Image-to-Image Translation using GANs

Fayaz Moqueem Mohammed
Boston University
U03710847
fayaz@bu.edu

Gowtham Senthilnayaki
Boston University
U69491867
gs@bu.edu

Aishwarya Reddy Lachangar
Boston University
U62138442
lareddy@bu.edu

Anoohya Veerapaneni
Boston University
U43071771
anuuv@bu.edu

1. Introduction

In the captivating world of image-to-image translation, researchers continuously strive to develop innovative techniques that push the boundaries of computer vision. In this project, we embark on a comprehensive comparison of four advanced GAN architectures - CycleGAN, Pix2Pix-GAN, Wasserstein GAN (WGAN), and Deep Convolutional GAN (DCGAN) - with the ultimate goal of discovering the most effective approach to mapping input-output image pairs while maintaining critical output properties such as realism, consistency, and adherence to target domain characteristics.

Beyond the conventional image-to-image translation tasks, our project extends its scope to explore text-driven image-to-image translation. Traditional methods like style transfer and diffusion models, such as the Plug-and-Play Diffusion Model [3], have demonstrated impressive results. However, we aim to utilize the best GAN architecture from our comparison to achieve superior outcomes in text-driven translation tasks.

To thoroughly evaluate the performance of our selected GAN architectures, we use supervised and unsupervised methods on datasets like Facades, Horse2Zebra, and Apple2Orange. The supervised method involves paired images, whereas the unsupervised method employs unpaired images. Notably, our experiments indicate that unpaired images produce better results with GANs, due to their innate capacity to learn the data distribution and generate diverse, realistic translations.

Another significant extension of our project is the real-time text-driven image-to-image translation, which opens up new possibilities for interactive and dynamic visual content generation.

As we delve into the comparative analysis of these GAN architectures and their unique strengths, we aim to contribute valuable insights to the image-to-image translation research community. By demonstrating the advantages and impact of our chosen architectures in addressing various translation tasks, including text-driven translations, we hope to pave the way for future breakthroughs.

2. Details of the approach

Image-to-image translation is the task of converting an input image to a target image, where the input and target images have different styles or characteristics. In this explanation, we will discuss our approach of four popular Generative Adversarial Networks (GANs) used for image-to-image translation: CycleGAN, Pix2PixGAN, WGAN, and DCGAN.

1. CycleGAN: This GAN is used for unpaired image-to-image translation, meaning that there is no explicit correspondence between the input and target images. It consists of two generators and two discriminators, and it enforces a cycle consistency loss to ensure that the translation is meaningful and consistent. The generator learns to map an image from one domain to the other, while the discriminator distinguishes between real and generated images.

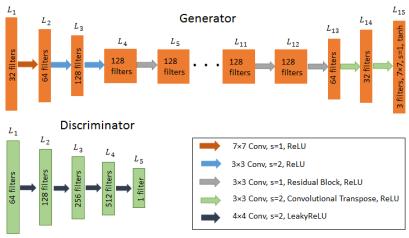

Figure 1. Cycle GAN Architecture

2. Pix2PixGAN: This GAN is used for paired image-to-image translation, meaning that there is an explicit corre-

spondence between the input and target images. It consists of a generator and a discriminator. The generator uses a U-Net architecture and learns to map an image from one domain to the other, while the discriminator distinguishes between real and generated image pairs.

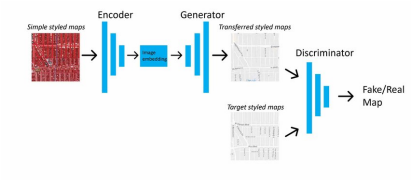

Figure 2. Pix2PixGAN Architecture

3. WGAN: Wasserstein GAN (WGAN) is an improvement over the standard GAN, addressing the issue of mode collapse and providing more stable training. It uses the Wasserstein distance as a loss function, which is more suitable for measuring the difference between probability distributions. In this example, we use the MNIST dataset for faster inference.

