Comparison of Image to Image Translation Techniques - InstaGAN and CycleGAN

Abhijeet Manoj Pal (200100107)
Electrical Engineering
IIT BOMBAY
Mumbai, India
200100107@iitb.ac.in

Nirmal S (20d070057)
Electrical Engineering
IIT BOMBAY
Mumbai, India
20d070057@iitb.ac.in

Kanak Yadav (20d070044)
Electrical Engineering
IIT BOMBAY
Mumbai, India
20d070044@iitb.ac.in

***Abstract*—Unsupervised image-to-image translation has been a prominent area of research, with models like CycleGAN demonstrating efficacy in various tasks. In this study, we compare the performance of CycleGAN and InstaGAN, particularly focusing on the challenging horse-to-zebra translation task using the widely adopted horse2zebra dataset. CycleGAN, known for its cycle-consistency mechanism, has been a benchmark in unpaired image translation. In contrast, InstaGAN introduces instance-awareness and context-preserving mechanisms, claiming superior results, especially in scenarios involving multiple instances and shape changes. Through extensive experiments on the horse2zebra dataset, we evaluate the capabilities of both models, considering factors such as image quality, preservation of individual instances, and overall translation robustness. Our findings shed light on the strengths and limitations of CycleGAN and InstaGAN in the context of unpaired image translation tasks.**

***Index Terms*—InstaGAN, CycleGAN, image translation, horse2zebra, translation, photographic translation**

I. Introduction

The field of image-to-image translation has witnessed significant advancements in recent years, driven by the emergence of powerful generative models. These models play a crucial role in tasks such as style transfer, domain adaptation, and object transfiguration. Among the notable approaches, InstaGAN and CycleGAN have demonstrated remarkable capabilities in transforming images from one domain to another.

Image-to-image translation is a fundamental problem in computer vision, with applications ranging from artistic content generation to practical tasks like image enhancement and domain adaptation. The choice of an appropriate model for a given task is crucial, as different models may exhibit varying performance characteristics, particularly in challenging scenarios involving multiple instances and shape changes.

This paper presents a comprehensive comparative analysis of three prominent image-to-image translation models: InstaGAN and CycleGAN. The primary focus is on evaluating their performance in scenarios where complex transformations, such as converting shirts to tshirts in fashion images or converting horse to zebra and vice versa are required in photography. By addressing the limitations of previous methods, InstaGAN introduces instance-awareness to enhance multi-instance transfiguration. We illustrate the pros and cons of each of these methods and where InstaGAN outperforms the other methods.

In the following sections, we delve into the architectures, methodologies, and unique features of each model. We conduct extensive experiments on horse2zebra dataset, emphasizing challenging cases that showcase the strengths and weaknesses of each approach. Through this comparative study, we aim to provide insights into the suitability of these models for specific image translation tasks and contribute to the understanding of their capabilities and limitations.

Contribution:
Project concept: Abhijeet Pal
Data Collection: Abhijeet Pal & Kanak Yadav
Coding : Kanak Yadav & Nirmal S
Report Writing : Abhijeet Pal & Nirmal S
Video Making : Nirmal S & Kanak Yadav

II. Background and Prior Work

The horse2zebra Kaggle dataset has been a widely used benchmark for evaluating image-to-image translation models, prominently featuring in CycleGAN implementations. We have also used Coco dataset for obtaining image segmented images. CycleGAN, proposed by Zhu et al., is renowned for its ability to learn mappings between unpaired image domains through the introduction of cycle-consistency loss. While CycleGAN has shown success in various applications, recent advancements in InstaGAN, as proposed by Mo et al., claim to outperform CycleGAN, particularly in scenarios involving multiple target instances and shape changes. InstaGAN's architectural innovations, including the incorporation of instance information and a context-preserving loss, contribute to its superior performance by addressing limitations observed in CycleGAN. Through our comparative analysis, we aim to delve into the specific architectural differences that make InstaGAN more effective in scenarios such as converting horse to zebra and vice versa. StyleGAN and Pix2Pix are influential models in generative image synthesis. StyleGAN introduces a novel generator architecture that operates on a disentangled latent space. The generator is composed of multiple mapping and synthesis layers, where the mapping layers transform a latent code into a style code, and the synthesis layers generate images from these styles. StyleGAN's contribution lies in its ability to separate high-level semantic content and stochastic

variations, providing fine-grained control over image synthesis. Mathematically, the generator can be represented as $G(z, w) = \text{Synthesis}(\text{MLP}(z), w)$, where z is the input latent code, w is the style code, MLP denotes the mapping layers, and Synthesis represents the synthesis layers.

