Beyond Depth: Attention-Driven GANs for Hyper-Realistic Indoor Scene Generation

Rajat Gade
rgade@umass.edu

Priyanka Gupta
pgupta@umass.edu

Meghana Maddipatla
meghanamaddi@umass.edu

Abstract

Recent advances in indoor scene synthesis have been driven by deep learning models leveraging depth information to enhance realism and spatial consistency. Existing methods often struggle to accurately capture depth and spatial relationships from 2D images alone, especially within diverse indoor layouts. To address this challenge, we develop a Generative Adversarial Network (GAN) architecture featuring a modified generator, building upon the principles of DepthGAN. Our novel approach explores architectural enhancements, including an integrated attention mechanism within the generator, enabling targeted focus on specific parts of the latent vector or depth map during image generation. This attention mechanism enhances detail sharpness, spatial coherence, and scene layout comprehension. The proposed model inputs a low-dimensional latent vector representing essential scene properties and outputs realistic RGB images with accurate spatial object relationships. Additionally, the model can generate corresponding depth maps for further analysis and manipulation. Through this work, we aim to experiment models for indoor scene synthesis, offering improved realism and spatial fidelity in generated scenes

1. Introduction

This project aims to develop a deep-learning model for generating realistic and spatially consistent indoor scenes. We take inspiration from the recent work "3D-Aware Indoor Scene Synthesis with Depth Priors" by Qifeng Chen et al. (2022), which introduced DepthGAN, a dual-path generative network architecture that leverages depth information for improved scene synthesis. However, we propose to push the boundaries of DepthGAN's performance by exploring novel architectural modifications and training strategies. Current methods for indoor scene synthesis from 2D images often struggle to capture the true depth and spatial relationships between objects. This limitation stems from the inherent challenge of inferring 3D information from purely 2D data, especially considering the diverse layouts and object arrangements found in indoor environments. We plan to develop a Generative Adversarial Network (GAN) based architecture with a modified dual-path generator inspired by DepthGAN. We will delve deeper by exploring potential improvements to the core architecture. The state architecture diffusion model plays an important role in managing the model's parameters and intermediate outputs during forward(noise addition) and reverse(denoising) processes. The dual-path generator, influenced by DepthGAN, maintains spatial consistency by handling depth and texture generation separately but synchronously. An integrated attention mechanism incorporated within the generator allows the network to focus on specific parts of the latent vector or the generated depth map when creating the RGB image. This could lead to sharper details, improved spatial relationships between objects, and a better understanding of the scene layout.

Input and Output The model will take a latent vector as input. This low-dimensional vector will represent the essential properties of the desired indoor scene, such as room size, furniture types, and overall style. The model will generate a realistic RGB image depicting an indoor scene with accurate spatial relationships between objects. The model can also produce a corresponding depth map for further analysis or manipulation.

2. Related Work

Shi et al. 2022 serves as the primary inspiration for our project. It introduces DepthGAN, a similar dual-path generative network architecture that leverages depth information for improved indoor scene synthesis. Depth-GAN utilizes a dual-path generator: one path creates depth maps, and the other generates the image conditioned on these depth features. Additionally, a special switchable discriminator helps differentiate real from fake images while also predicting depth from the generated scene. This feedback loop between depth and image creation allows DepthGAN to synthesize indoor scenes with impressive visual quality and a higher degree of 3D consistency compared to previous methods.

Mildenhall et al. 2020 paper introduces a novel method

for rendering realistic and complex scenes from a set of 2D images. It achieves this by learning a continuous function that predicts not only the color but also the density of light rays along each viewing direction. This allows NeRF to capture the intricate details of a scene, including lighting effects and object boundaries, resulting in high-fidelity renderings from various viewpoints.

Noguchi and Harada 2020 learn to create 3D representations of objects from regular images by generating images with color and depth information. It achieves this by training on natural images and enforcing consistency between images generated from different viewpoints, all without needing 3D labels. This method represents a step towards understanding 3D shapes from regular photos.

Dhariwal and Nichol 2021 demonstrates that diffusion models have emerged as a superior alternative to Generative Adversarial Networks (GANs).The model offers great stability and diversity in image synthesis, unlike GANs, which often suffer from issues like mode collapse, diffusion models maintain a more consistent performance across various datasets.

