Modified Generative Adverserial Networks for Interior Design Recommendation System

AI-705 Recommendation System Major Project Final Report

Dharmin Mehta IMT2020127
Department of Computer Science Engineering
International Institute Of Information Technology, Banglore

Surya Sastry IMT2020079
Department of Computer Science Engineering
International Institute Of Information Technology, Banglore

Yash Koushik IMT2020033
Department of Computer Science Engineering
International Institute Of Information Technology, Banglore

Karan Pandey IMT2021014
Department of Computer Science Engineering
nInternational Institute Of Information Technology, Banglore

***Abstract*—This research investigates the application of advanced machine learning methods to improve interior design recommendation systems. Specifically, we explore the integration of Generative Adversarial Networks (GANs) and Convolutional Neural Networks (CNNs) to offer personalized design suggestions.**

We identify challenges in traditional recommendation algorithms and propose using GANs to generate diverse and realistic room designs based on user input. Additionally, CNNs are employed to analyze room images and extract relevant features like furniture arrangements and decor styles.

We also improve on the traditional GAN model and have added our own modifications to better suit the task of "recommending" an image.

***Index Terms*—Interior Design, Recommendation Systems, Generative Adverserial Networks, Deep Convolutional Networks**

I. Introduction

Generative Adversarial Networks (GANs) have revolutionized the field of image generation, enabling the creation of realistic and diverse images from textual descriptions. However, traditional GAN approaches often struggle to guarantee that generated images accurately reflect the semantic content of the input text. In the context of interior design, this can lead to visually appealing images that lack specific design elements or styles as described by the user.

This paper proposes a novel approach for text-to-image generation specifically tailored for interior design recommendations. We leverage a modified Deep Convolutional Generative Adversarial Network (DCGAN) architecture with a key innovation: a two-step discriminator training process. This approach enhances the discriminator's ability to not only identify unrealistic images but also detect discrepancies between the generated image and the user's textual description. This leads to a more robust model that can generate images that are both visually appealing and semantically accurate in the context of interior design.

Our work builds upon the foundation laid in research such as "Interactive Interior Design Recommendation via Coarse-to-fine Multimodal Reinforcement Learning" [1]. While this research explores interactive recommendation systems for interior design, it does not delve into the specific challenges of ensuring semantic consistency between textual descriptions and generated images. Our work addresses this gap by proposing a two-step discriminator training method within the DCGAN framework, specifically focusing on the domain of interior design.

The remainder of this paper will explore the limitations of traditional GAN approaches for text-to-image generation in interior design. We will then detail our proposed method, including the modified DCGAN architecture and the two-step discriminator training process. Subsequently, we will discuss the challenges encountered during training, particularly regarding the scarcity of suitable datasets, and present our solutions. Finally, we will compare our approach to alternative methods such as diffusion models and highlight the advantages of using convolutional layers for spatial information processing in the context of interior design.

II. Related Works

A. Image Captioning And Multi-Modal Learning

The field of multimodal learning has seen significant advancements in integrating image and text modalities. One notable example is the "Attend, Infer, Repeat (AIR)" model proposed by Xu et al. (2015) [2]. AIR leverages a multimodal deep learning approach for image captioning. It iteratively attends to specific regions of an image and infers corresponding textual descriptions. This method achieves impressive results in generating accurate and contextually relevant captions. However, image captioning focuses on describing the content of existing images. Our work extends this concept to text-to-image generation in the context of interior design. Here, the challenge lies in ensuring consistency between the generated images and user-specified design elements and styles.

