AVIAN ARTISTRY

CREATING IMAGES OF BIRDS FROM TEXTUAL DESCRIPTIONS

GROUP-A1

Akshara Nair (MT22008)

Aryan Chaudhary (MT22019)

Medha (MT22110)

Dept. of CSE, IIITD

INTRODUCTION

- This project aims to generate bird images from textual descriptions using GANs, addressing the limitations of traditional GANs and improving the efficiency and stability of the model.
- The project begins by reviewing the state-of-the-art GAN architectures and examining their limitations.
- The project proposes new techniques for generating images of birds, on the basis of given text prompt as input.
- The results are being tested on a benchmark dataset of bird images on the basis of evaluation metrics.
- Diverse applications as wildlife conservation, virtual reality gaming and animation.

DATASET

- The dataset used in this project is the **CUB-200-2011** dataset, which contains images of **200** bird species.
- It consists of a total of **11,788** images, with each image containing a single bird.
- Along with the images, the dataset contains textual descriptions of the birds, such as the common name, scientific name, and a brief description of the bird's appearance and behavior.
- The textual descriptions are provided in a structured format, with each bird having **10 attribute labels** describing its appearance and behavior, such as "has a long beak" or "perches on branches".

BASELINE

BASELINE- 1

DC-GAN :- Deep Convolutional Generative Adversarial Networks

DC-GAN conditioned on text features encoded by a hybrid character-level Convolutional RNN. Both the **generator** and the **discriminator** network perform feed-forward inference conditioned on the text feature.

BASELINE- 2

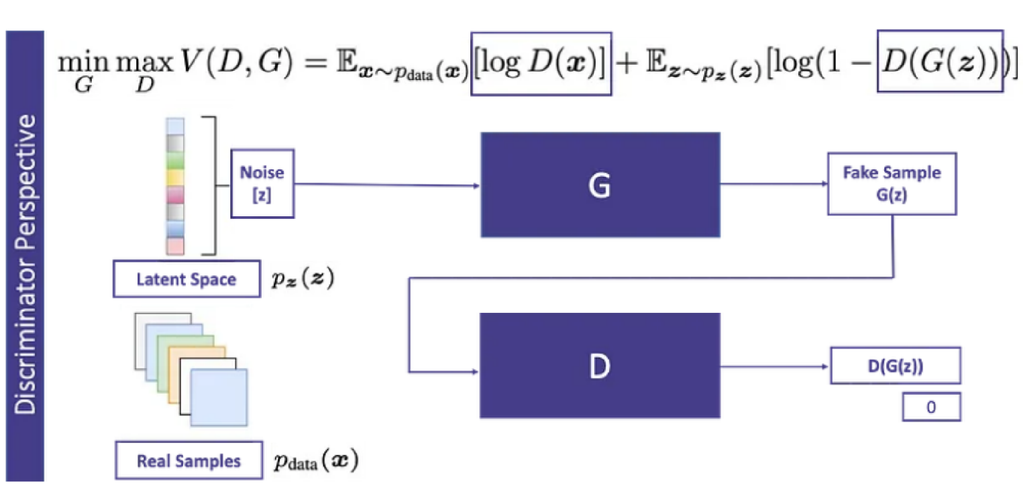
StackGANs - Stacked Generative Adversarial Networks

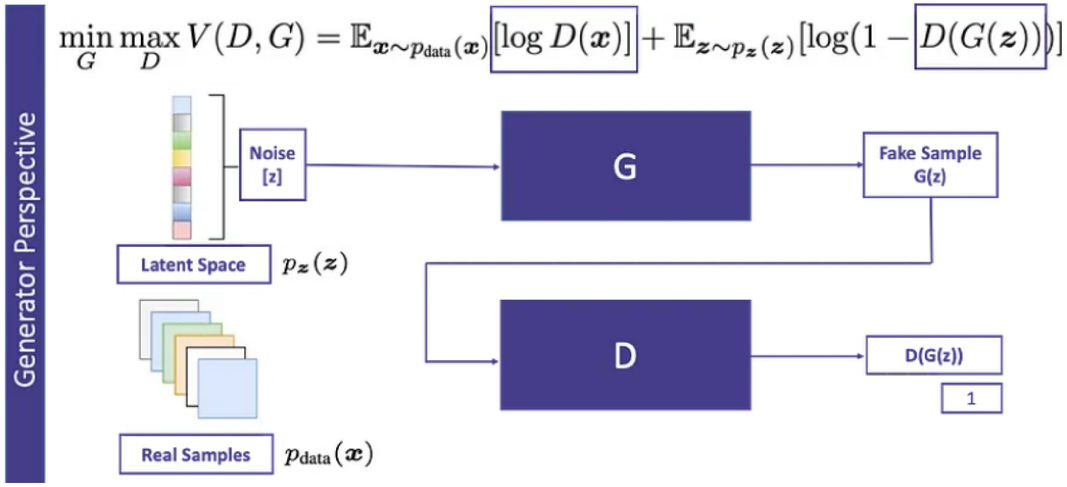
StackGAN has a two-stage architecture, where the first stage generates a low-resolution image from a textual description and the second stage refines the low-resolution image to a high-resolution image.

