, in each training iteration, the noise vector from the assumed base (Gaussian) distribution P_{noise} is sampled and transformed using the generator $\mathcal{G}$ to obtain the weights of the target NeRF model F_θ. Furthermore, the target model is employed to render 2D images from a variety of perspectives. The generator $\mathcal{G}$ is responsible for creating a 3D representation that enables the generation of 2D images that will be indistinguishable from those created by the true data distribution P_{data}. In contrast, the discriminator $\mathcal{D}$ is designed to distinguish between fake renderings and authentic 2D images drawn from the data distribution. Formally, this minimax game is given by the following expression:

$$\min_{\mathcal{G}} \max_{\mathcal{D}} V(\mathcal{D}, \mathcal{G}), \tag{1}$$

where

$$V = \mathbb{E}_{x \sim P_{data}} \log \mathcal{D}(x) + \mathbb{E}_{z \sim P_{noise}} \log(1 - \mathcal{D}(\mathcal{G}(z))). \tag{2}$$

A diagrammatic summary of the architectural structure of HyperNeRFGAN is presented in Figure 1.

The key distinction between our proposed model and existing state-of-the-art solutions in the field is the use of a special 3D-aware NeRF representation F_θ (instead of SIREN) that differs from the original NeRF in a few aspects. Firstly, in contrast to the standard linear architecture, F_θ employs factorized multiplicative modulation (FMM) layers, as seen in INR-GAN. The FMM layer with an input of size n_{in} and an output of size n_{out} can be defined as follows:

$$y = W \odot (A \times B) \cdot x_{in} + b = \tilde{W} \cdot x_{in} + b, \tag{3}$$

where W and b are matrices that share the parameters across 3D representations, while A and B are two modulation matrices (created by the generator $\mathcal{G}$) with dimensions $n_{out} \times k$ and $k \times n_{in}$, respectively. The parameter k controls the rank of $A \times B$ and exerts an influence on the expressiveness and memory usage. In this regard, higher values of k result in an increase in the expressiveness of the FMM layer, but also in an increase in the amount of memory required by the hypernetwork. In our approach, we always use $k = 10$.

Secondly, to reduce the computational expense of training, we do not optimize two networks as in the original NeRF. Instead, we reject the larger "fine" network and employ only 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.

Finally, in contrast to the standard NeRF architecture, our approach does not utilize the viewing direction. Instead, our NeRF representation is a single MLP that takes the spatial location $\mathbf{x} = (x, y, z)$ and transforms it to the emitted color $\mathbf{c} = \mathbf{c}(\mathbf{x})$ and volume density $\sigma = \sigma(\mathbf{x})$, i.e.:

$$F_\theta \colon \mathbf{x} \to (\mathbf{c}, \sigma). \tag{4}$$

This is due to the fact that the images utilized for training lack view-dependent characteristics, such as reflections. (However, while our solution does not currently employ viewing direction data, there is no inherent reason to prohibit its use in datasets that would benefit from this additional information.) It should be noted that the use of such an architectural approach allows for the training of our model on a variety of data types that do not provide access to the camera position associated with images or to a substantial number of images for a single object. (In fact, the use of a single unlabeled view per object is sufficient for training.) This allows for the training of our model on medical data, as demonstrated in the following section.

4. Experiments

In this section, we conduct an empirical evaluation of our HyperNeRFGAN method in comparison to the state-of-the-art solutions. Firstly, we undertake a comparative analysis of the quality of generated 3D objects produced by a range of models trained on two distinct datasets. The first is a dataset comprising 2D images of 3D objects obtained from ShapeNet (Zimny et al., 2022), and the second is the