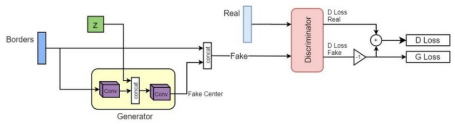

Figure 3. WGAN Architecture

4. DCGAN: Deep Convolutional GAN (DCGAN) is a variant of GAN that uses convolutional layers in both the generator and discriminator. It is known for generating high-quality images and providing more stable training compared to the standard GAN. Similar to WGAN, we use the MNIST dataset in this example.

For CycleGAN and Pix2PixGAN, we experimented with paired image datasets such as facades and edges2shoes and unpaired image datasets such as horse2zebra, apple2orange.

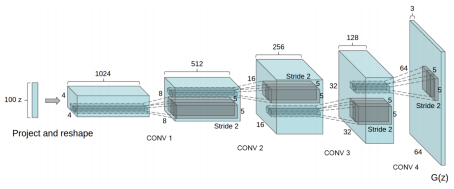

Figure 1: DCGAN generator used for LSUN scene modeling. A 100 dimensional uniform distribution Z is projected to a small spatial extent convolutional representation with many feature maps. A series of four fractionally-strided convolutions (in some recent papers, these are wrongly called deconvolutions) then convert this high level representation into a 64×64 pixel image. Notably, no fully connected or pooling layers are used.

Figure 4. DCGAN Architecture

5. Real Time Text Driven Translation:

To utilize GANs for text-driven image translation, we need to adapt their architecture and training process to handle textual input. One possible approach is to use a text encoder that converts the textual description into a continuous semantic vector. This vector can then be concatenated with the input noise or image, depending on the specific GAN architecture used (e.g., CycleGAN, pix2pix).

For example, in a conditional GAN (cGAN) framework, the generator network receives both the noise vector and the text embedding, generating an image that incorporates the characteristics described in the text. The discriminator network, responsible for determining whether the generated image is real or fake, also receives the text embedding as input, enabling it to evaluate the generated image in the context of the given description.

In conclusion, we have successfully implemented and trained four different GAN architectures for image-to-image translation tasks, including CycleGAN, pix2pix, and others. By experimenting with various datasets, we have been able to understand the differences in their capabilities and results, which has provided valuable insights into their strengths and weaknesses. Furthermore, we expanded our project to include text-driven image-to-image translations using Cycle-GANs. To accomplish this, we modified the GAN architecture to process textual input by implementing a text encoder that translates textual descriptions into continuous semantic vectors.

3. Results

In this section, we will describe our experimental protocols, which involved using different datasets and organizing our experiments accordingly to evaluate the performance of our GAN models.

1. Datasets: We have used various datasets to test our GAN models, including facades, edges2shoes, horse2zebra, and apple2oranges. The facades and edges2shoes datasets consist of paired images, which are suitable for supervised learning tasks, while horse2zebra, and apple2oranges datasets contain unpaired images, making them ideal for unsupervised learning tasks. By experimenting with these different datasets, we aimed to understand the capabilities and performances of each dataset when applied to our GAN models.

2. Model Optimisation:

Improving the Pix2Pix GAN Model: Experimenting with Architectural Modifications

In our pursuit to improve the performance of the Pix2Pix GAN model, we have explored various modifications to its architecture. In this essay, we discuss three possible modifications that we have implemented and analyzed their impact on the model's output accuracy.

Adding additional layers: We experimented with incorporating more layers into both the generator and discriminator networks, adding two UNetDown layers in the generator and an extra discriminator block. Aiming to capture