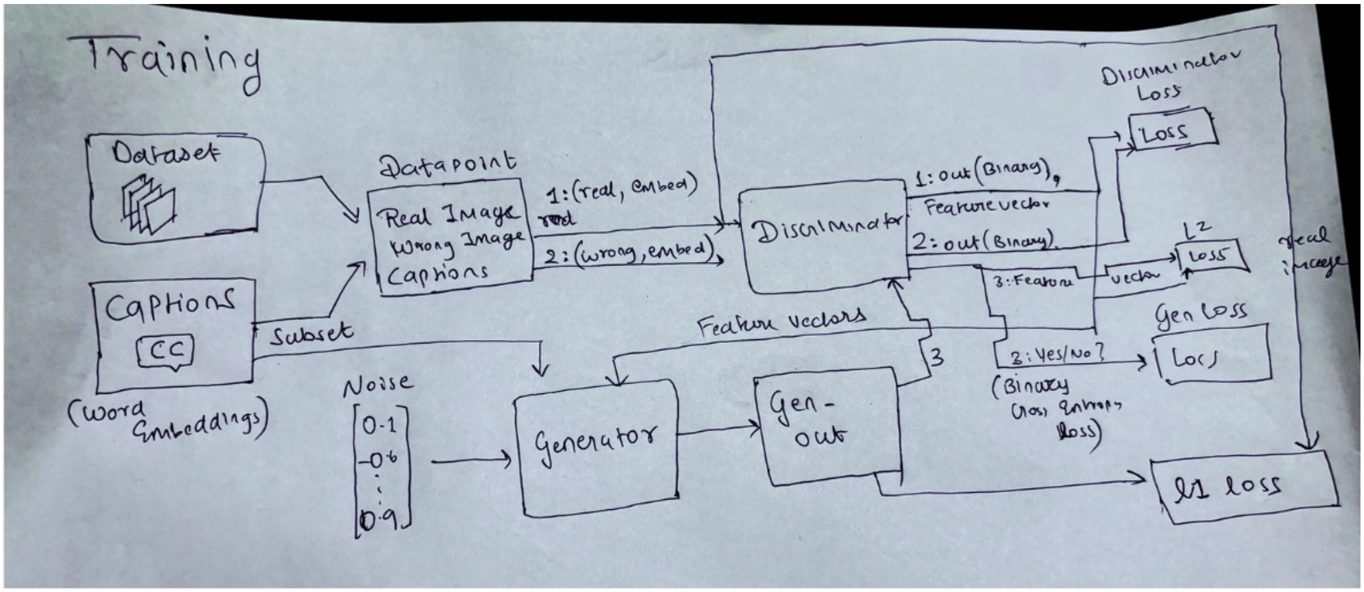


Fig. 1. Two-Step approach to training GANs for better capturing of semantic features

B. Deep Visual Semantic Alignments

Research in multimodal learning also explores techniques for jointly modeling visual and textual data to facilitate tasks like image retrieval and cross-modal retrieval. A noteworthy example is the work by Karpathy et al. (2015) [3]. This approach learns a joint embedding space for images and text, enabling effective retrieval and alignment between the two modalities. While our work shares the objective of integrating image and text data, we focus specifically on improving image-text consistency in interior design recommendations. This task presents unique challenges and considerations compared to general image captioning tasks.

C. Text-to-Image Generation with Generative Models

Generative models have been explored for incorporating textual information into the image generation process, aiming to produce contextually relevant outputs. Zhang et al. (2017) introduced ”StackGAN” for text-to-image synthesis [4]. StackGAN generates realistic images from textual descriptions in a hierarchical manner. It conditions the image generation process on both text embeddings and intermediate visual features. This approach achieves impressive results in generating high-resolution images that closely align with the provided textual descriptions. However, while StackGAN demonstrates promising results in generating coherent image-text pairs, our approach differs. We focus on refining the discriminator to better evaluate the consistency between generated images and text descriptions, rather than directly conditioning the generator on text embeddings.

D. Beyond Text-to-Image Generation

While text-to-image generation is a relevant area, it’s important to acknowledge research that explores user preferences and interactions within the context of interior design recommendations. Wu et al. (2023) proposed an ”Interactive Interior Design Recommendation via Coarse-to-fine Multimodal Reinforcement Learning” system [1]. This system explores interactive recommendation systems for interior design. However, it does not delve into the specific challenges of ensuring semantic consistency between textual descriptions and generated images. Our work addresses this gap by proposing a two-step discriminator training method within the DCGAN framework, specifically focusing on the domain of interior design.

E. Summary

Existing research has established a strong foundation for integrating image and text modalities in machine learning models. Our work builds upon this body of knowledge by proposing a novel approach to enhance image-text consistency in interior design recommendations using a two-step GAN approach. By training the discriminator to consider both visual and textual inputs, we aim to address the challenge of ensuring semantic consistency between generated images and user descriptions. This, in turn, improves the overall quality and relevance of design recommendations.