BASELINE-1

DC-GAN for Text to Image Synthesis

(Objective Function)

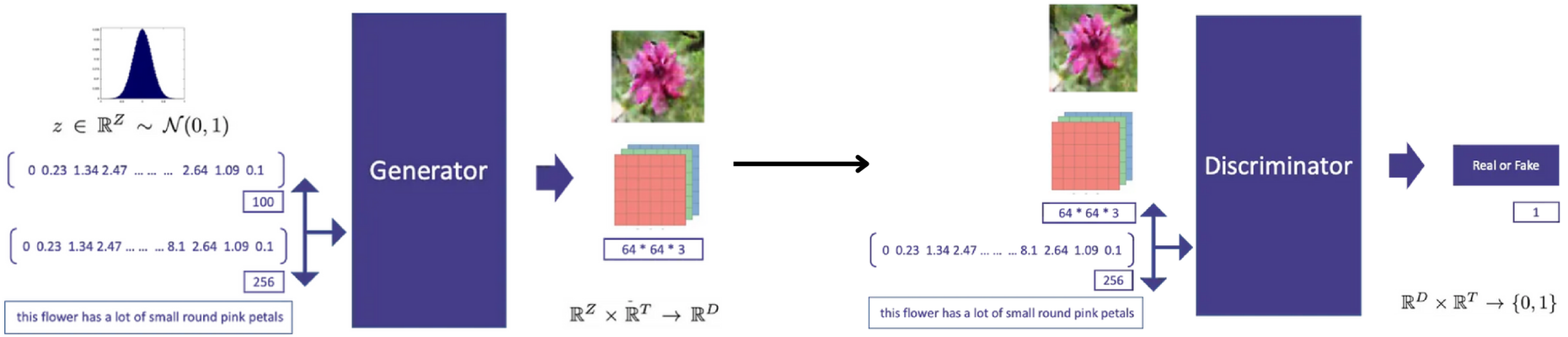




Discriminator wants to drive the likelihood of D(G(z)) to 0. Hence it wants to maximize (1-D(G(z))) whereas the Generator wants to force the likelihood of D(G(z)) to 1 so that Discriminator makes a mistake in calling out generated sample as real. Hence Generator wants to minimize (1-D(G(z)).

BASELINE-1

DC-GAN for Text to Image Synthesis



The following input(s) are used to realize the responsibility of the Generator to generate images which are real and aligned with the text -

- Pair of **(Real Image, Real Caption)** as input and target variable is set to 1
- Pair of **(Wrong Image, Real Caption)** as input and target variable is set to 0
- Pair of **(Fake Image, Real Caption)** as input and target variable is set to 0

(Fake Image, Real Caption) -> 0 by Discriminator, 1 for Generator Loss.

BASELINE-1

DC-GAN for Text to Image Synthesis

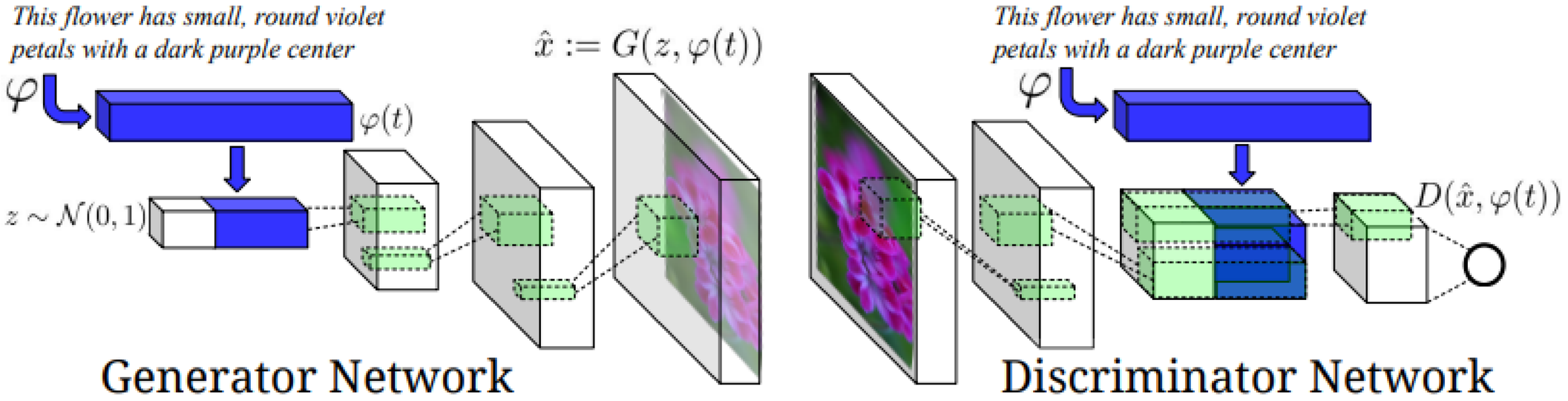


Figure 2. Our text-conditional convolutional GAN architecture. Text encoding $\varphi(t)$ is used by both generator and discriminator. It is projected to a lower-dimensions and depth concatenated with image feature maps for further stages of convolutional processing.