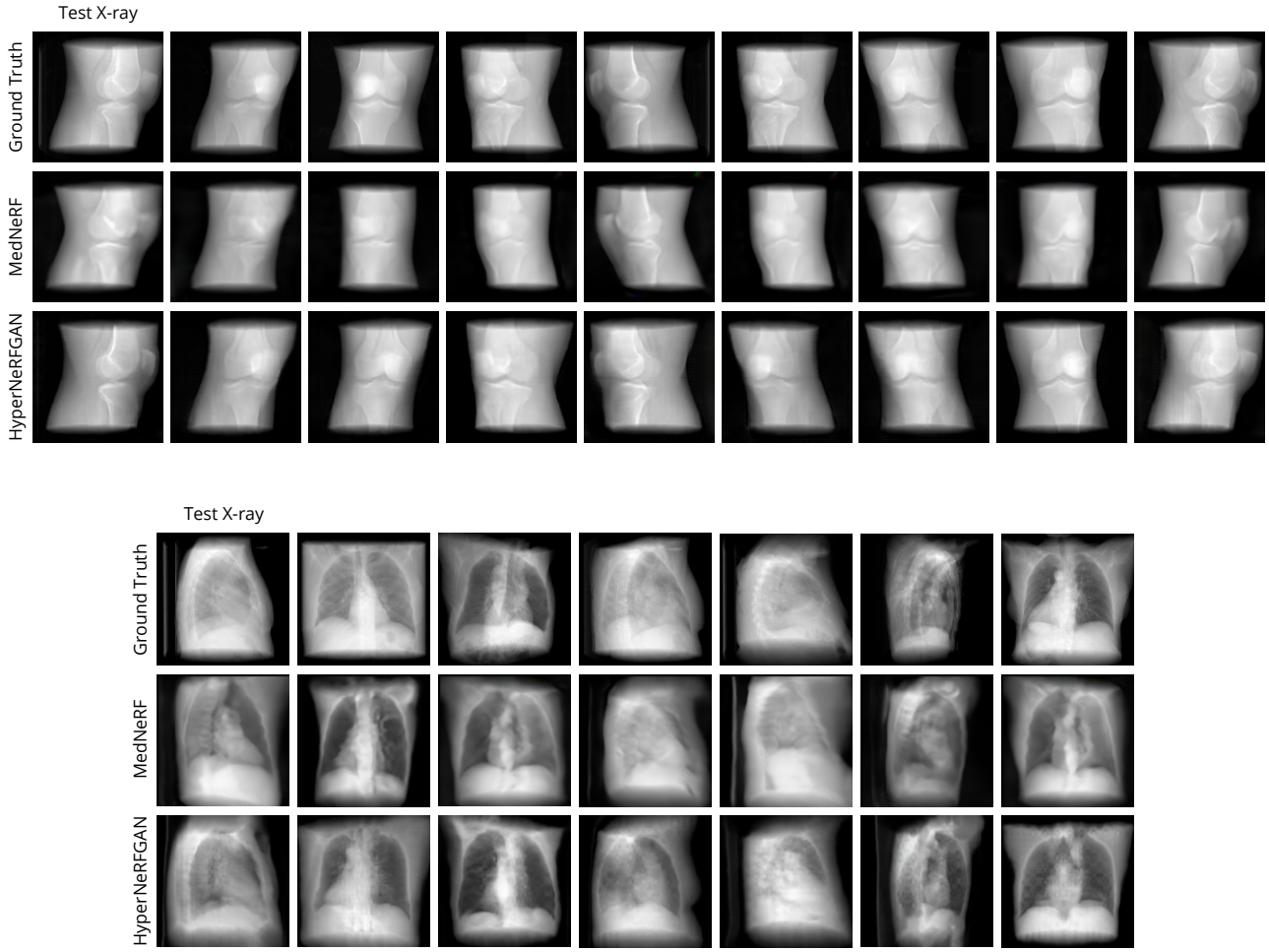

Figure 5. Qualitative comparison between HyperNeRFGAN (our) and MedNeRF trained on the medical dataset consisting of digitally reconstructed radiographs (DRR) of knees and chests (Corona-Figueroa et al., 2022). Note that our qualitative comparison shows that our model shows a significant improvement in the quality of reconstruction of CT projections.

	Chest dataset		Knee dataset	
	$\downarrow$ FID ($\mu \pm \sigma$)	$\downarrow$ KID ($\mu \pm \sigma$)	$\downarrow$ FID ($\mu \pm \sigma$)	$\downarrow$ KID ($\mu \pm \sigma$)
GRAF	68.25 ± 0.954	0.053 ± 0.0008	76.70 ± 0.302	0.058 ± 0.0001
pixelNeRF	112.96 ± 2.356	0.084 ± 0.0012	166.40 ± 2.153	0.158 ± 0.0010
MedNeRF	60.26 ± 0.322	$\mathbf{0.041 \pm 0.0005}$	76.12 ± 0.193	0.052 ± 0.0004
UMedNeRF	60.25 ± 0.642	0.043 ± 0.0011	70.73 ± 1.665	$\mathbf{0.041 \pm 0.0012}$
HyperNeRFGAN	$\mathbf{53.53 \pm 0.917}$	0.043 ± 0.0015	$\mathbf{57.65 \pm 0.689}$	0.044 ± 0.0007