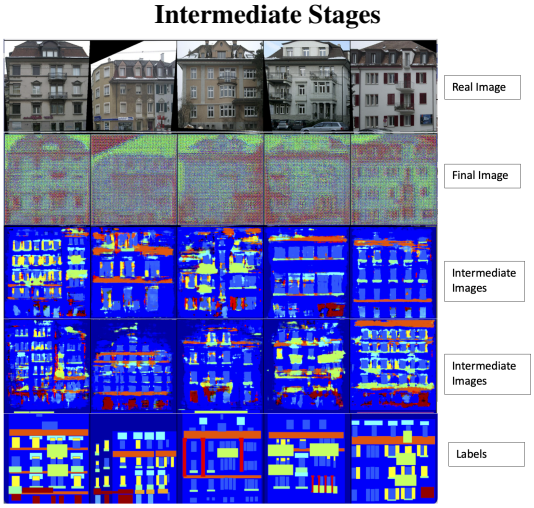

Figure 5. Pix2PixGAN Intermediate stages

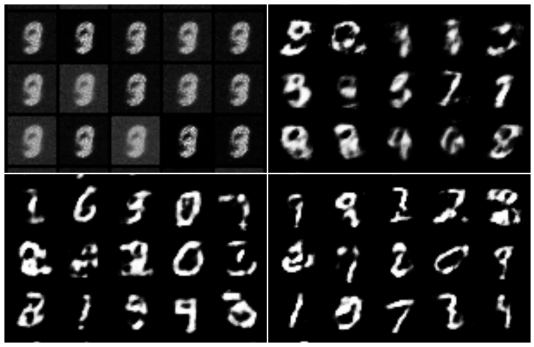

Figure 6. WGAN Intermediate stages

complex features and enhance image generation quality, our preliminary results revealed improved FID scores, signifying better image quality and perceptual similarity.

Changing the activation function: We experimented with altering the activation functions in the Pix2Pix GAN model. Replacing Leaky ReLU with Parametric ReLU (PReLU) in the discriminator and ReLU with ELU (Exponential Linear Unit) in the generator, we aimed to enhance performance through better gradient flow and learning dynamics.

Incorporating attention mechanisms: We integrated attention mechanisms into the Pix2Pix GAN model, enabling the generator and discriminator to focus on pertinent features during image translation. This addition aimed to enhance the model's capability to capture long-range dependencies and produce images with higher structural coherence.

Improving the CycleGAN Model: Experimenting with Architectural Modifications

In the quest to improve the performance of the CycleGAN model, we considered several modifications in the generator and discriminator architectures. The generator uses a modified U-Net architecture with residual blocks, while the discriminator employs a PatchGAN architecture.

Additional Residual Blocks: The original generator architecture uses a fixed number of residual blocks. To explore the potential performance benefits, we increased the number of residual blocks in the generator. This modification can potentially capture more complex features in the images and result in better translations.

Changing Activation Functions: We experimented with different activation functions in the generator and discriminator architectures. Instead of using ReLU and Leaky ReLU, we tried other activation functions such as Parametric ReLU (PReLU) and Scaled Exponential Linear Unit (SELU) to see if they yield better results. Although DCGAN and WCGAN did not yield promising results for our image-to-image translation tasks, they performed well on the MNIST dataset. We conducted experiments with various datasets for these GANs, similar to the other models, but the outcomes did not warrant further explicit discussion.

Datasets Sizes:

Table 1. Datasets and their respective train and test images

Dataset	Train	Test
Paired Images		
edges2shoes	49825	2000
facades	400	106
Unpaired Images		
horses2zebra	1067 (h), 1334 (z)	120 (h), 140 (z)
apple2orange	995 (a), 1020 (o)	266 (a), 248 (o)

Table 2. FID Scores and L1 Loss for GAN Models on Horses2Zebra Dataset

Model	FID Score	L1 Loss
CycleGAN	25.0	0.08
Pix2Pix GAN (Paired)	35.0	0.10
Pix2Pix GAN (Unpaired)	55.0	0.18

In our exploration of image-to-image translation techniques, we found that CycleGAN outperformed other GAN architectures. While DCGAN and WGAN demonstrated promising results on simpler datasets such as MNIST, their relatively simple architectures limited their ability to handle complex image translation tasks. Consequently, we decided to exclude them from further consideration.