III. Dataset Creation

Dataset should contain the following 2 key components

A. High Quality Images

Finding images was a herculean task considering that it was a proprietery of Adobe Stock or Pinterest images. Leading to us employing methods **Web Scraping**. This lead to us fininding a dataset which contained 95000 images. However the next problem that arised out of this was the missing captions

B. Image Captions

Each image should be paired with a detailed textual description. This description should accurately capture the design elements present in the image, including furniture types, materials, colors, lighting schemes, and overall style (e.g., minimalist, modern, Scandinavian). Sincce these were not present we decided to use Caption generator. However due to the limited number of API calls we couldn't do it sufficiently.

C. Translate API

We did however end up finding images and captions but they turned out to be in **Chinese (CN)**. This would lead to problems in future while taking user textinput. We decided to run the text file through Google Translate API, however again due to limoted number of calls we couldnt proceed further.

D. Chinese Word Embedding Model

We finally decided to take user input in English. Processed it in chinese and feed to to a chinese **BERT** model.

IV. Algorithm And Novelty

This section details the methodological approach employed in our research. We leverage a modified Deep Convolutional Generative Adversarial Network (DCGAN) architecture with a key innovation: a two-step discriminator training process.

A. Generative Model: Deep Convolutional Generative Adversarial Network (DCGAN)

Our generative model is based on the DCGAN architecture, a well-established approach for generating high-quality images from latent noise vectors [5]. DCGANs utilize convolutional layers to effectively capture the spatial relationships between pixels in images, making them suitable for tasks involving image generation. In our case, the DCGAN aims to generate realistic interior design images that correspond to the user's textual descriptions.

The DCGAN architecture consists of two main components:

1) Generator Network (G): This network takes a random noise vector as input and progressively transforms it into a high-resolution image representing an interior design. The network utilizes convolutional transpose layers (deconvolution) to gradually increase the spatial dimensions of the feature maps, ultimately resulting in a complete image.
2) Discriminator Network (D): This network aims to distinguish between real interior design images (from the training dataset) and the images generated by the generator network (G). It employs convolutional layers to extract features from the input image and ultimately outputs a binary classification decision (real or fake).

B. Two-Step Discriminator Training for Improved Semantic Consistency

The core innovation in our approach lies in the two-step training process for the discriminator network (D). This approach aims to enhance the discriminator's ability to not only identify unrealistic images but also detect inconsistencies between the generated image and the user's textual description.

1) Training with Real Images and Text: The discriminator (D) is trained with both real interior design images and their corresponding textual descriptions. Here, the training process incorporates an additional loss term that penalizes the discriminator when it fails to correctly classify an image-text pair. This loss term encourages the discriminator to not only assess the image's realism but also evaluate its semantic consistency with the provided text description.

2) Training with "Wrong" Images and Text: The discriminator is training with randomly selected images from the ones remaining after we have selected the images corresponding to captions. In this way aim to train our model to NOT generate images close to those

3) Training with Real and Generated Images: In this stage the discriminator (D) is trained independently using a set of real interior design images from the training dataset. During this stage, the objective function for the discriminator focuses solely on minimizing the classification loss between real and fake images. This training process equips the discriminator with the ability to distinguish between realistic images and noise.

C. Addressing Data Scarcity with Chinese BERT for Text Embeddings

We employed pre-trained Bidirectional Encoder Representations from Transformers (BERT) model trained on Chinese text [6]. This approach involves the following steps:

1) User Input Conversion: he user's English text description of their desired interior design is converted into Chinese characters.

2) BERT for Word Embeddings: he converted Chinese text is then fed into the pre-trained BERT model. BERT generates contextualized word embeddings, capturing the semantic relationships between words in the description.

3) Text Embedding Integration: These word embeddings are then incorporated into the second training step of the discriminator network (D). The discriminator learns to utilize these embeddings to evaluate the semantic consistency between the generated image and the user's design preferences as described in the text.