Table 1. Quantitative evaluation of HyperNeRFGAN (our) in comparison with GRAF (Schwarz et al., 2020), and pixelNeRF(Yu et al., 2021), MedNeRF (Corona-Figueroa et al., 2022), and UMedNeRF (Hu et al., 2023), in terms of the FID (lower is better) and KID (lower is better) metrics. trained on the medical dataset consisting of digitally reconstructed radiographs (DRR) of chests and knees (Corona-Figueroa et al., 2022). All scores are the average of 3 runs. The obtained results prove the superior performance of our solution with respect to the baseline methods.

Figure 6. Sample meshes generated by the HyperNeRFGAN model (our) trained on the CARLA dataset (Dosovitskiy et al., 2017) and two classes (car and plane) from the ShapeNet dataset (Zimny et al., 2022).

CARLA dataset (Dosovitskiy et al., 2017), which includes images of cars. It should be noted that these datasets are particularly well-suited to our purposes, as each object is presented from only a few perspectives. Secondly, the classical CelebA dataset (Liu et al., 2015), which contains photographs of faces, is employed to assess the performance of different state-of-the-art generative models designed for 3D-aware image synthesis (Chan et al., 2022). It should be noted that from the perspective of 3D generation, this task presents a significant challenge, given that the only available source data are the photographed front sides of faces. Finally, the effectiveness of our model is evaluated using a dataset comprising digitally reconstructed radiographs (DRR) of chests and knees (Corona-Figueroa et al., 2022), in order to ensure its comparability with existing methods.

3D object generation In the initial experiment, a ShapeNet-based dataset comprising 50 images of each object from the plane, chair, and car categories was utilized. The data were obtained from (Zimny et al., 2022), where the authors propose the Point2NeRF model, which thus serves as a natural baseline for our method. Figure 3 presents the 3D objects produced by HyperNeRFGAN, while Figure 7 additionally displays the results of linear interpolation. It is evident that our approach produces high-quality renders. This is corroborated by the findings of our quantitative study, detailed in Table 2.

In the second experiment, we evaluate the performance of our model on the CARLA dataset (Dosovitskiy et al., 2017) in comparison to other GAN-based models, namely HoloGAN (Nguyen-Phuoc et al., 2019), GRAF (Schwarz et al., 2020) and π-GAN (Chan et al., 2021). It should be noted that CARLA comprises only a single image per object (a car), but that we have nonetheless access to photographs of the objects captured from a range of perspectives. The qualitative comparison is presented in Figure 2, while Table 3 delivers the results of the quantitative comparison in terms of the Fréchet Inception Distance (FID) (Heusel et al., 2017), Kernel Inception Distance (KID) (Bińkowski et al., 2018), and Inception Score (IS) (Salimans et al., 2016). It is evident that HyperNeRFGAN outperforms all the competitors. Furthermore, as illustrated in Figure 4, our method allows for the effective modeling of transparency in car windows.

ShapeNet	Car	Plane	Chair
Points2NeRF	82.1	239	129.3
HyperNeRFGAN (our)	**29.6**	**33.4**	**22.0**

Table 2. Quantitative evaluation of HyperNeRFGAN (our) in comparison with the autoencoder-based Point2NeRF model (Zimny et al., 2022) in terms of the FID metric (lower is better). The models were trained on three datasets consisting of 50 images from the car, plane, and chair classes of ShapeNet. The obtained results clearly demonstrate the superiority of our proposed solution.

CARLA	↓ FID	↓ KID×100	↑ IS
HoloGAN	67.5	3.95	3.52
GRAF	41.7	2.43	3.60
$\pi-$GAN	29.2	1.36	**4.27**
HyperNeRFGAN (our)	**20.5**	**0.78**	4.20

Table 3. Quantitative evaluation of HyperNeRFGAN (our) in comparison with HoloGAN (Nguyen-Phuoc et al., 2019), GRAF (Schwarz et al., 2020), and π-GAN (Chan et al., 2021), in terms of the FID (lower is better), KID (lower is better), and IS (greater is better) metrics. The models were trained on the CARLA dataset (Dosovitskiy et al., 2017). The obtained results demonstrate that our proposed solution is superior (or at least comparable) to the baseline methods.