Vaswani et al. 2023 proposes a new architecture called the Transformer for sequence-to-sequence tasks like machine translation. Unlike prior models that rely on recurrence or convolution, the Transformer uses an attention mechanism to understand relationships between elements in a sequence. This allows for more efficient processing and achieves state-of-the-art results on various tasks, demonstrating that attention is a powerful tool for modeling long-range dependencies in sequences.

3. Network Architecture

The proposed network architecture is designed for the generation and discrimination of RGB-D scenes, utilizing a generator and a discriminator with sophisticated components aimed at enhancing scene synthesis and discrimination capabilities.

3.1. Generator Architecture

Encoder: The generator's encoder module receives concatenated RGB and depth images as input. It comprises a series of convolutional layers, each followed by leaky Rectified Linear Unit (ReLU) activation functions. These convolutional layers are responsible for extracting hierarchical features from the input data. Notably, self-attention mechanisms are strategically integrated within the encoder architecture to enable the network to selectively focus on informative regions in the feature maps, promoting the generation of realistic and contextually coherent scenes.

- **Self-Attention Mechanism:** The self-attention layers are employed to capture long-range dependencies by computing attention scores that indicate the importance of different regions in the feature maps. This mechanism helps in enhancing the feature representation by allowing the network to focus on critical areas, thereby improving the quality of the generated scenes.

Decoder: The decoder module of the generator operates in tandem with the encoder to reconstruct the scene from the extracted features. Utilizing transposed convolutional layers, the decoder progressively upsamples the encoded feature maps to generate high-resolution RGB-D scenes. Each transposed convolutional layer is accompanied by batch normalization and ReLU activation functions to stabilize training and introduce non-linearity.

3.2. Discriminator Architecture

Convolutional Layers: The discriminator architecture consists of a sequence of convolutional layers designed to process concatenated RGB-D input images. Similar to the generator, these convolutional layers are equipped with leaky ReLU activation functions to introduce non-linearity and facilitate feature extraction. By analyzing both RGB and depth information, the discriminator learns to effectively distinguish between real and fake scenes.

Fully Connected Layers: Following the convolutional layers, the discriminator employs fully connected layers to further process the extracted features. These layers transform the flattened feature tensor into a single output, representing the probability of the input being real or fake. By using a sigmoid activation function, the discriminator produces a probability score indicating the likelihood of the input scene being real.

3.3. Self-Attention Mechanism in Generator

The self-attention mechanism is an integral part of the generator architecture. It allows the network to compute attention scores for different regions in the feature maps, facilitating the capture of long-range dependencies and global contextual information. This mechanism enhances the network's ability to generate scenes by focusing on the most relevant regions in the input data.

- **Attention Scores:** The attention scores are calculated by comparing the similarity between different regions in the feature maps. Higher scores indicate more important regions, guiding the network to prioritize these areas during the generation process.

- **Integration in Generator:** In the generator, self-attention layers are placed within both the encoder and decoder. This integration allows the network to maintain spatial coherence and capture detailed features throughout the entire generation process.

The adversarial training of the generator and discriminator encourages continuous improvement in the quality of the generated RGB-D scenes. The incorporation of self-attention mechanisms in the generator fosters better feature learning and generation by capturing global contextual information and long-range dependencies effectively.

4. Limitations of GANs and Possible Solutions

4.1. Limitations

- **Mode Collapse**: Mode collapse remains a prevalent challenge in GANs, where the generator produces a limited variety of samples, ignoring the diversity present in the dataset. This phenomenon restricts the generative capacity of the model, hindering its ability to capture the full spectrum of data distribution. Mode collapse often occurs due to adversarial training dynamics, where the generator and discriminator engage in a competitive process.

- **Training Instability**: The training of GANs is notoriously unstable, characterized by oscillations in the loss functions and difficulties in convergence. This instability poses a significant hurdle in achieving high-quality sample generation and can lead to suboptimal results. The complex interplay between the generator and discriminator networks, coupled with the non-convex nature of the objective function, exacerbates training instability.

- **Evaluation metrics**: Evaluating the performance of GANs poses a significant challenge due to the lack of reliable metrics that can accurately measure the quality and diversity of generated samples. Traditional metrics like Inception Score (IS) and Frechet Inception Distance (FID) have limitations in capturing the perceptual quality of generated images. Additionally, these metrics may not always correlate with human judgment, further complicating the evaluation process.