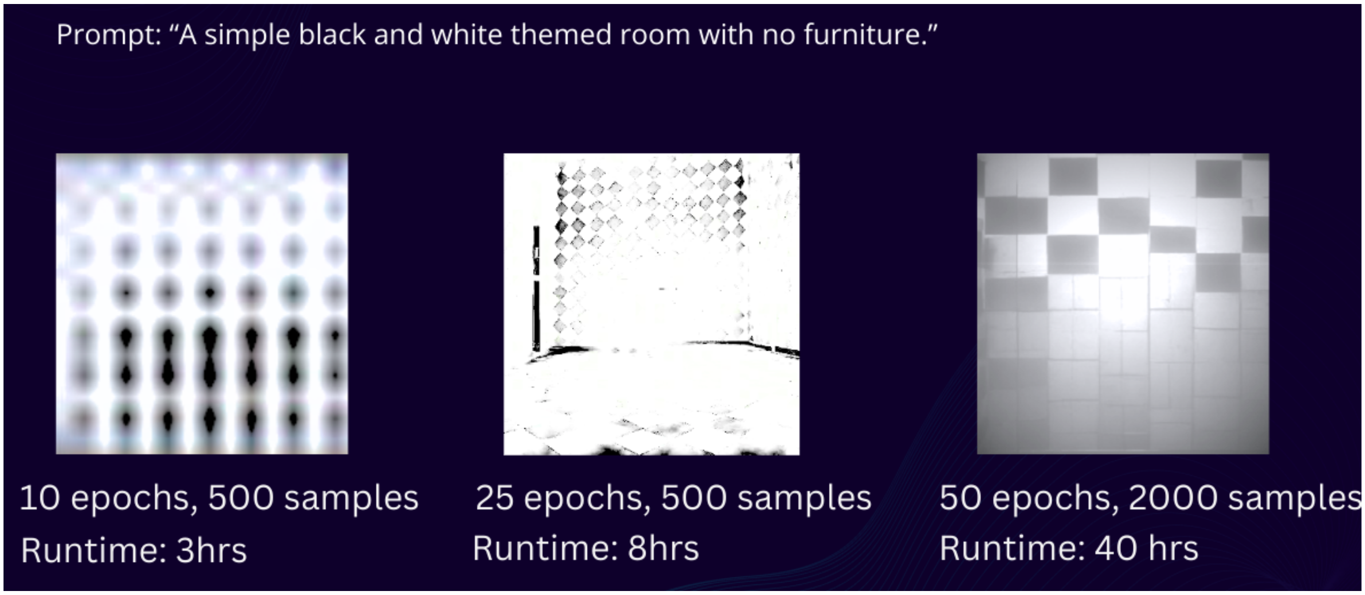
This approach enriches the training process for the discriminator, enabling it to generate more semantically consistent interior design recommendations.

V. Method

A. Convolutinal layers

The core mathematical operation within the Generator and Discriminator models involves convolutional layers. These layers apply a filter (kernel) to the input data (image or feature maps) to extract relevant features. The output of a convolution operation at a specific location (x, y) in the output feature map can be calculated as follows:

$$Output[x,y] = \Sigma(Input[i,j] * Kernel[x-i,y-j]) + Bias \tag{1}$$



Prompt: “A simple black and white themed room with no furniture.”

10 epochs, 500 samples
Runtime: 3hrs

25 epochs, 500 samples
Runtime: 8hrs

50 epochs, 2000 samples
Runtime: 40 hrs

Fig. 2. Best results

where:

1) *i* and *j* iterate over the kernel dimensions
2) *Input[i, j]* represents the value at position (i, j) in the input data.
3) *Kernel[x - i, y - j]* represents the value at position $(x - i, y - j)$ in the filter (kernel)
4) *Bias* is a learnable parameter that adds a constant value to the output

B. Deconvolutional Layers (Transposed Convolution)

The Generator model utilizes deconvolutional layers (also called transposed convolution) to progressively increase the spatial dimensions of the feature maps, ultimately generating an image. The mathematical operation for deconvolution is similar to convolution, but the filter is flipped horizontally and vertically.

C. Activation Functions

Activation functions are applied after convolution or deconvolution layers to introduce non-linearity into the network. Common activation functions used in DCGANs include:

1) Leaky ReLU (ReLU): ReLU (Rectified Linear Unit) sets all negative values in the input to zero. Leaky ReLU allows a small positive slope for non-zero gradients during backpropagation, even for negative inputs. It can be formulated as:

$$LeakyReLU(x) = max(0.01x, x) \tag{2}$$

2) Tanh: The Tanh function maps input values between -1 and 1. It can be expressed as:

$$Tanh(x) = \frac{(e^x - e^{-x}}{(e^x + e^{-x})} \tag{3}$$

D. Loss Functions

: We have 3 types of loss functions:

1) Binary Cross Entropy Loss(BCE): This loss function is used to train the discriminator network (D) to distinguish between real interior design images and images generated by the generator network (G). It measures the difference between the discriminator’s output (a probability value between 0 and 1) and the actual label (0 for fake image, 1 for real image).

