Image translation using GANs

Lucy Kumari
Information Technology
National Institute of Technology, Karnataka
Surathkal, India
lucykumari.211it037@nitk.edu.in

Nandini A C
Information Technology
National Institute of Technology, Karnataka
Surathkal, India
nandiniac.211it043@nitk.edu.in

Shamitha M Naik
Information Technology
National Institute of Technology, Karnataka
Surathkal, India
shamithamnaiknitkeduin.211ec248@nitk.edu.in

I. Introduction

The paper explores advancements in Generative Adversarial Networks (GANs) to enhance their performance in generating complex scene images, characterized by multiple objects and intricate layouts. Traditional GAN models often struggle with such images due to the high structural complexity, which challenges the discriminator's ability to effectively distinguish between real and fake images. To overcome this, the paper introduces two innovative approaches: self-supervised learning and transfer learning. The self-supervised approach enhances the discriminator's multi-scale representations by learning useful visual features from the images themselves without requiring labeled data, thereby improving the generative capability of the GAN. Additionally, transfer learning leverages pretrained models from scene understanding tasks to further bolster the discriminator's effectiveness, particularly in complex scenes. The study reports significant improvements, achieving up to a 63%. better Fréchet Inception Distance (FID) score compared to state-of-the-art models, demonstrating the potential of these methods to substantially improve GAN performance in synthesizing diverse and realistic scene images.

II. Literature Review

[1]This paper investigates the challenges associated with generating unseen complex scenes using GANs and other generative models. The authors critically evaluate the state-of-the-art methods and discuss their limitations in capturing the intricate structures and relationships inherent in complex scenes.

a) Merits: Provides a comprehensive analysis of the difficulties in generating complex scenes, identifying key areas for improvement. Offers insights into the limitations of existing models, which is valuable for guiding future research in this area.

b) Demerits: The paper primarily focuses on identifying issues rather than proposing concrete solutions. While the analysis is thorough, it lacks empirical results or new methodologies that directly address the highlighted challenges.

[2]This paper presents a comprehensive survey of object detection techniques developed over the past two decades. The authors provide a detailed overview of the evolution of object detection models, highlighting key milestones and innovations in the field.

c) Merits: The paper serves as an extensive resource for understanding the progression of object detection methods, offering a historical context that is beneficial for both newcomers and experienced researchers.It effectively categorizes the different approaches, making it easier to identify trends and emerging techniques in object detection.

d) Demerits: While the survey is exhaustive, it may overwhelm readers with the breadth of information, making it challenging to focus on specific advancements or current state-of-the-art techniques. The paper does not delve deeply into the practical implementation aspects of the models discussed, which could limit its utility for practitioners seeking to apply these methods.

[3]This paper provides a survey on semantic segmentation using deep learning techniques. It systematically reviews the various models and approaches developed for semantic segmentation, analyzing their performance across different datasets and tasks.

e) Merits: The survey is thorough in its coverage of deep learning methods for semantic segmentation, offering a clear comparison of the strengths and weaknesses of different approaches. It includes a detailed analysis of the performance of various models, which helps in understanding their applicability to different types of datasets and segmentation tasks.

f) Demerits: The paper could benefit from more emphasis on the limitations of current models and potential areas for future research, which would provide a more balanced view. It primarily focuses on the technical aspects of the models, with less attention given to the practical challenges of deploying these models in real-world applications.

[4]This paper introduces a method for multimodal unsupervised image-to-image translation. The authors propose a framework that allows for translating an image from one domain to another without paired data, addressing the challenge of multimodal generation in an unsupervised setting.

g) Merits: The proposed method significantly advances the field of unsupervised image-to-image translation by allowing for multimodal outputs, which is crucial for applications requiring diverse and realistic image generation. The framework is versatile and can be applied to a wide range of image translation tasks, making it a valuable contribution to the field.

h) Demerits: The complexity of the model might make it difficult to reproduce and apply in practice, particularly for those without extensive experience in deep learning. While

979-8-3503-6052-3/24/$31.00 © 2024 IEEE

the paper demonstrates impressive results, it lacks a detailed exploration of the limitations and potential failure cases of the proposed method. [5]This paper introduces an improved precision and recall metric for assessing the performance of generative models. The authors argue that traditional evaluation metrics are insufficient and propose a new approach that better captures the quality of generated samples.

i) Merits: The proposed metric provides a more nuanced and accurate assessment of generative models, addressing some of the shortcomings of existing evaluation techniques. It offers a practical tool for researchers and practitioners to better evaluate and compare different generative models, leading to more reliable conclusions about their performance.

j) Demerits: The paper introduces a new metric but does not provide extensive empirical validation across a wide range of generative models, which would strengthen the case for its adoption. The metric, while innovative, may be challenging to integrate into existing workflows, especially for researchers accustomed to traditional evaluation methods.