BASELINE-2
StackGAN

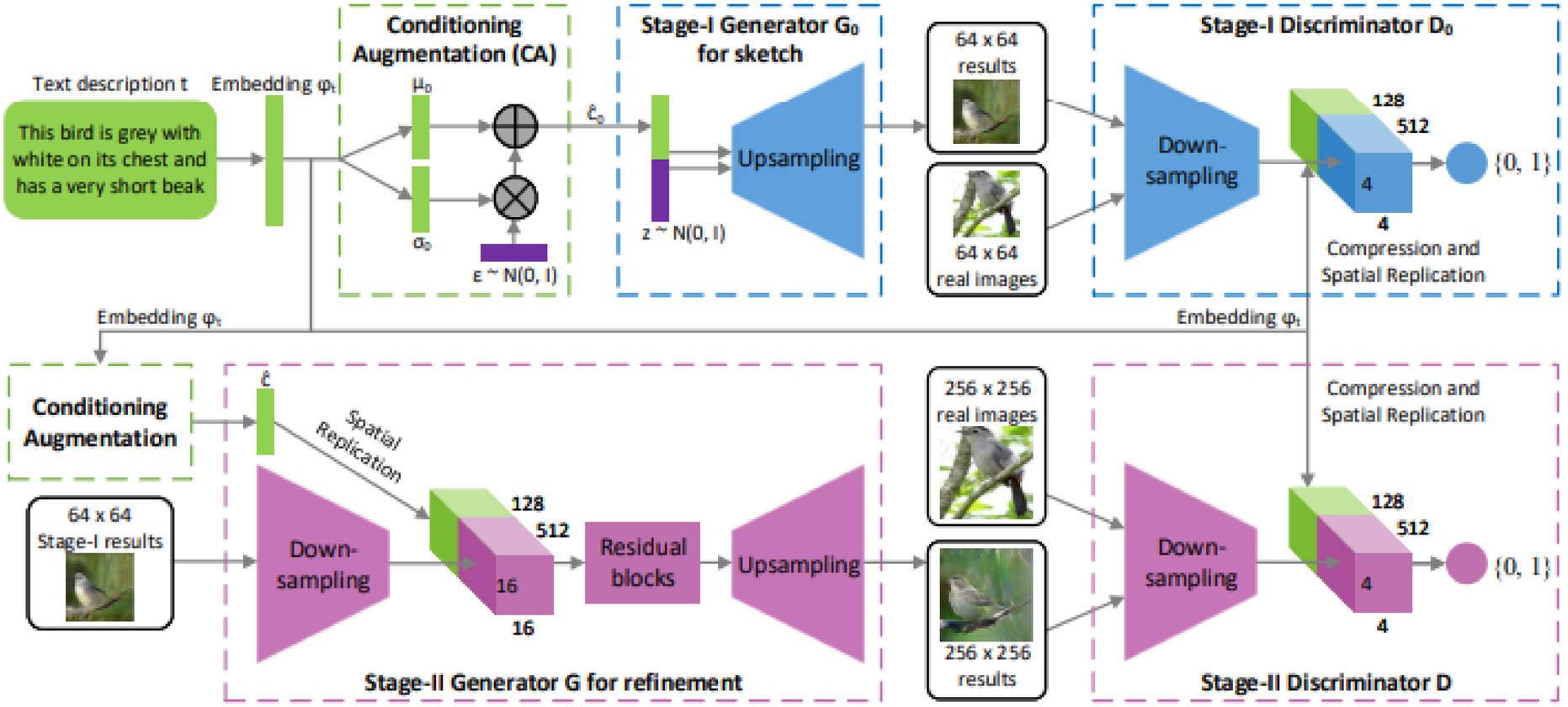


Figure 2. The architecture of the proposed StackGAN. The Stage-I generator draws a low-resolution image by sketching rough shape and basic colors of the object from the given text and painting the background from a random noise vector. Conditioned on Stage-I results, the Stage-II generator corrects defects and adds compelling details into Stage-I results, yielding a more realistic high-resolution image.

LIMITATIONS OF TRADITIONAL GANs

Two major limitaion of traditional conditional GANs are

- GANs are models with millions of parameters that require extensive computational resources, increases training instability and undermines replicability.
- Processing multi-channel data often loses intra - channel spatial relations.

PROPOSED NOVELTY

PROPOSAL-1

The use of **Quaternionic Generative Adversarial Networks (QGANs)** to generate high-quality color images from textual descriptions, which reduces the number of parameters needed for efficient image generation. Effectively reduces the value of FID

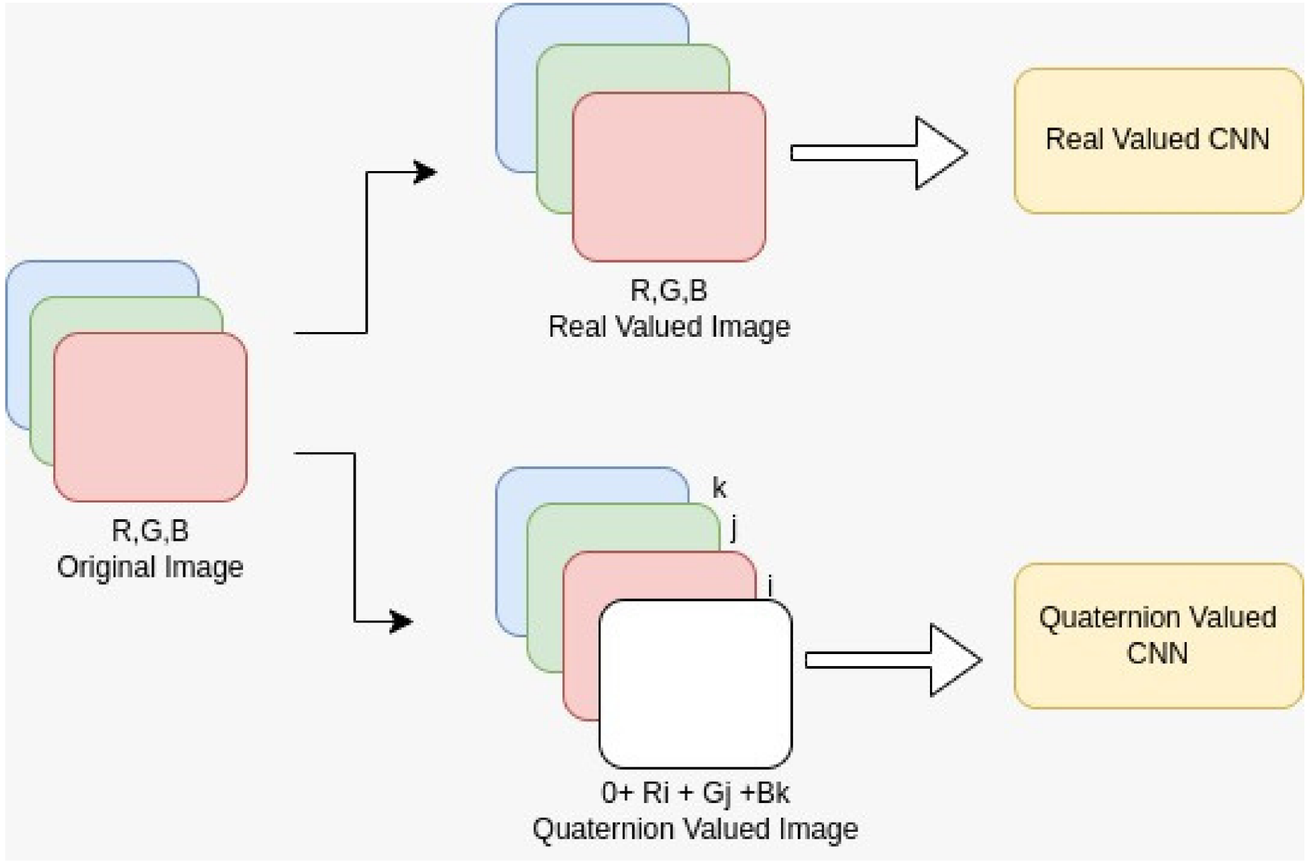
PROPOSAL-2

Spectral Normalization is a regularization technique that can be used with Generative Adversarial Networks (GANs) to improve their stability and prevent mode collapse.