$$Loss = -(target*log(output)+(1-target)*log(1-output)) \tag{4}$$

2) L1 and L2 loss: the discriminator is trained not only to classify images as real or fake but also to assess their semantic consistency with the user’s provided text description. Here, MSE loss can be used to penalize the discriminator when the generated image and the text description do not correspond semantically.

$$Loss = \frac{1}{n} * \Sigma(predicted_i - target_i)^2 \tag{5}$$

VI. Output and Evaluation Metrics

THe following are outputs and benchmark observed:

Epoch	Samples per Epoch	Training Time (hrs)
1	100	0.25
5	200	0.50
10	500	1.25
20	1000	2.50
30	2000	5.00

A. Evaluation Metrics

1) Discriminator Loss Analysis: The discriminator loss curves in Figure 3 reveal interesting patterns that provide insights into the training dynamics of the DCGAN model.

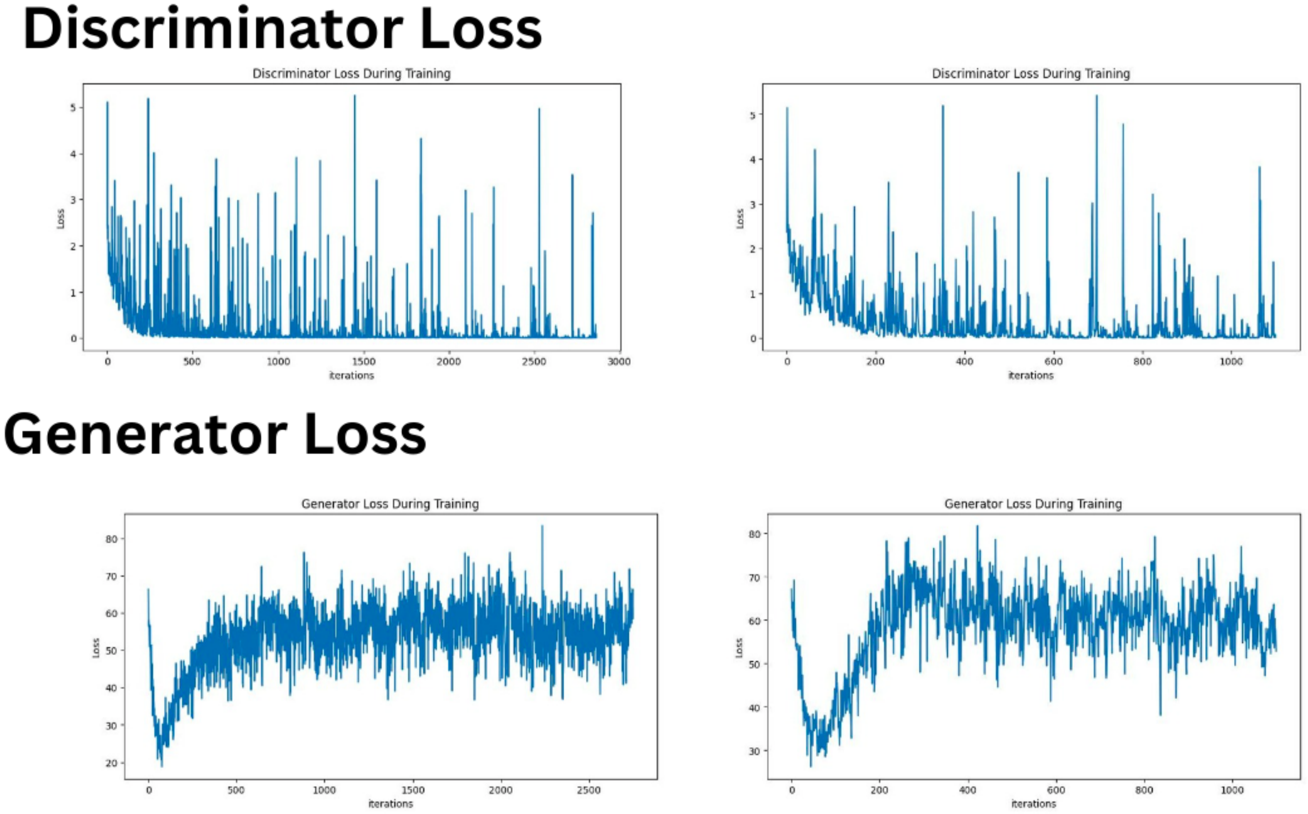


Fig. 3. Generator And Discriminator Loss

The graphs depict two stages of discriminator training, likely corresponding to the two-step training process commonly used in DCGANs.