It should be noted that the HyperNeRFGAN model, due to its ability to represent NeRF implicitly, can produce high-quality meshes of 3D objects. This is demonstrated in Figure 6, which presents meshes generated by the CARLA-trained model and two classes of ShapeNet.

3D-aware image synthesis In the third experiment, the same models are further compared by modifying the setup to focus on face generation. For this objective, the CelebA dataset (Liu et al., 2015) is employed, comprising 200,000 high-resolution images of 10,000 different celebrities. The images are cropped from the top of the hair to the bottom of the chin and resized to 128×128 resolution, as in π-GAN. The quantitative results are presented in Table 4. It is evident that our model and π-GAN achieve similar performance, which can also be observed in Figure 8.

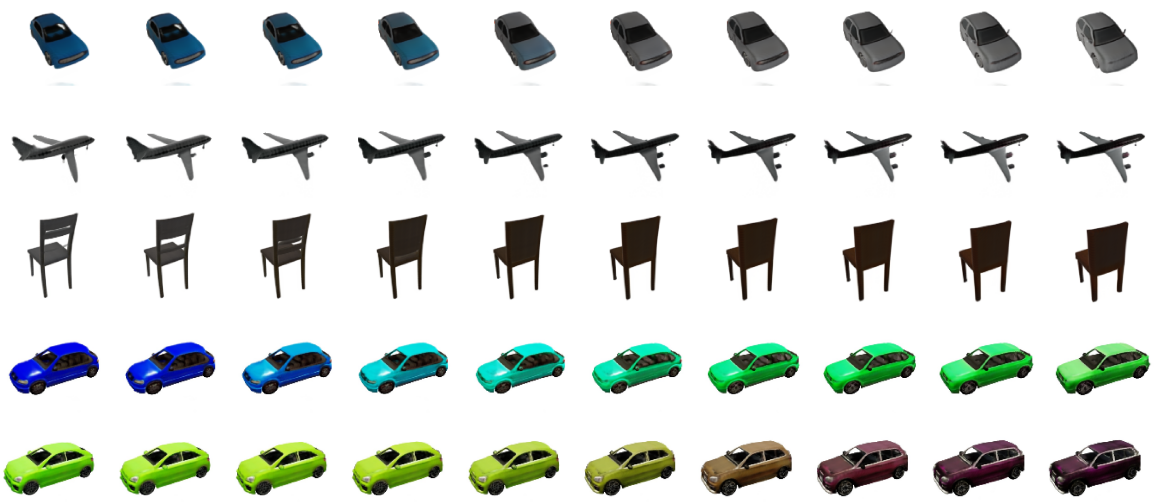

Figure 7. Linear interpolation examples generated by the HyperNeRFGAN model (our) trained on the ShapeNet-based dataset proposed in (Zimny et al., 2022) (first three rows) and the CARLA dataset (Dosovitskiy et al., 2017) (last two rows).

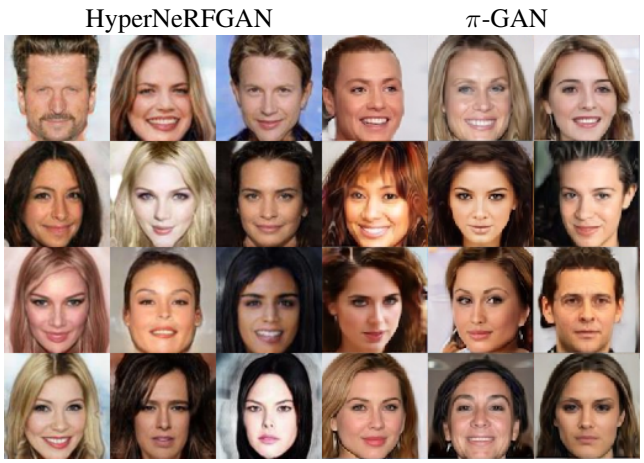

Figure 8. Qualitative comparison between HyperNeRFGAN (our) and π-GAN trained on the CelebA dataset (Liu et al., 2015). It can be seen that both methods demonstrate comparable performance.