[6]This paper introduces Least Squares Generative Adversarial Networks (LSGANs) to address the vanishing gradients problem in regular GANs. The authors demonstrate that LSGANs not only generate higher quality images but also improve stability during the training process.

k) Merits: Enhances image quality and training stability by using the least squares loss function. Provides empirical evidence showing the superiority of LSGANs over regular GANs on multiple datasets.

l) Demerits: The study is limited to simpler datasets, leaving the applicability to more complex datasets unexplored. Further integration with more advanced GAN architectures is required to fully realize the potential of LSGANs.

[7]This paper explores the use of conditional adversarial networks for image-to-image translation tasks. The authors propose a general-purpose approach that eliminates the need for hand-engineering loss functions, making it applicable across a wide range of tasks.

m) Merits: Successfully applies a single framework to various image-to-image translation tasks, demonstrating flexibility. Reduces the need for manual loss function design, streamlining the development process.

n) Demerits: Potential dependence on specific types of data, which might limit generalizability. May require significant computational resources, especially for highly complex tasks.

[8]This paper focuses on improving GAN performance by modifying consistency regularization. The authors propose several enhancements to the regularization procedure, resulting in better FID scores for both unconditional and conditional image synthesis.

o) Merits: Achieves improved FID scores across different GAN architectures, demonstrating the effectiveness of the proposed modifications. Applicable to both unconditional and conditional image synthesis, showcasing versatility.

p) Demerits: Consistency regularization can introduce artifacts, requiring careful application. The modifications need fine-tuning, adding to the complexity of the method.

[9]This paper investigates knowledge transfer in generative models, specifically GANs. The authors introduce MineGAN, a method that effectively transfers knowledge from pretrained GANs to new target domains, even with limited target data.

q) Merits: Enables effective knowledge transfer, improving fine-tuning efficiency with limited target domain data. Can leverage multiple pretrained GANs, enhancing the flexibility of the method.

r) Demerits: Requires access to pretrained GANs, which may not always be available. Managing knowledge transfer from multiple GANs adds complexity to the implementation.

[10]This paper proposes a method to enhance GAN discriminators by leveraging self-supervised learning and transfer learning. The approach is designed to improve the generation of complex scene images by boosting the discriminator's multi-scale representations.

s) Merits: Significantly improves the generation performance of GANs on complex scenes. Utilizes self-supervised and transfer learning techniques to enhance discriminator effectiveness.

t) Demerits: Increases computational complexity due to the integration of multiple expert models. The approach is still evolving, with challenges in fully combining the proposed methods.

III. Novelty

- Unlike traditional GANs requiring a separate model for each domain pair, this project employs a single model for multiple domains, reducing computational overhead and enhancing scalability.

IV. Methodology

A. Problem Definition

The goal of this project is to train a Conditional Generative Adversarial Network (cGAN) for generating images of celebrities from the CelebA dataset, conditioned on specific attributes like "Black Hair", "Blond Hair", "Male", etc. The project uses a deep learning approach to generate images that match the target attributes, leveraging GANs to produce high-quality, diverse images.

B. Dataset

The **CelebA dataset** is used, which contains 202,599 images of celebrities, each labeled with 40 different attributes. For this project, we focus on five selected attributes:

- Hair color
- Gender
- Age

These attributes allow the model to condition the generated images on specific characteristics, enabling control over the generated output.

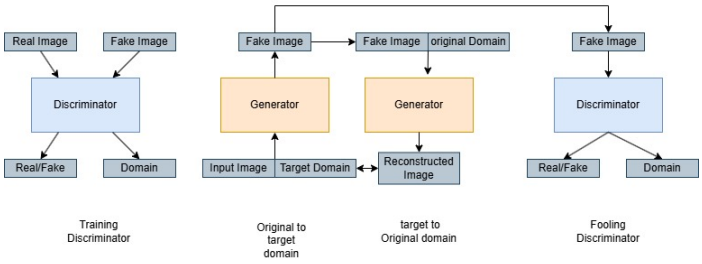

Fig. 1. Block Diagram

C. Data Preprocessing

To prepare the images for the GAN model:

- **Resize**: The images are resized to a fixed height and width of 128x128 pixels.
- **Random Crop**: Random cropping is applied to each image to introduce variability during training.
- **Normalization**: Each image is normalized to the range [-1, 1] for compatibility with GAN training.
- **Data Augmentation**: The dataset is augmented with random horizontal flips to improve the generalization of the model.

The data is split into training and validation sets, with images loaded in batches using PyTorch's `DataLoader` class. The training images are transformed according to the specified transformations. The block diagram is shown in fig. 1. Here the discriminator and a generator is present. discriminator learns to distinguish between real and fake images and classify the real images to its corresponding domain. Generator takes in as input both the image and target domain label and generates a fake image. The target domain label is spatially replicated and concatenated with the input image. Generator tries to reconstruct the original image from the fake image given the original domain label.Next the generator tries to generate images indistinguishable from real images and classifiable as target domain by discriminator.

D. Model Architecture

1) Generator: The **Generator** network is a deep convolutional neural network (CNN) that generates new images based on a noise vector and the target attributes. It uses:

- **Residual Blocks** to ensure stable training.
- The network takes a latent vector (noise) and condition information (selected attributes) as input, generating a high-quality image.