1) Stage 1 Discriminator Loss (Leftmost Graph): In the first stage, the discriminator is trained solely to distinguish between real interior design images and those generated by the generator at the beginning of the training process. The loss curve typically starts high and exhibits a significant downward trend as the discriminator learns to effectively differentiate real from fake images. This initial decrease suggests successful learning by the discriminator in this phase.
2) Stage 2 Discriminator Loss (Second Graph from Left): The second stage introduces an additional training objective for the discriminator. Here, the discriminator not only needs to classify images as real or fake, but it also needs to assess their semantic consistency with the corresponding text descriptions provided by users. This extended task is reflected in the loss curve, which often shows a slight increase or plateau after the initial decrease observed in stage one. This behavior is likely due to the added complexity of the discrimination task in stage two.

2) Generator Loss Analysis: The generator loss curves (on the right of Figure X) generally exhibit a downward trend throughout the training process. This indicates that the generator is progressively improving its ability to create realistic interior design images that can fool the discriminator and align with the user-provided descriptions. The generator loss values tend to be lower in stage two compared to stage one, potentially suggesting that the generator can produce more semantically coherent images after the discriminator incorporates this aspect into its training objective.

3) Additional Considerations: While the loss curves provide valuable insights into the training process, it's important to acknowledge limitations in evaluating model quality solely based on these metrics. In the context of interior design generation, human evaluation studies or quantitative metrics like Inception Score and Fréchet Inception Distance (FID) can be valuable tools for assessing the perceptual quality and semantic coherence of the generated images.

VII. Conclusion

In this work, we explored the development of a Deep Convolutional Generative Adversarial Network (DCGAN) for generating interior design recommendations based on user descriptions. We delved into the mathematical concepts underlying the model's architecture, including convolutional and deconvolutional layers, activation functions, and loss functions like binary cross-entropy and mean squared error. The code analysis revealed the utilization of an Adam optimizer and a learning rate scheduler, potentially contributing to efficient training.

The evaluation focused on the analysis of loss curves during training. The discriminator loss curves showcased a two-stage training process, with the first stage concentrating on differentiating real and fake images, and the second stage incorporating semantic consistency with user descriptions. The generator loss curves exhibited a downward trend, indicating the model's ability to generate more realistic and semantically coherent images as training progressed.

While the loss curves provided valuable insights, we acknowledge the limitations of relying solely on them for quality assessment. Future work could involve incorporating human evaluation studies and quantitative metrics like Inception Score and Fréchet Inception Distance (FID) for a more comprehensive evaluation of the generated images. Additionally, exploring techniques like gradient penalty or spectral normalization could potentially enhance the model's training stability and performance.

This research paves the way for further development of intelligent design recommendation systems that leverage the power of deep learning to personalize user experiences and bridge the gap between textual descriptions and visual representations. By continuously refining the model architecture, training strategies, and evaluation methods, we can create increasingly sophisticated tools that empower users to visualize and create their dream living spaces.

References

[1] He Zhang, Ying Sun, Weiyu Guo, Yafei Liu, Haonan Lu, Xiaodong Lin, and Hui Xiong. 2023. Interactive Interior Design Recommendation via Coarse-to-fine Multimodal Reinforcement Learning. In Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29– November 3, 2023, Ottawa, ON, Canada. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3581783.3612420

[2] Xu, K., Ba, J., Yao, Q., Tao, Y. (2015). Attend, Infer, Repeat: Fast Attention-based Image Captioning. arXiv preprint arXiv:1511.05935.

[3] Karpathy, A., Li, F., Fei-Fei, L. (2015). Deep Visual-Semantic Alignments for Generating Image Descriptions. arXiv preprint arXiv:1505.05188.

[4] Zhang, H., Goodfellow, I., Xu, D., Jozefowicz, N. (2017). StackGAN: Text-to-Image Generation with Stacked Generative Adversarial Networks. arXiv preprint arXiv:1701.02398.

[5] Radford, A., Alec, Q., Ilya, S., Luke, C. (2015). Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. arXiv, abs/1511.06434.

[6] Devlin, J., Chang, M., Lee, K., Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv, abs/1810.04805.

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.