PROPOSAL-3

Proposing the combination of **Quaternionic + Spectral Normalization** as a novel architecture to as **QSNGANs.**

Quaternion Convolutional Layers (QCNN)



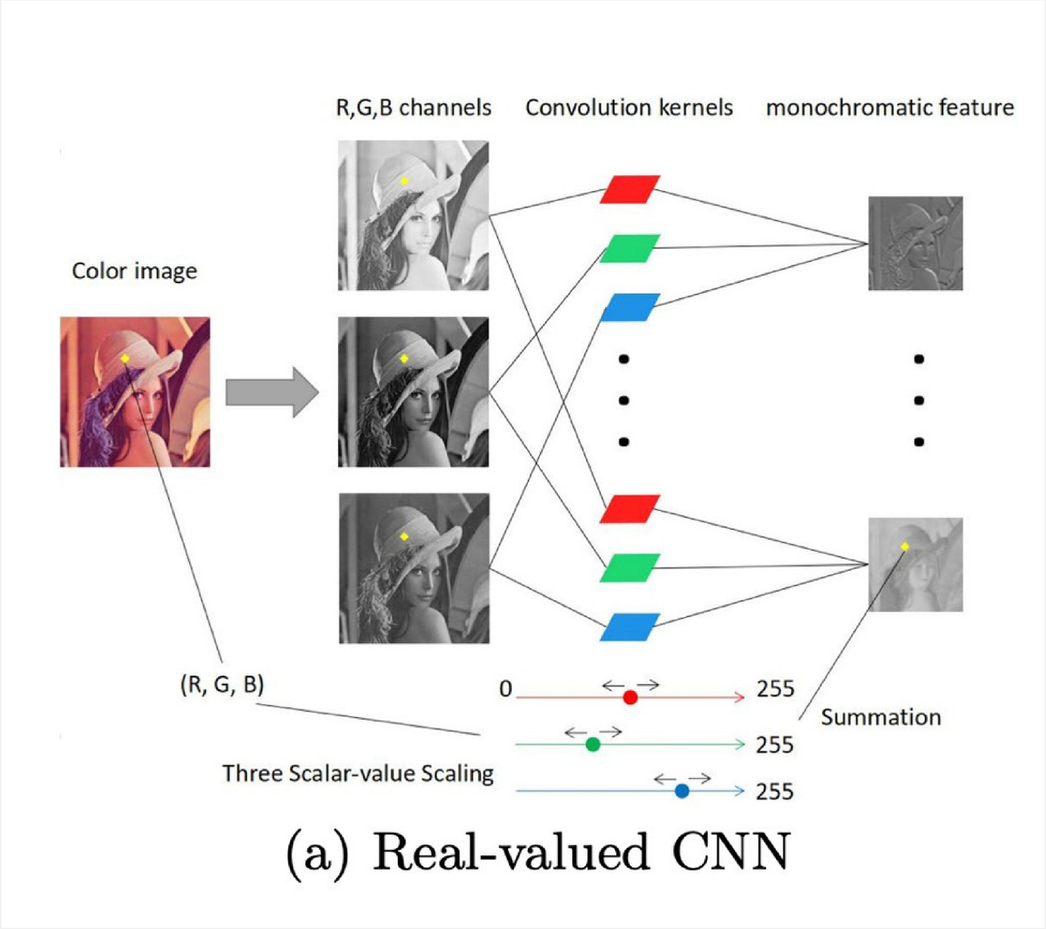
Quaternion Product, commonly known as Hamilton Product is described as below -

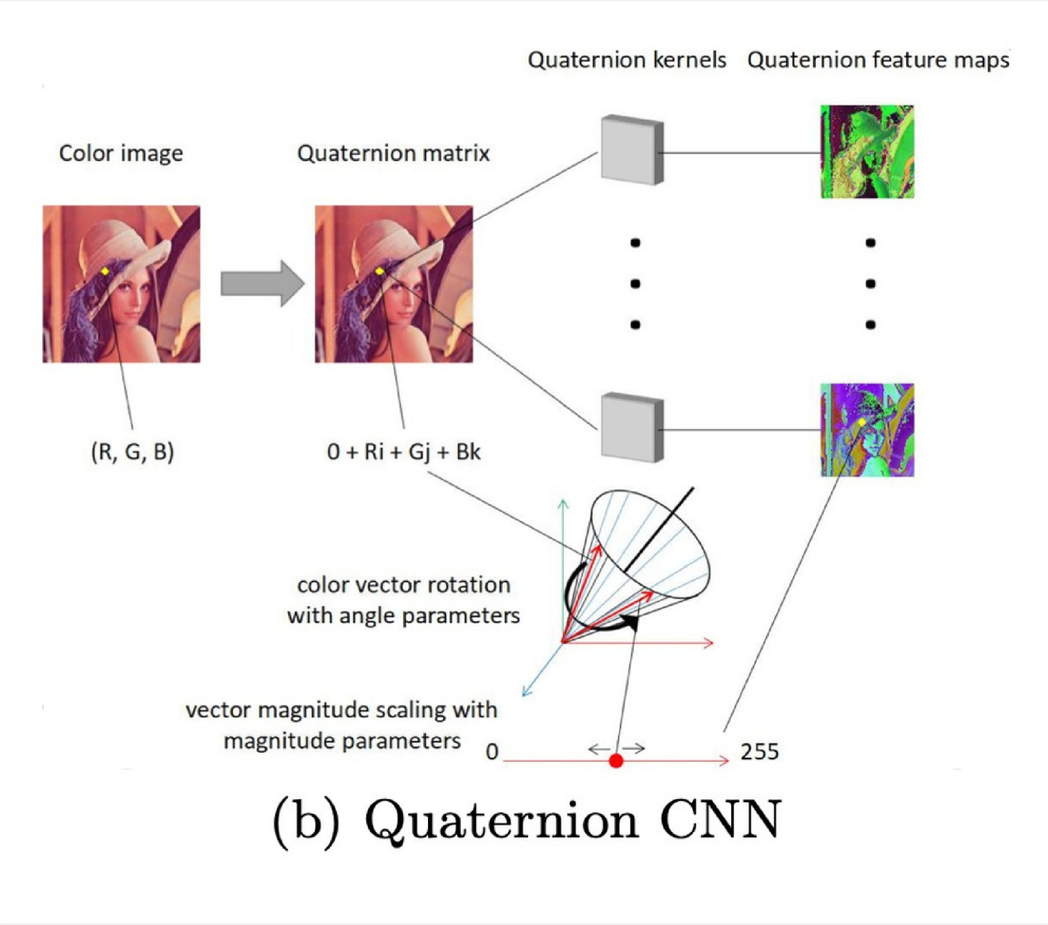
$$y = \phi(Wr * xr + br) \text{ - (1)}$$