- **Data Efficiency**: GANs typically require a large amount of data to train effectively, which may not always be available, especially in specialized domains or niche applications. This reliance on extensive training data limits the scalability and applicability of GANs in real-world scenarios where data scarcity is a prevalent issue.

4.2. Solutions

- **Mode Collapse Mitigation**: To address mode collapse, several strategies have been proposed, including architectural modifications and training techniques. Methods such as minibatch discrimination, feature matching, and adding noise to inputs can encourage the generator to produce more diverse samples, mitigating mode collapse and enhancing the overall quality of generated outputs.

- **Training Stability Enhancement**: Stabilizing the training process of GANs is crucial for achieving convergence and generating high-quality samples consistently. Techniques such as progressively growing GANs (PGGANs) Karras et al. 2018, spectral normalization Miyato et al. 2018, and gradient penalties (e.g., Wasserstein GANs) Arjovsky, Chintala, and Bottou 2017 have been introduced to mitigate training instability and promote smoother optimization trajectories.

- **Improved Evaluation Metrics**: Developing more reliable evaluation metrics is essential for accurately assessing the performance of GANs. Recent research efforts focus on devising metrics that align more closely with human perception, such as Precision and Recall scores or leveraging human evaluations. These metrics aim to provide a more comprehensive understanding of the visual quality and diversity of generated samples.

- **Data-Efficient Approaches**: To overcome the data efficiency limitations of GANs, techniques such as transfer learning and self-supervised learning have been explored. Transfer learning enables the fine-tuning of pre-trained generators on smaller datasets, leveraging knowledge from larger datasets to improve performance on limited data Kornblith, Shlens, and Le 2019. Similarly, self-supervised learning techniques can enhance data efficiency by utilizing unsupervised signals for training, reducing the reliance on annotated data.

5. Comparison with State-of-the-Art Diffusion Models:

The proposed method in the paper includes depth information as 3D prior to help generate indoor scenes more accurately. As indoor spaces vary widely and lack a consistent structure, it uses 3D geometry to create scenes that look realistic. On the other hand, diffusion models focus on modeling data distribution directly without considering 3D information. While diffusion models excel at making high quality samples, they might not capture the detailed 3D structure of indoor scenes as well as methods that take depth into account. Controlling the spatial arrangement of objects in the scene can be difficult with diffusion models, as they mainly operate in a pixel space without explicit knowledge of scene geometry. Additionally, diffusion models use an iterative process of denoising data to generate images while the proposed solution uses a dual path generator framework. The proposed method includes a switchable discriminator that not only distinguishes between real and fake domains

but also predicts the depth from a given input, while diffusion models do not involve discriminators in their training process as they focus on modeling the data distribution directly rather than using an adversarial framework. If we compare the quality of produced samples, the proposed method generates impressively good quality indoor scenes. While diffusion models are known for producing high quality samples too, they might not be effective in capturing the 3D consistency and structure of indoor scenes as compared to methods explicitly used for 3D aware synthesis. Regarding the computational resources involved with both the models, diffusion models may require fewer computational resources as compared to the proposed method. This is because the proposed method uses depth information and a dual path generator which requires additional computational resources compared to a simpler generation process based on iterative denoising steps during the training phase in diffusion models . In terms of the training time, diffusion models may have an advantage over the proposed method involving depth calculations and dual path generation. Overall, while diffusion models may offer advantages in terms of computational efficiency and training time, the proposed method's additional complexity allows for the incorporation of depth information and potentially results in more accurate and realistic scene synthesis compared to diffusion models.

6. Experimental Evaluation

In this section, we present the experimental results and evaluations of the proposed RGB-D scene generation and discrimination model. We assess the performance of the model using quantitative metrics and qualitative analyses.

6.1. Dataset and Training Setup

We trained our model on a subset of the LSUN bedroom dataset. The RGB images in this dataset were passed through LeRES (Yin et al. 2022) to obtain depth images. The rgb and depth images are then passed on to our model. We trained and evaluated our model using a large-scale RGB-D dataset containing diverse indoor and outdoor scenes. The dataset was split into training, validation, and test sets with ratios of 80%, 10%, and 10%, respectively. The model was trained for 100 epochs using an Adam optimizer with a learning rate of 0.0002. We employed a batch size of 8 and utilized an attention mechanism within the generator and discriminator architectures to enhance feature learning and scene synthesis.

