Adversarial Learning for Text Image Semantic Consistency Using Deep Fusion (DF-GAN)

Sujata S. Virulkar[1,2] and Kanchan Tiwari[3]

[1]*E&TC Engg., AISSMS IOIT, Pune, India*
[2]*I2IT, Pune, India*
[3]*E&TC Engg., MESCOE, Pune, India*

Keywords: Generative, Adversarial, Networks, Deep, Fusion, Text, Images.

Abstract: Painting is not just a visual art, but also a human creation. Researchers have been hard at work developing AI systems that can mimic human intellect and carry out tasks previously thought impossible, such as facial recognition, text production, and even artistic creation. Meanwhile, deep convolutional generative adversarial networks (GANs) have started producing visually arresting pictures in select categories. To achieve these goals, we present a Deep Fusion Generative Adversarial Networks that is both easier to implement and more successful in its applications (DF-GAN). To be more precise, we propose (i) exotic fusion block of Deep Text-Image, which make possible understand the fusion process to make a full fusion between text and images, (ii) a exotic Target-aware discriminator combination of One-Way Output and matching aware penalty gradient, that improves the semantic consistency for text-image not either introducing additional networks, and (iii) exotic one-stage text-image We find that the proposed DF-GAN performs the state-of-the-art algorithms on popular datasets, while being more straightforward and efficient in its ability to generate natural-looking and text-matching synthetic pictures.

1 INTRODUCTION

Impressive advancements have been achieved in converting text to photo-realistic images with the use of deep neural networks recently (Brock, Donahue, et al. , 2019). Generative adversarial networks (GANs) demonstrate superiority in creating high-quality pictures when compared to other state-of-the-art networks for text-to-image synthesis. Multiple methods have been suggested to enhance GAN's training process stability and picture resolution (Vries, Strub, et al. , 2017). None of these methods, however, guarantees that all of the information from the input text is fully reflected in the produced picture, making it more difficult to deduce the original text from the image. Their primary emphasis is on better resolution and photo-realism. Rather than concentrating primarily on increasing resolution, it is more important to ensure that all relevant information is extracted from the input text. Let's provide an information theoretic description of this issue. The most crucial part of the text to image synthesis pipeline is the generator, which is used to produce a false picture given a phrase containing words describing certain photographs. We want to Maximize the mutual information of the input sentence and the false picture to motivate the fake image to communicate the information of the input phrase as much as feasible. To get better results from the text, however, producing images with to make high-quality photos, therefore we present a novel neural network to approximate the mutual information and optimize with it.

2 RELATED WORK

In recent years, deep neural networks have seen a lot of success, especially in the natural language processing (NLP) and computer vision (CV) fields. Recent fashionable methods in language modelling and deep generative models provide the backbone of much existing work in text to picture synthesis. For the first time, Variational Autoencoders (VAE) (Cheng, Song, et al. , 2021) used a probabilistic graphical model and continuous latent variables to solve this challenge, with the objective being to optimise the lower limit of data likelihood. The Deep

Virulkar, S. S. and Tiwari, K.
Adversarial Learning for Text Image Semantic Consistency Using Deep Fusion (DF-GAN).
DOI: 10.5220/0013595900004664

In *Proceedings of the 3rd International Conference on Futuristic Technology (INCOFT 2025) - Volume 2*, pages 514-522
ISBN: 978-989-758-763-4

Recurrent Attention Writer (DRAW) (Ding, Yang, et al. , 2021) was developed using a later method that also featured a unique differentiable attention mechanism

But the created photos are very low quality and seem cartoonish compared to real life. Since then, several tweaks and extensions to Generative Adversarial Networks (GAN) (Brown, Mann, et al. , 2020) have been suggested to better optimise its ability to generate crisper images. Several methods (Cheng, Wu, et al. , 2020) have been offered to stabilise the training process and provide convincing results in light of GAN's unsteady training dynamics. Image creation from text has shown some encouraging results using conditional GANs (cGANs), which leverage conditional information for both the discriminator and generator (Cheng, Song, et al. , 2021). An impressive use of GAN, (Cheng, Song, et al. , 2021) may produce pictures that are almost photographic in quality. They also presented other alternatives by using various goal functions to impose smoothness on the language manifold and lessen the likelihood of overfitting. The StackGAN architecture was used as the foundation for our design. To optimise the information expressed, we don't only produce high-resolution, photo-realistic pictures; we also maximize the mutual information between the input text and the output image.