$$\begin{aligned} W * x = {} & (W0 * x0 - W1 * x1 - W2 * x2 - W3 * x3) \\ & + (W1 * x0 + W0 * x1 - W3 * x2 + W2 * x3)\imath\hat{} \\ & + (W2 * x0 + W3 * x1 + W0 * x2 - W1 * x3)\jmath\hat{} \\ & + (W3 * x0 - W2 * x1 + W1 * x2 + W0 * x3)\kappa\hat{}. \end{aligned}$$

- (2)

Quaternion GANs(A2)



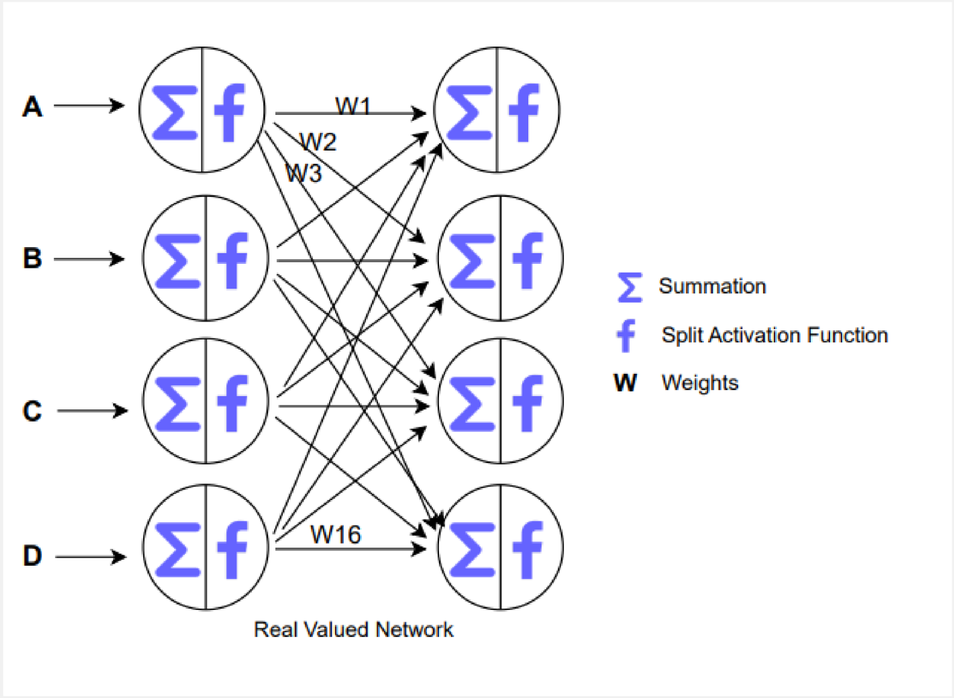


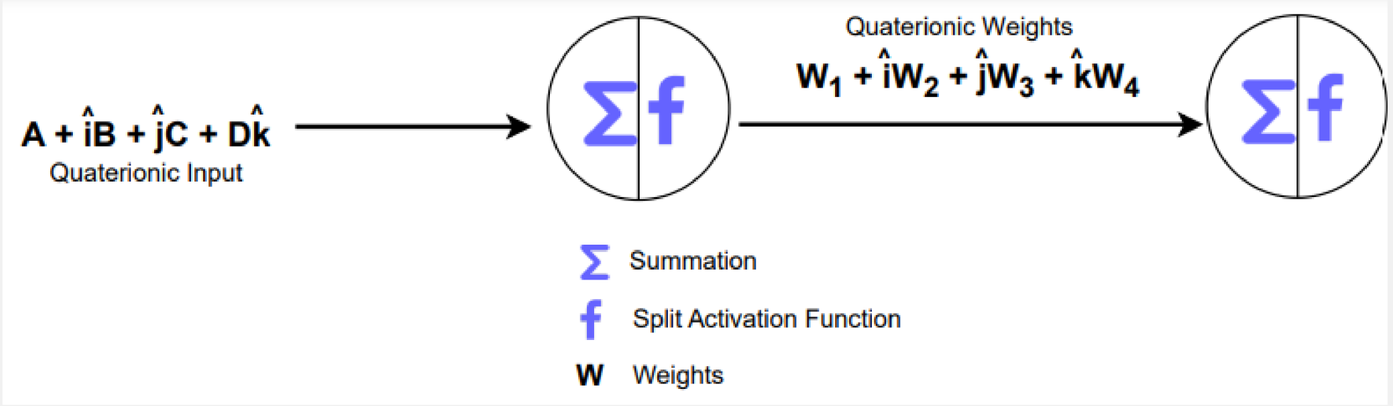
TRAINING QGANs

$$\frac{\delta \mathcal{L}}{\delta \mathbf{W}} = \frac{\delta \mathcal{L}}{\delta \mathbf{W}_0} + \frac{\delta \mathcal{L}}{\delta \mathbf{W}_1}\hat{\imath} + \frac{\delta \mathcal{L}}{\delta \mathbf{W}_2}\hat{\jmath} + \frac{\delta \mathcal{L}}{\delta \mathbf{W}_3}\hat{\kappa}.$$