Pix2Pix, on the other hand focuses on paired image-to-image translation tasks using conditional adversarial networks. The architecture consists of a generator G and a discriminator D in a conditional GAN setup. The generator takes an input image from one domain and transforms it into a corresponding output image in another domain. The adversarial loss encourages the generator to produce realistic-looking images, while a pixel-wise L1 loss enforces pixel-level similarity between the generated and ground truth images. Mathematically, the objective function for Pix2Pix is given by $\mathcal{L}_{\text{Pix2Pix}}(G, D) = E_{x,y}[\log D(x, y)] + \lambda E_{x,y,\hat{y}}[\|y - \hat{y}\|_1] - \lambda E_{x,\hat{y}}[\log D(x, \hat{y})]$, where x is the input image, y is the ground truth, $\hat{y}$ is the generated output, and λ is a hyperparameter controlling the balance between adversarial and L1 loss terms.

III. Approach

The horse2zebra dataset consist of 1187 Horse and 1474 Zebra Images. Images in both the domains are split into train and test subsets. CycleGAN consists of two generators, $G : X \to Y$ and $F : Y \to X$, where X and Y represent the input and output domains. The generators are responsible for transforming images from one domain to another, incorporating a cycle-consistency mechanism that ensures reversibility. Two discriminators, D_X and D_Y, distinguish between real and generated images in their respective domains. Adversarial training is employed, where generators aim to produce realistic images to deceive discriminators, while discriminators strive to distinguish between real and generated images. The overall objective involves minimizing adversarial losses and enforcing cycle-consistency constraints.

InstaGAN distinguishes itself from CycleGAN through key architectural differences designed to address limitations in handling multiple target instances and shape changes. Unlike CycleGAN, InstaGAN incorporates instance information, such as object segmentation masks, facilitating effective multi-instance transfiguration. To maintain the identity function outside of target instances, InstaGAN introduces a context-preserving loss, ensuring non-target regions remain unchanged during translation. Additionally, InstaGAN proposes a sequential mini-batch inference/training technique to handle multiple instances efficiently, addressing challenges posed by limited GPU memory.

The two different architectures were implemented on the horse2zebra dataset and Coco Dataset. The results obtained by both of these models were compared based on visual inspection (quality) and metrics.

CycleGAN employs four distinct loss functions for training its generators and discriminators. The adversarial loss ($\mathcal{L}_{\text{GAN}}(G, D_Y, X, Y) = E_{y \sim p_{\text{data}}(y)}[\log D_Y(y)] + E_{x \sim p_{\text{data}}(x)}[\log(1 - D_Y(G(x)))]$) encourages the generator to produce images that can deceive the discriminator, promoting realism in the generated images. The cycle-consistency loss ($\mathcal{L}_{\text{cyc}}(G, F) = E_{x \sim p_{\text{data}}(x)}[\|F(G(x)) - x\|_1] + E_{y \sim p_{\text{data}}(y)}[\|G(F(y)) - y\|_1]$) enforces a one-to-one mapping between domains, ensuring that translating an image from one domain to another and back results in the original image. The identity loss ($\mathcal{L}_{\text{idt}}(G, F) = E_{y \sim p_{\text{data}}(y)}[\|F(y) - y\|_1] + E_{x \sim p_{\text{data}}(x)}[\|G(x) - x\|_1]$) preserves the content of an image within the same domain, maintaining the original characteristics. The total generator loss ($\mathcal{L}_{\text{total}}(G, F, D_X, D_Y, X, Y) = \mathcal{L}_{\text{GAN}} + \lambda \cdot \mathcal{L}_{\text{cyc}} + \lambda_{\text{idt}} \cdot \mathcal{L}_{\text{idt}}$) combines the adversarial, cycle-consistency, and identity losses, providing the overall objective for training the generators.