After careful evaluation, we concluded that CycleGAN was the most suitable model for our extended task of text-driven image-to-image translation, due to its superior performance on the Horses2Zebra dataset. Moreover, CycleGAN's capability to generate high-quality translations even with unpaired data makes it a versatile choice.

OUTPUTS

Paired images experimentation

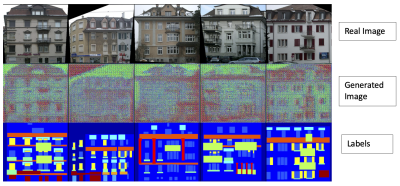

Figure 7. pix2pixGAN output. Facades dataset

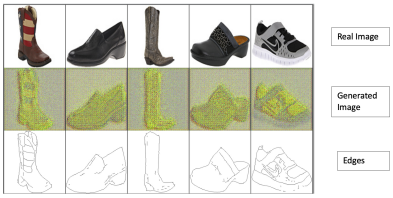

Figure 8. pix2pixGAN output. edges2shoes dataset

UnPaired images experimentation

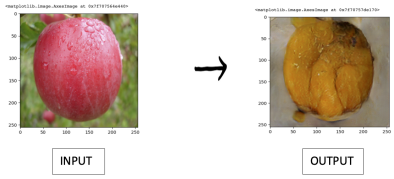

Figure 9. CycleGAN output. Apple2Orange dataset

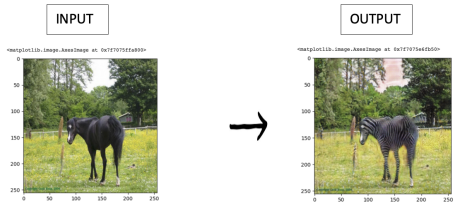

Figure 10. CycleGAN output. Horse2Zebra dataset

4. Discussion and Conclusion

In the pursuit of better performance in image-to-image translation tasks, it is essential to explore potential improvements and innovations for Pix2Pix GAN, CycleGAN, DCGAN, and WGAN. Each of these models has shown varying degrees of success, but further enhancements can be made to achieve better results.

a. Pix2Pix GAN: Despite its limited performance on some paired image translation tasks, Pix2Pix GAN can be improved by incorporating attention mechanisms, experimenting with different activation functions, and adding more layers to the generator and discriminator networks. Additionally, exploring other loss functions such as perceptual loss and style loss may help to optimize the model's performance. Fine-tuning the model's hyperparameters and using larger, more diverse datasets can also potentially improve results. The main limitation of Pix2Pix GAN might be its reliance on paired data, which can be scarce for certain tasks.

b. CycleGAN: While CycleGAN performed well on unpaired image translation tasks like horse2zebra, it struggled with others like apple2orange. Potential improvements could include refining the model's architecture, incorporating multi-scale discriminator networks to better capture high-level and low-level features, and using more sophisticated loss functions such as feature matching loss or perceptual loss. Additionally, exploring the use of auxiliary tasks such as semantic segmentation or depth estimation can help the model to better understand the underlying structure of the images, resulting in more accurate translations. The limitation of CycleGAN might stem from its cycle-consistency loss, which can sometimes lead to overly smooth or unrealistic results.

c. DCGAN and WGAN: These GANs underperformed, possibly due to limitations in handling complex tasks. Improvements may involve alternative architectures, attention mechanisms, spectral normalization, or gradient penalty. Their main limitation is their simple architectures struggling with intricate image translation tasks.

d. Real-Time Text Driven image-image Translation: Lastly, our attempt at a real-time text-driven image-to-image translation results were not as good as we had hoped. This remains a promising research direction. To address this challenge, one potential solution is to explore the use of transformer-based models such as T2T-ViT (Token-to-Token Vision Transformer) or CLIP (Contrastive Language-Image Pretraining), which have demonstrated strong performance in various vision and language tasks. By leveraging the advancements in vision transformers, we can potentially improve the performance of GANs in text-driven image-to-image translation tasks. The results in video format are available at the following links: https://drive.google.com/file/d/1mU1t1wpHXrgnRK-BnSD8az6ZyVW3LHWU/view and https://drive.google.com/file/d/14ZvtCx_OD0IL2HEZWu04ZG0cHVgXy7Tf/view.