CelebA	↓ FID	↓ KID×100	↑ IS
HoloGAN	39.7	2.91	1.89
GRAF	41.1	2.29	2.34
π-GAN	**14.7**	**0.39**	2.62
HyperNeRFGAN (our)	15.04	0.66	**2.63**

Table 4. Quantitative evaluation of HyperNeRFGAN (our) in comparison with HoloGAN (Nguyen-Phuoc et al., 2019), GRAF (Schwarz et al., 2020), and π-GAN (Chan et al., 2021), in terms of the FID (lower is better), KID (lower is better), and IS (greater is better) metrics. The models were trained on the CelebA dataset (Liu et al., 2015). The obtained results show that our proposed solution achieves similar performance to π-GAN (the best of the competitors).

Medical applications In order to evaluate our model for medical applications and ensure its comparability with existing methods, we utilize a dataset employed by the authors of (Corona-Figueroa et al., 2022), which contains digitally reconstructed radiographs (DRRs). The dataset comprises 20 examples of the chest and 5 examples of the knee, with each example consisting of 72 128×128 images captured at 5-degree intervals, encompassing a full 360-degree vertical rotation for each patient. To account for differences between the synthetic and medical datasets, the sampling angle is modified to encompass a single axis, as in MedNeRF (Corona-Figueroa et al., 2022), and the model configuration is altered to exclude the assumption of a white background for the training data. The model was trained on a single example for each experiment. As illustrated

in Figure 5, our qualitative comparison demonstrates that HyperNeRFGAN exhibits a notable enhancement in the quality of reconstruction of CT projections compared to MedNeRF (Corona-Figueroa et al., 2022). This observation is further substantiated by the quantitative results presented in Table 1.

5. Conclusions

In this paper, we present HyperNeRFGAN, a novel generative adversarial network (GAN)-based approach to generating 3D-aware representations from 2D images. Our model employs a hypernetwork paradigm and a simplified NeRF representation of a 3D scene. In contrast to the conventional NeRF architecture, HyperNeRFGAN does not utilize viewing directions during training. This enables its successful deployment in diverse datasets where camera position estimation may be challenging or impossible, particularly in the context of medical data. The outcomes of the conducted experiments illustrate that our solution outperforms (or is at least on a comparable level to) existing state-of-the-art methods.

References

Bińkowski, M., Sutherland, D. J., Arbel, M., and Gretton, A. Demystifying mmd gans. In *International Conference on Learning Representations*, 2018.

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.

Corona-Figueroa, A., Frawley, J., Bond-Taylor, S., Bethapudi, S., Shum, H. P., and Willcocks, C. G. Mednerf: Medical neural radiance fields for reconstructing 3d-aware ct-projections from a single x-ray. In *2022 44th annual international conference of the IEEE engineering in medicine Biology society (EMBC)*, pp. 3843–3848. IEEE, 2022.

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.

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.

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.

Ha, D., Dai, A., and Le, Q. V. Hypernetworks. *arXiv preprint arXiv:1609.09106*, 2016.

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. *arXiv preprint arXiv:2005.08482*, 2018.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local Nash equilibrium. In *Adv. in Neural Information Processing Systems*, NeurIPS, pp. 6626–6637, 2017.

Hu, J., Fan, Q., Hu, S., Lyu, S., Wu, X., and Wang, X. Umednerf: Uncertainty-aware single view volumetric rendering for medical neural radiance fields. *arXiv e-prints*, pp. arXiv–2311, 2023.

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.

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.

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.

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.

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.

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.

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.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

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.

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. 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.

Tran, N.-T., Tran, V.-H., Nguyen, N.-B., Nguyen, T.-K., and Cheung, N.-M. On data augmentation for gan training. *IEEE Transactions on Image Processing*, 30: 1882–1897, 2021. ISSN 1941-0042. doi: 10.1109/tip.2021.3049346. URL http://dx.doi.org/10.1109/TIP.2021.3049346.

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.

Yu, A., Ye, V., Tancik, M., and Kanazawa, A. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4578–4587, 2021.

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.

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., Przewiezlikowski, 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 9. Linear interpolation between latent codes with model trained on CARLA.

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

Figure 11. Elements generated by model trained on planes from ShapeNet.