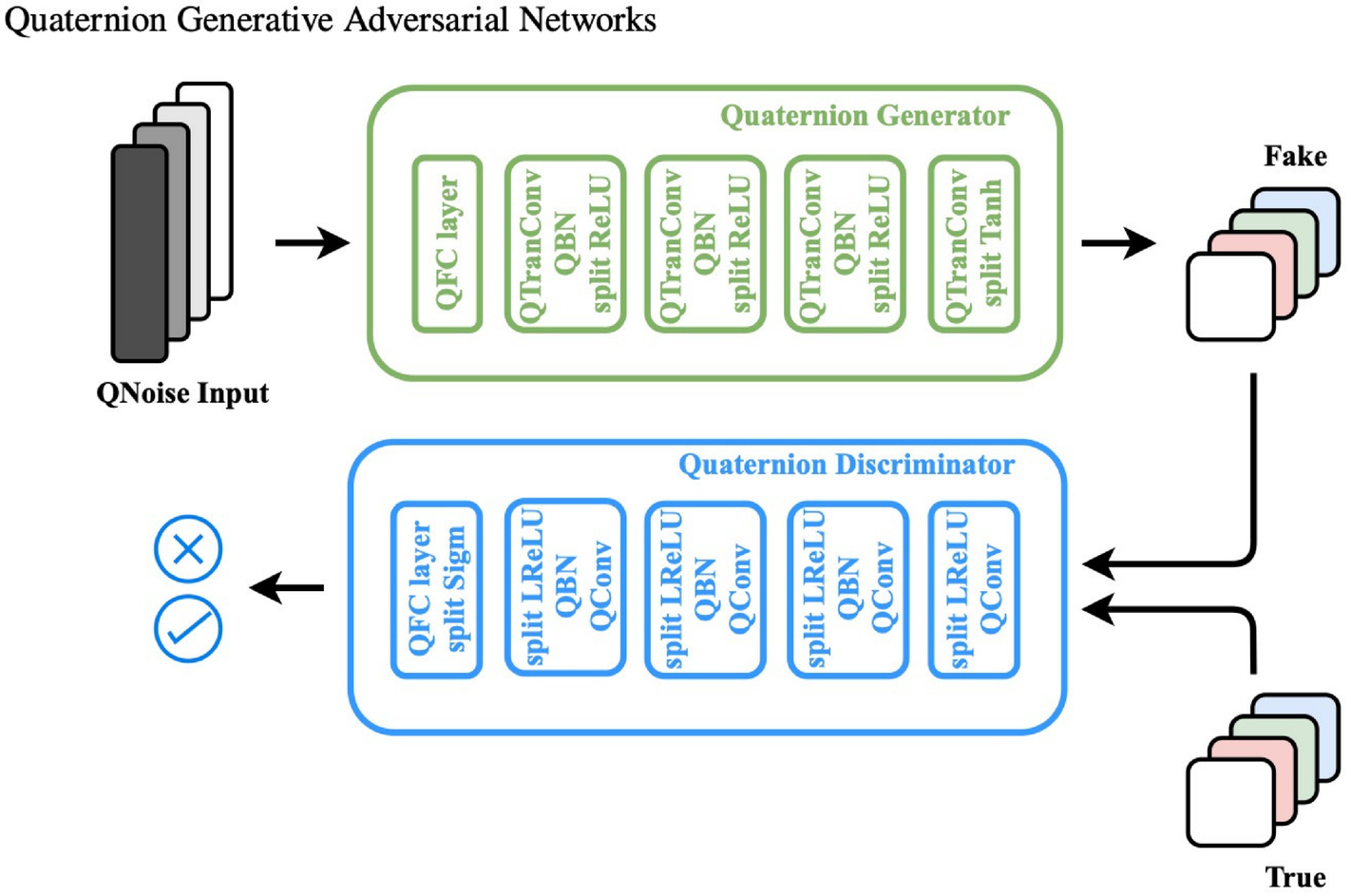
- The forward phase of a QNN is the same as its real-valued counterpart, Input flows from the first to the last layer of the network
- The gradient of a general quaternion loss function L is computed for each component of the quaternion weight matrix W using the above equation.
- This gradient is then propagated back through the network following the chain rule, allowing for the QNN to be trained via backpropagation, just like a real-valued neural network.

How Quaternion reduces parameters ?





Quaternion + GANs = QGANs



Spectral Normalised GANs(A3)

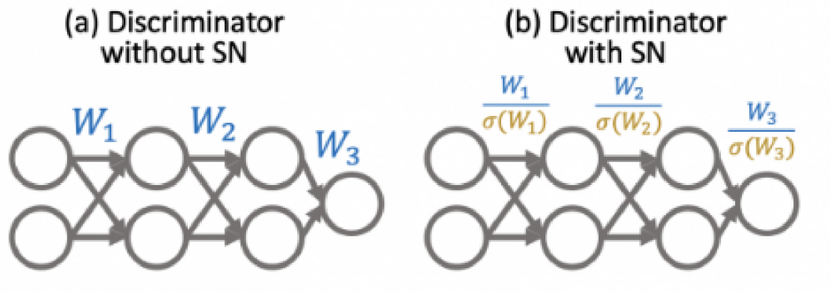


Figure 2: Spectral normalization divides the weights W_i by their spectral norms $\sigma(W_i)$ (i.e., the largest singular value of W_i).

$$\|\text{gradient}\|_{\text{Frobenius}} \leq \sqrt{\text{number of layers}} \cdot \|\text{input}\|.$$

$$\text{Var}(W) = (\text{fan-in of the layer})^{-1},$$

- Exploding and vanishing gradients describe a problem in which gradients either grow or shrink rapidly during training. Major reason for the instability of GANs.
- Limits the ability of weight tensors to amplify inputs in any direction. More precisely, when the **spectral norm of weights = 1** (as ensured by spectral normalization), and the activation functions are 1-Lipschitz (e.g., (Leaky)ReLU), the gradient is bound as in Eq above.
- Analogous to the concept introduced 1996, LeCun, Bottou, Orr, and Müller introduced a new initialization technique (commonly called **LeCun initialization** as in Eq.2)