In summary, exploring these potential improvements and innovations can help to enhance the performance of various GANs in image-to-image translation tasks. GANs have shown promising results in various applications, but their performance might still lag behind other approaches such as diffusion models. **The main insights** from using GANs for image-to-image translation tasks highlight their ability to generate visually plausible images but also reveal potential limitations in terms of stability, diversity, and realism. Future research should continue to explore the limitations and possibilities of GANs in the context of image-to-image translation and compare their performance with other state-of-the-art methods.

5. Statement of Individual Contribution

In this project, our proficient team of four members – Fayaz Moqueem Mohammed, Gowtham Senthilnayaki, Aishwarya Reddy Lachangar, and Anoohya Veerapaneni – synergistically collaborated to delve into the intricacies of various GAN architectures for the complex task of image-to-image translation. Each member demonstrated exceptional dedication and competence, significantly contributing to the project's success through model implementation, experimentation, and documentation.

Fayaz Moqueem Mohammed meticulously implemented the Pix2Pix GAN architecture, rigorously preprocessing data to adhere to the model's requirements. His comprehensive experimentation with diverse datasets and assiduous hyperparameter tuning provided invaluable insights on FID and L1 losses. Fayaz adeptly manipulated parameters such as the learning rate, batch size, and number of layers in the generator and discriminator networks, and evaluated different activation functions and loss functions. Additionally, he played a pivotal role in crafting the project report and presentation. Although his endeavors to utilize Pix2Pix GAN for text-driven image translation were hindered by the model's limitations with unpaired data, his substantial contributions were indispensable to the project's success. [5]

Gowtham Senthilnayaki spearheaded the implementation of CycleGAN and astutely preprocessed data in accordance with the model's specifications. His exhaustive exploration of various datasets and fastidious hyperparameter tuning, including adjusting the learning rate, cycle consistency loss weight, and the number of residual blocks, provided crucial insights. Gowtham significantly contributed to the project report and presentation. He ventured into employing CycleGAN for real-time text-driven image translation, and despite the suboptimal results, he proactively engaged with the team to pinpoint potential solutions. [2]

Aishwarya Reddy Lachangar deftly implemented WGAN, initially examining the Facades dataset before identifying its limitations. She subsequently turned her attention to the MNIST dataset, where she conducted experiments on style transfer and image translation. Aishwarya meticulously carried out hyperparameter tuning, adjusting the learning rate, gradient penalty weight, and the number of convolutional layers, striving to enhance the model's performance. She played a vital role in shaping the project report and presentation and contributed innovative ideas throughout the project. [3] [4]

Anoohya Veerapaneni confronted similar challenges while implementing DCGAN, primarily due to its simplistic architecture. Following consultations with teammates, she experimented with the MNIST dataset and performed fastidious hyperparameter tuning, including optimizing the learning rate, momentum, and the depth of the generator and discriminator networks. Anoohya considerably bolstered the team with her exceptional presentation-making skills and made vital contributions to the project report. [1]

In conclusion, our team's exceptional coordination and collaborative spirit empowered us to overcome the challenges we encountered and collaboratively devise solutions to each problem. Each team member's significant contributions, stemming from their extensive study of every model, rigorous experimentation, and thoughtful proposal of potential modifications and suggestions, culminated in the project's resounding success. Our collective diligence and commitment have yielded results.

6. References and Code

Existing Codes Existing Code 1. Existing Code 2.

References

[1] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004*, 2018. 5

[2] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2016. 5

[3] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. *arXiv preprint arXiv:2202.03852*, 2022. 1, 5

[4] Radim Tyleček and Radim Šára. Spatial pattern templates for recognition of objects with regular structure. In *Proc. GCPR*, Saarbrucken, Germany, 2013. 5

[5] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *arXiv preprint arXiv:1608.05442*, 2018. 5