3 THE PROPOSED DF-GAN

In this study, we present a shallow convolutional neural network (CNN) model for text-to-image synthesis (DF-GAN). In order to generate visuals that are both realistic and compatible with the surrounding text, we propose: I a cutting-edge method for immediately synthesising high-resolution pictures without visual feature entanglements by using a text-to-image backbone. (ii) a unique Target-Aware Discriminator that improves text-image semantic consistency without adding additional networks, made up of Gradient Penalty Matching Aware(GP-MA) and output w.r.t One Way. (iii) an innovative Deep Fusion text-image Block (DF-Block) that integrates textual and visual characteristics to a greater extent.

Through the employment of several Deep text-image Fusion Blocks (DFBlock) in UPBlocks, DFGAN creates high resolution pictures directly from a single discriminator and generator.

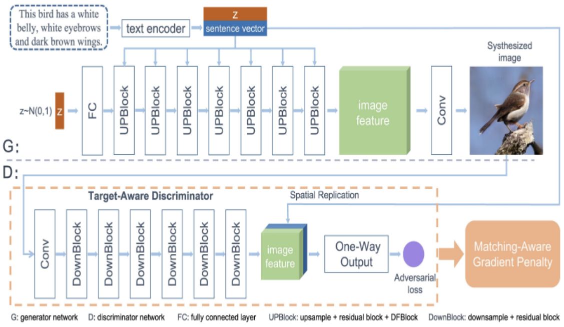


Figure 1: DF-GAN for text-to-image synthesis.

3.1 Model Overview

As can be seen in Figure 1, the proposed DF-GAN consists of three main parts: Generator, Discriminator, and a Pre-Trained Text Encoder. Generator takes two different inputs 1) a phrase encoded vector by a text encoder and from Gaussian Distribution a sampled noise vector —to guarantee that the resulting pictures are diverse. The first step involves reshaping the noise vector by feeding it into a fully linked layer. The discriminator encourages the generator to synthesise pictures with improved quality and text-image semantic coherence by discriminating created images from genuine samples. In order to derive semantic vectors from the provided text description, a Text Encoder is a bidirectional Long Short-Term Memory (LSTM). The AttnGAN pre-trained model is used directly.

3.2 One-Stage Text-to-Image Backbone

Previous text-to-image converters have failed because to the GAN model's inconsistency. As a standard practice, GANs produce high-resolution pictures from low-resolution inputs using a layered architecture [56,57We proposed a single-stage text-image framework that may immediately synthesis highly resolute pictures via a single pair of generator and discriminator, drawing inspiration from previous works on unconditional image production. To maintain consistency during adversarial training, we make use of the hinge loss. It sidesteps potential tangles caused by several generators by having only one in the single-stage backbone Our one-step technique with hinge loss is formulated as follows (Lim and Chul, 2017):

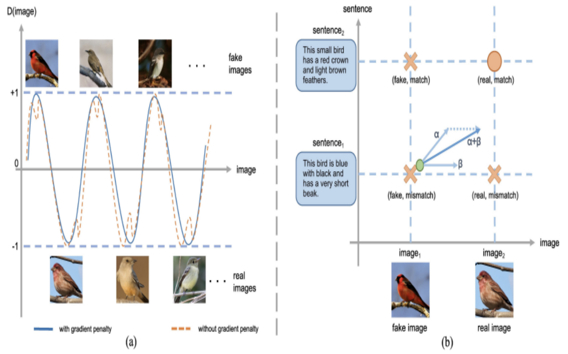


Figure 2: (a) The impact of the gradient penalty on the loss landscape is compared. The gradient penalty helps generator convergence by flattening the discriminator loss surface. A MA-GP schematic (b). We need to use the information we have (actual, compatible). MA-GP.