Spectral Normalised GANs(A3)

- For Fully Connected Layer

$$W \in \mathbb{R}^{m \times n}$$

$$\text{Var}\left(\text{spectrally-normalized } W\right) \quad \text{is on the order of} \quad \left(\max\left\{m, n\right\}\right)^{-1}$$

$$\max\left\{m, n\right\} = m = n = \text{fan-in of the layer.}$$

- For Convolutional Layer

$$W \in \mathbb{R}^{c_{out} c_{in} k_w k_h}, \text{ where } c_{out}, c_{in}, k_w, k_h$$

$$\frac{W}{\sigma\left(W_{c_{out} \times \left(c_{in} k_w k_h\right)}\right)}$$

$\sigma\left(W_{c_{out} \times \left(c_{in} k_w k_h\right)}\right)$ is the spectral norm on the reshaped weight, $\max\left\{m, n\right\} = \max\left\{c_{out}, c_{in} k_w k_h\right\} = c_{in} k_w k_h = \text{fan-in of the layer.}$ Therefore, spectrally-normalized convolutional layers also maintain the same desired variances as LeCun initialization!

EVALUATION METRICS

HASH BASED METRICS

- Average Hashing (M1)
- Perceptual Hash or Discrete Cosine Transform Hashing (M2)
- Difference Hashing (D-HASH) (M3)

The hash value is calculated for the Original and Generated Images and the difference of Hash Value for two images, corresponding to a given text is averaged over the entire validation test.

FID SCORE

- Fréchet Inception Distance (FID)
- A lower FID score indicates that the generated images are closer to the real images in feature space and thus are of higher quality.
- The FID score is based on the distance between the distributions of the real images and the generated images in feature space, where the features are extracted from a pre-trained Inception-v3 network.
- M4 and M5 include the results for original validation and quaternionic domain based validation images.

RESULTS

RESULT-1

Models	M-1	M-2	M-3
A-1	31.265	31.316	31.661
A-2	31.858	31.349	31.742
(A-3)	32.425	31.228	31.957
A-4)	31.438	31.257	31.589

Table 1: Performance of the proposed novelties as compared to the Baseline on the basis of Hash Function Based Evaluation Metrics.

RESULT-2

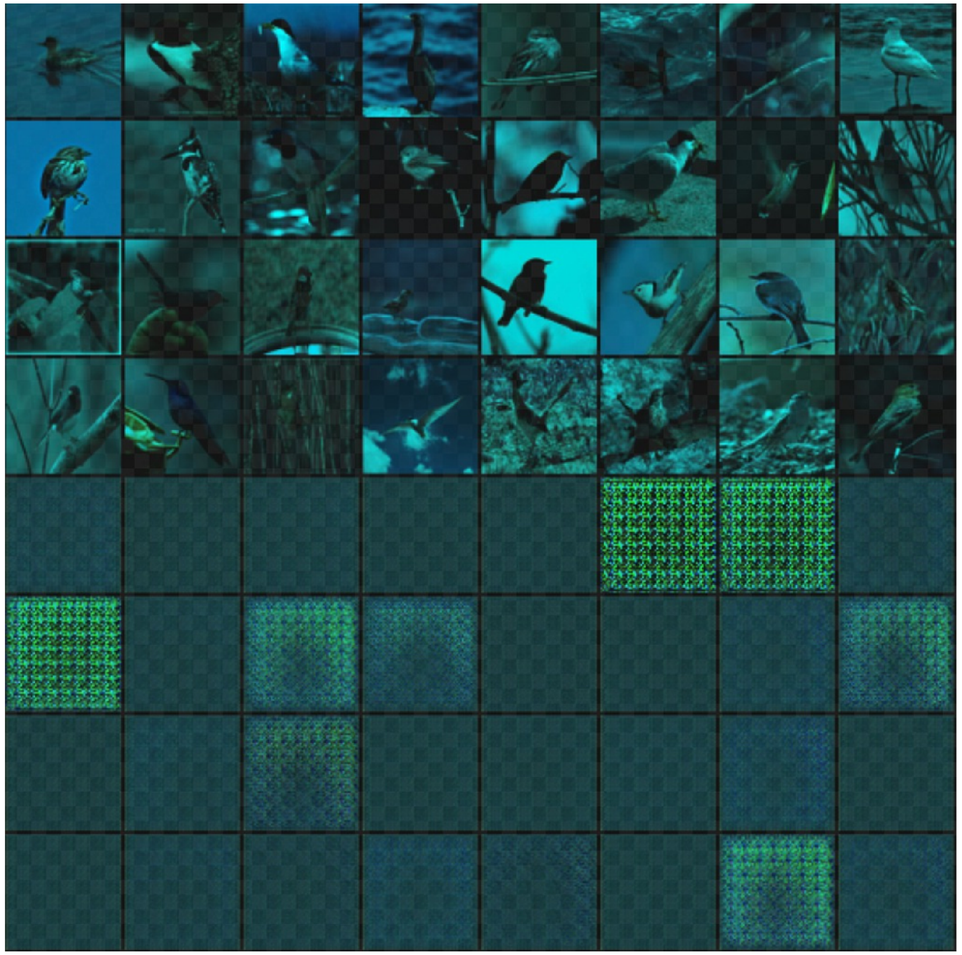
Models	M-4	M-5
(A-1)	132.465	177.490
(A-2)	220.544	114.094
(A-3)	230.650	275.3
(A-4)	352.262	323.598