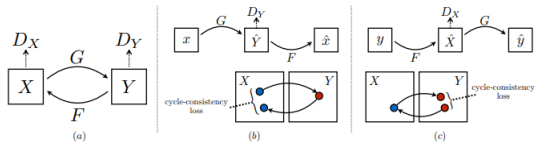

Fig. 1. CycleGAN Architecture

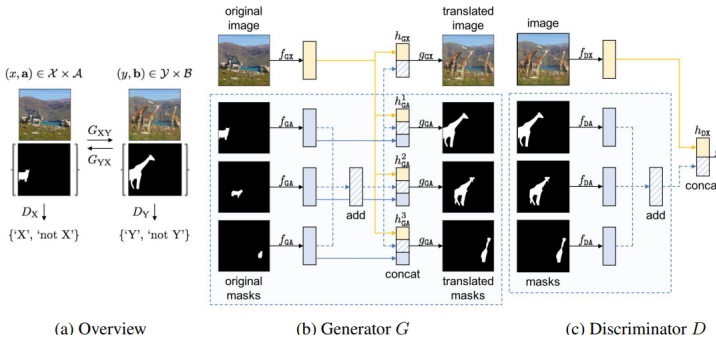

Fig. 2. InstaGAN Architecture

In InstaGAN, the loss function diverges from CycleGAN by introducing a context-preserving loss ($\mathcal{L}_{\text{context}}$) to ensure the preservation of the identity function outside of target instances. This additional term is incorporated to discourage distortion in non-target regions during image translation. Furthermore, InstaGAN includes specific terms to handle instance-aware translation, encouraging the model to translate both images ($\mathcal{L}_{\text{GAN}}$) and their corresponding sets of instance attributes ($\mathcal{L}_{\text{inst}}$). The instance-aware translation term maintains permutation invariance during the translation process. The modified loss function for InstaGAN can be expressed as:

$$\mathcal{L}_{\text{InstaGAN}} = \mathcal{L}_{\text{GAN}} + \lambda_{\text{context}} \cdot \mathcal{L}_{\text{context}} + \lambda_{\text{inst}} \cdot \mathcal{L}_{\text{inst}}$$

Here, λ_{context} and λ_{inst} are hyperparameters controlling the influence of the context-preserving and instance-aware terms, respectively. These modifications contribute to InstaGAN's claimed superiority, particularly in scenarios involving complex transformations and multiple instances.

IV. Experiments and Results

The architectures of CycleGAN and InstaGAN were run for multiple epochs. The images of the horse2zebra and Coco

dataset were normalized and standardized with 0 mean and 1 standard deviation. The results shown below are obtained after 50 epochs. The optimizer used is Adam. The results obtained by the models are as shown below. The results

Fig. 3. CycleGAN Results — Left to right : Original Horse, Fake Zebra, Original Zebra, Fake Horse

obtained by InstaGAN on the same dataset 'horse2zebra' are shown in fig 4. We observe that the quality of prediction by InstaGAN is much better compared to CycleGAN. Zebra to horse translation is shown in fig 5