6.2. Quantitative Evaluation

To quantitatively evaluate the performance of the model, we computed several metrics including generator and discriminator losses, as well as the Frechet Inception Distance (FID) score. The generator loss and discriminator loss were monitored throughout the training process to assess the convergence and stability of the adversarial training. Additionally, FID scores were calculated on the validation set periodically to measure the similarity between real and generated scenes in terms of feature distributions.

Figure 1. Results from the model

6.3. Qualitative Evaluation

We conducted qualitative evaluations by visually inspecting the generated RGB-D scenes. Sample scenes were generated using the trained generator and compared against ground truth scenes from the dataset. We analyzed the quality, diversity, and realism of the generated scenes, focusing on the ability of the model to synthesize contextually coherent RGB-D representations.

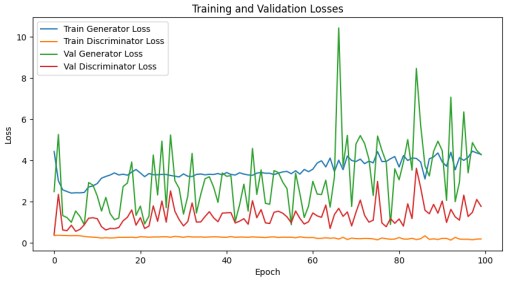

Figure 2. Training and Validation plot

6.4. Results and Discussion

Our experimental results demonstrate the efficacy of the proposed model in generating realistic RGB-D scenes. The

generator loss and discriminator loss exhibited stable convergence during training, indicating effective adversarial learning. Although at times, the discriminator learned to distinguish between real and fake images accurately, which led to increased generator loss. Moreover, the FID scores consistently decreased over epochs on the validation set, suggesting that the generated scenes closely resemble real scenes in terms of visual features.

Qualitatively, the generated RGB-D scenes exhibit diverse textures, colors, and structures, capturing intricate details present in real-world scenes. The attention mechanisms integrated within the generator architectures enhance the model's ability to focus on relevant regions and improve spatial coherence. Due to limited time and computing resources, the architecture and training had to be scaled down, but with enough resources, the model could generate even more high fidelity and realistic scenes with lower FID scores.

Overall, the experimental evaluations validate the effectiveness of the proposed RGB-D scene generation and discrimination model, showcasing its potential applications in computer vision tasks such as virtual environment synthesis, augmented reality, and scene understanding.

References

Arjovsky, Martin, Soumith Chintala, and Léon Bottou (2017). *Wasserstein GAN*. arXiv: 1701.07875 [stat.ML].

Dhariwal, Prafulla and Alex Nichol (2021). *Diffusion Models Beat GANs on Image Synthesis*. arXiv: 2105.05233 [cs.LG].

Karras, Tero et al. (2018). *Progressive Growing of GANs for Improved Quality, Stability, and Variation*. arXiv: 1710.10196 [cs.NE].

Kornblith, Simon, Jonathon Shlens, and Quoc V. Le (2019). *Do Better ImageNet Models Transfer Better?* arXiv: 1805.08974 [cs.CV].

Mildenhall, Ben et al. (2020). *NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis*. arXiv: 2003.08934 [cs.CV].

Miyato, Takeru et al. (2018). *Spectral Normalization for Generative Adversarial Networks*. arXiv: 1802.05957 [cs.LG].

Noguchi, Atsuhiro and Tatsuya Harada (2020). *RGBD-GAN: Unsupervised 3D Representation Learning From Natural Image Datasets via RGBD Image Synthesis*. arXiv: 1909.12573 [cs.CV].

Shi, Zifan et al. (2022). *3D-Aware Indoor Scene Synthesis with Depth Priors*. arXiv: 2202.08553 [cs.CV].

Vaswani, Ashish et al. (2023). *Attention Is All You Need*. arXiv: 1706.03762 [cs.CL].

Yin, Wei et al. (2022). "Towards Accurate Reconstruction of 3D Scene Shape from A Single Monocular Image". In: *TPAMI*.