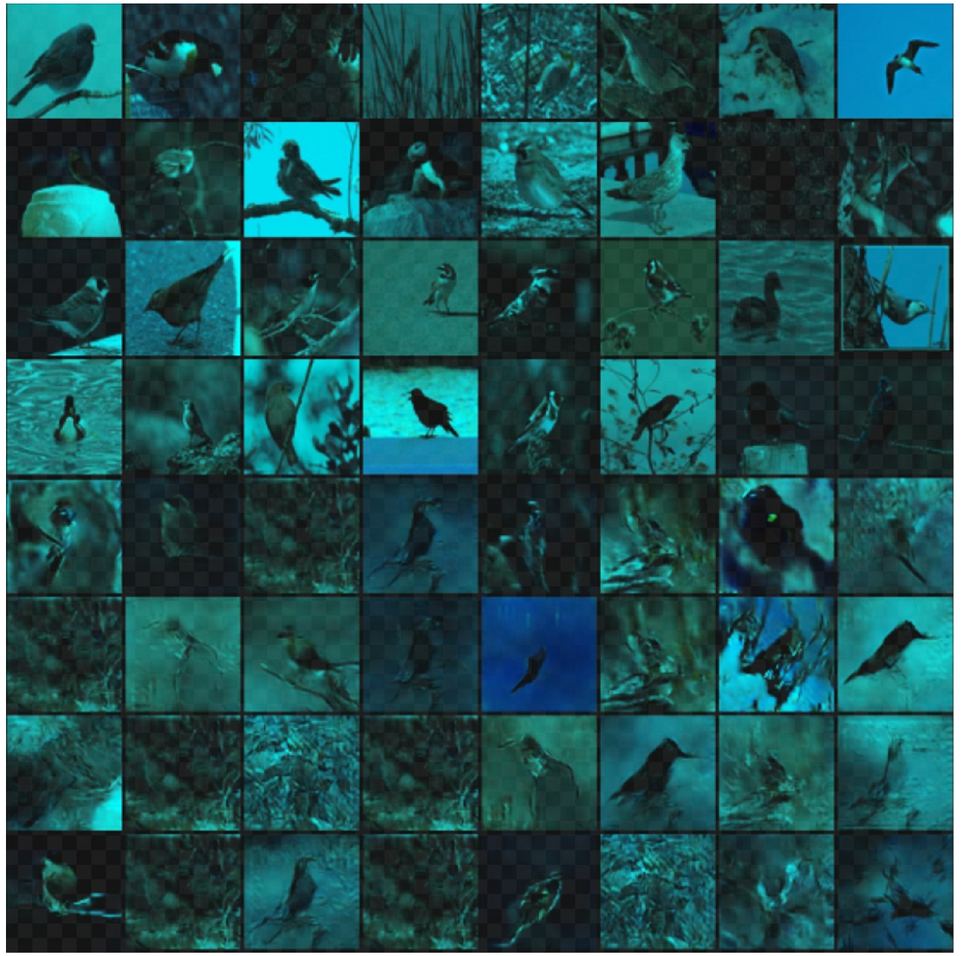
Table 2: Performance of the proposed novelties as compared to the Baseline on the basis of FID-Score.

RESULT

There is a significant decrease in the value of **FID Score, after Quaternionic Convolution** is incorporated in GANs and hence is a significant improvement over the Baseline

Results for Quaternion GANs





ERROR ANALYSIS

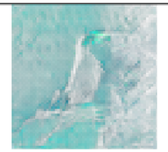
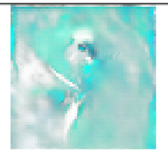
	1	2	3	4	5
Text	Text - A brilliantly orange colored bird with black head, nape and tail, and black wings has white wing bars.	Text - This small bird has a grey bill and crown and grey wings with white wing bars	This bird has a medium beak with mostly yellow feathers.	This small bird has a light brown breast and belly and a small belly-pointed beak.	This bird has a large beak and a long neck.
Image					

Table 3: Table with Images which are generated according to the intent of the user in the input prompt.

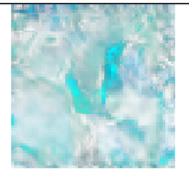
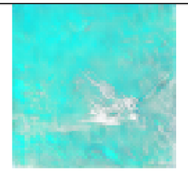
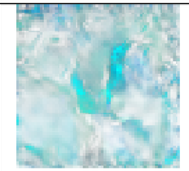
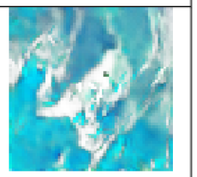
	6	7	8	9	10
Text	Text- This bird has wings that are grey and has a white belly.	Text - A small bird with a black head and yellow underbelly.	This bird is brown black in color with a light pink beak, and brown eye rings.	This bird has wings that are brown and has a yellow belly.	This bird is grey with white and has a very short beak.
Image					

Table 4: Table with Images which are generated contrary to the the intent of the user in the input prompt.

CONTRIBUTION

Aryan designed the novel approach to generative modeling using quaternion algebra. Medha and Akshara extended the project by adding spectral normalization to GAN and QGANs to experiment with potential improvements in GAN training. Aryan, Medha, Akshara conducted experiments to demonstrate the effectiveness of QGANs, SNGANs, and QSNGANs, respectively. Additionally, all team members contributed to the writing and editing of the report to ensure clear and accurate presentation of findings.

FUTURE WORK

1. Design and implement Triplet-Loss based GANs for better stability and training process of Discriminator

2. Extend the explainability using GRAD-CAM for DIscriminator for enhanced feedback.

3. StackGANs can be incorporated with the proposed work for better architecture.

REFERENCES

REFERENCE 1

Reed, Scott, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. "Generative adversarial text to image synthesis." In International conference on machine learning, pp. 1060-1069. PMLR, 2016.

REFERENCE 2

Grassucci, Eleonora, Edoardo Cicero, and Danilo Comminiello. "Quaternion generative adversarial networks." In Generative Adversarial Learning: Architectures and Applications, pp. 57-86. Cham: Springer International Publishing, 2022.

REFERENCE 3

Miyato, Takeru, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. "Spectral normalization for generative adversarial networks." arXiv preprint arXiv:1802.05957 (2018).

REFERENCE 4

Zhang, Han, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N. Metaxas. "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks." In Proceedings of the IEEE international conference on computer vision, pp. 5907-5915. 2017.

THANK YOU