2) Discriminator: The **Discriminator** network distinguishes between real and generated images, conditioned on the target attributes. It uses:

- **Convolutional layers** to classify whether an image is real or fake.

The discriminator is trained with the binary cross-entropy loss, encouraging it to correctly classify images.

3) Cycle Consistency Loss: Since we use cycle-consistency (common in image-to-image translation tasks), we add an additional loss function, **L1 loss**, which ensures that when an image is generated and then reconstructed, it closely matches the original input image. This is particularly useful when the model is conditioned on attributes and needs to maintain the original image content.

4) Gradient Penalty (WGAN-GP): To stabilize training, a **gradient penalty** is applied to the discriminator, penalizing large gradients during backpropagation, which is a characteristic of WGAN-GP (Wasserstein GAN with Gradient Penalty).

E. Training Process

The model is trained over multiple epochs using the following procedure:

- **Adversarial Training**: The generator and discriminator are trained alternately. The generator tries to fool the discriminator by generating realistic images, while the discriminator learns to distinguish real images from generated ones.
- **Optimization**: Adam optimizers are used for both the generator and discriminator with a learning rate of 0.0002, and betas (0.5, 0.999).
- **Loss Functions**:
  - Binary Cross-Entropy for adversarial loss (discriminator).
  - L1 loss for cycle consistency.
  - WGAN-GP loss for regularizing the discriminator.

The models are updated after every batch, and the generator is trained less frequently than the discriminator (specified by `n_critic`), which helps maintain a balance between the two models during training.

F. Evaluation

- **Sample Images**: Periodically, sample images are generated and saved to visually assess the progress of the model.
- **Model Checkpoints**: The model is periodically saved to avoid losing progress. This allows us to resume training from a previous checkpoint if needed.

The evaluation of the model is done by comparing the generated images with real images from the validation dataset. Since this is a conditional GAN, we assess how well the generator captures the target attributes by visually inspecting the images.

G. Hyperparameters

The following key hyperparameters are defined:

- **Learning rate**: 0.0002
- **Batch size**: 16
- **Lambda weights** for loss functions:
  - Lambda for classification loss: 1
  - Lambda for reconstruction loss: 10
  - Lambda for gradient penalty: 10

V. Results

After training the model for a specified number of epochs, we expect the generator to produce high-quality images that accurately reflect the target attributes. The generated images are evaluated. Results are shown in fig. 2, 3.

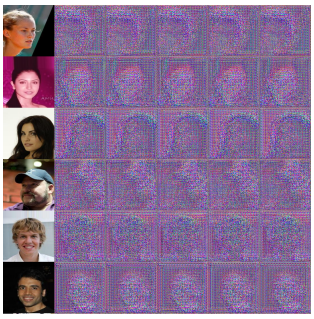

Fig. 2. output for 1st epoch

Fig. 3. output

A. Comparision

The model offers significant advantages over DIAT and CycleGAN in multi-domain image translation. Unlike these models, which require multiple generator-discriminator pairs for each domain pair, StarGAN uses a single pair for all domains, reducing parameter count and enhancing training efficiency. Additionally, StarGAN's capability for multi-attribute control within a single transformation simplifies complex edits, enhancing both visual quality and application range. Together, these architectural strengths make it a more versatile, efficient, and scalable solution for multi-domain image tasks.

VI. conclusion

on training a Conditional GAN on the CelebA dataset to generate images conditioned on specific attributes. The use of advanced GAN techniques like WGAN-GP and cycle consistency ensures that the model produces high-quality, realistic images while maintaining the necessary control over the generated attributes.

References

[1] A. Casanova, M. Drozdzal, and A. Romero-Soriano, "Generating Unseen Complex Scenes: Are We There Yet?" (2020)

[2] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object Detection in 20 Years: A Survey" (2019)

[3] F. Lateef and Y. Ruichek, "Survey on Semantic Segmentation Using Deep Learning Techniques," Neurocomputing (2019)

[4] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal Unsupervised Image-to-Image Translation," ECCV (2018)

[5] T. Kynkäänniemi, T. Karras, S. Laine, J. Lehtinen, and T. Aila, "Improved Precision and Recall Metric for Assessing Generative Models," NeurIPS (2019)

[6] Improving Complex Scene Generation by Enhancing Multi-Scale Representations of GAN Discriminators

[7] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks,"inProc.IEEEInt.Conf.Comput. Vis. (ICCV), Oct. 2017, pp. 2794–2802.

[8] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 1125–1134.

[9] Z. Zhao, S. Singh, H. Lee, Z. Zhang, A. Odena, and H. Zhang, "Improved consistency regularization for GANs," in Proc. AAAI Conf. Artif. Intell., vol. 35, no. 12, 2021, pp. 11033–11041.

[10] Y. Wang, A. Gonzalez-Garcia, D. Berga, L. Herranz, F. S. Khan, and J. Van De Weijer, "MineGAN: Effective knowledge transfer from GANs to target domains with few images," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 9332–9341.