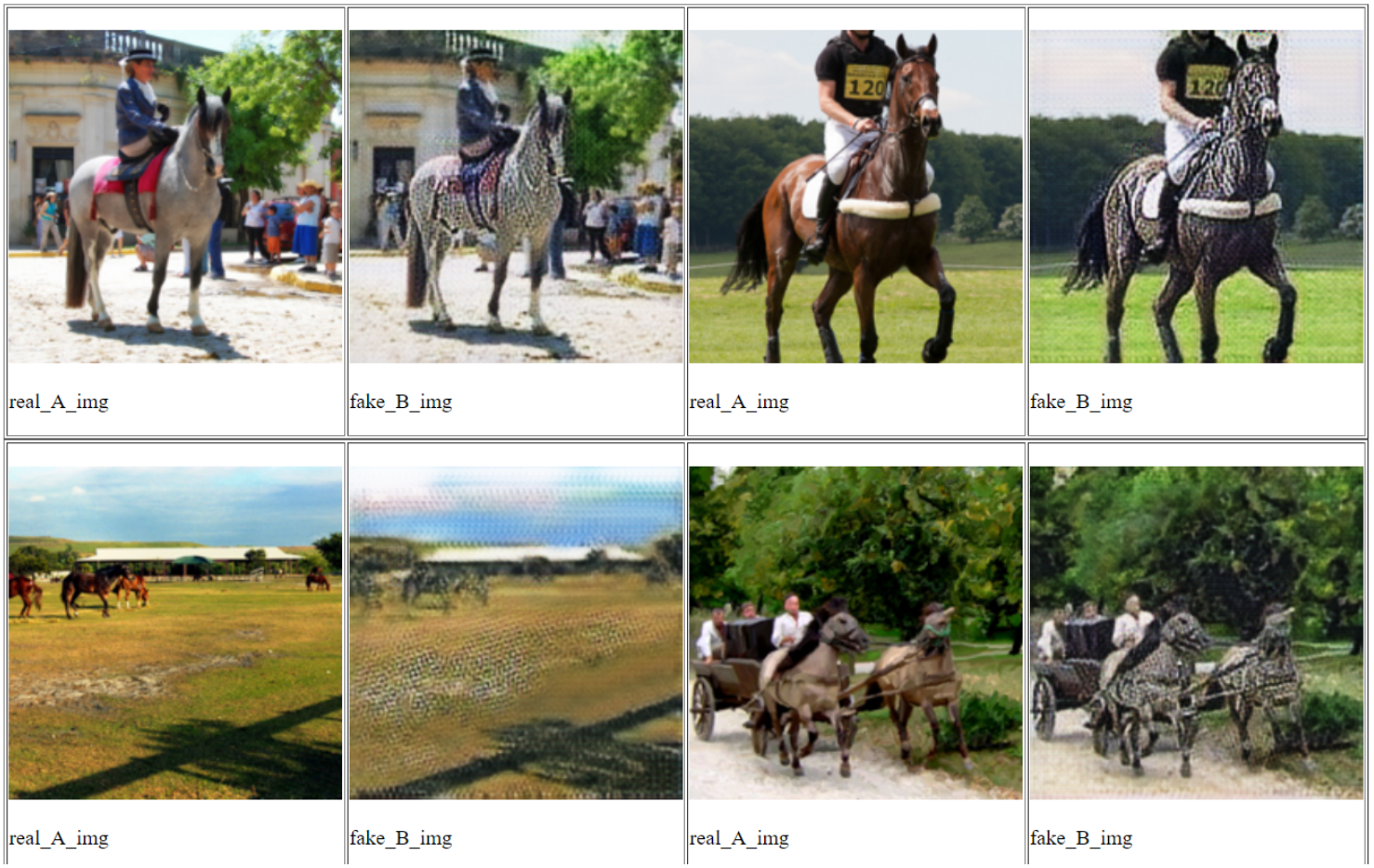

Fig. 4. Horse to Zebra Translaton by InstaGAN

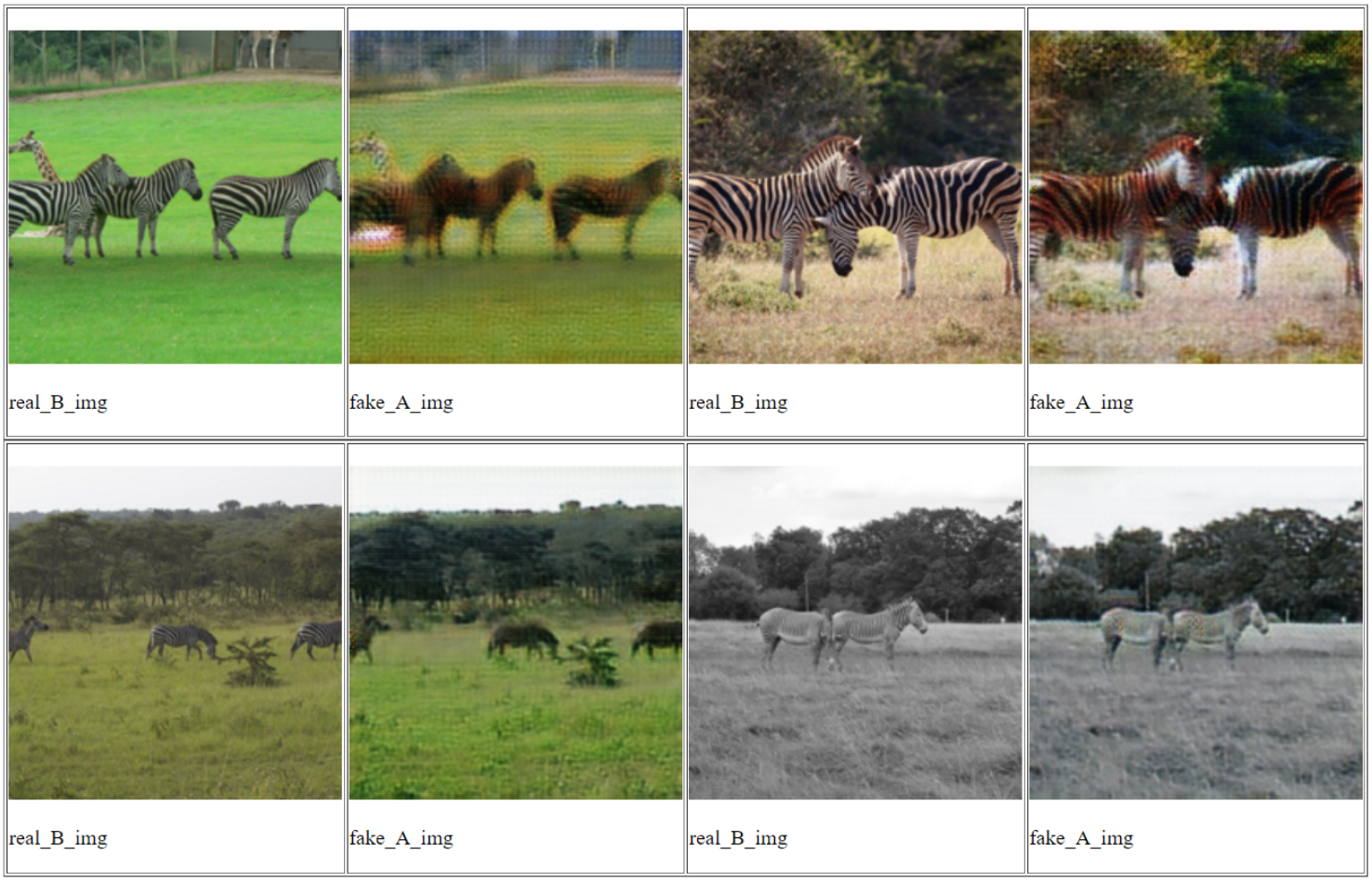

Fig. 5. Zebra to Horse Translation by InstaGAN

V. Discussion and Conclusions

In this experiment, we find that InstaGAN gives much better results in image to image translation on horse to zebra and zebra to horse. This can be mainly attributed to its incorporation of instance information such as image segmentation masks, which enables the model to perform multi-instance transfiguration effectively. It also introduces a context-preserving loss, helping the model to work much better compared to CycleGAN.

VI. Key Links

Github Repo :- Link
Demo Video :- Link

References

[1] Mo. Sangwoo, Cho Minsu, Shin Jinwoo 'InstaGAN: Instance-Aware image-to-image Translation'
[2] Zhu. Jun-Yan, Park Taesung, Isola Philip, Efros Alexei 'Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks'
[3] Horse2Zebra Dataset Kaggle
[4] CycleGAN Keras Blog
[5] Tensorflow CycleGAN tutorial