$$LD = -Ex\sim \Pr\left[\min\left(0, -1 + D(x,e)\right)\right] - \left(\frac{1}{2}\right) EG(z)\sim pg\left[\min\left(0, -1 - D(G(z),e)\right)\right]$$

$$-(1/2)\ Ex(z)\sim pmis\left[\min\left(0, -1 - D(x,e)\right)\right]$$

$$LG = -EG(z)\sim \mathrm{Pg}\left[D(G(z),e)\right]$$

[1]

where z is the vector of Gaussian random noise and e is the vector of sentences. Synthetic data distribution is denoted by Pg, actual data distribution by Pr, and mismatched data distribution by Pmis.

3.3 Target-Aware Discriminator

The proposed Target-Aware Discriminator is described in full below; it consists of the Gradient Penalty Matching Aware(GP-MA) and the output of single way. The Target Aware discriminator encourages the generator to produce more natural and semantically consistent across text and picture representations.

3.4 Matching-Aware Gradient Penalty

To improve text-image semantic consistency, we developed a novel method called the Matching-Aware zero-centered Gradient Penalty (MAGP). Here, we apply the unconditional gradient penalty (Mescheder et al., 2018) to our MA-GP for the text-to-image creation problem after first demonstrating it from a fresh and understandable angle. Figure 2(a) demonstrates how the target data (actual pictures) in unconditional image synthesis have a minimal discriminator loss with incorporating the perspective inside the process of creating images from text. In case of text to image generation, the aspect of discriminator takes in 4 different types of input, as shown in Figure 2(b): fake images with matching Text (match, fake), fake image with the mismatched text (mismatch, fake), real images with matching text (match, real), and real images with mismatched text (mismatch, real). Our whole model formulation using MA-GP reads as follows:

$$LD = -Ex\sim \Pr\left[\min\left(0, -1 + D(x,e)\right)\right] - \left(\frac{1}{2}\right) EG(z)\sim pg\left[\min\left(0, -1 - D(G(z),e)\right)\right]$$

$$-(1/2)\qquad Ex(z)\sim pmis\left[\min\left(0, -1 - D(x,e)\right)\right]$$

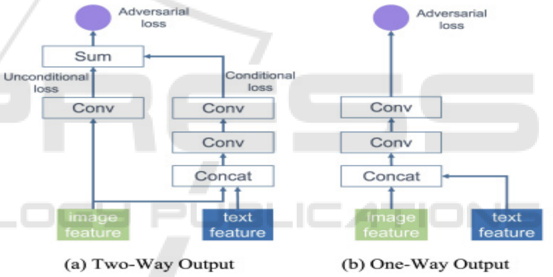
$$+kEx\sim pr\left[\left\|\Delta x D(x,e)\right\| + \left\|\Delta e D(x,e)\right)\right\|$$

$$LG = -EG(z)\sim \mathrm{Pg}\ [\mathrm{D}(\mathrm{G}(z),\mathrm{e})]$$

[2]

where k and p = Hyper parameters to balance the effectiveness of penalty gradient.

Our model is able to more closely approximate the text matching actual data via the use of the MA-GP loss as a regularization on the discriminator, allowing for the generation of more text matching pictures.



(a) Two-Way Output (b) One-Way Output

Figure 3: Our One-Way Output compared to the Two-way Output. The ultimate adversarial loss is predicted by adding the predicted conditional loss to the predicted unconditional loss in the Two-Way Output. To further elaborate on (b), our One-Way Output accurately estimates the total adversarial loss.

3.5 One-Way Output

Previously Text to Image GAN's typically employ images features retrieved by the discriminators in the two ways (Figure 2(a)): one assesses if the picture is genuine or false, and the other concatenates the image features and phrase vectors to evaluated the text to image semantic coherence. the other concatenates the image feature and phrase vector to evaluate text-image semantic coherence Specifically, as seen in Figure 2(b), following back propagation the conditional loss produces a point gradient the conditional loss produces a gradient pointing to both the matched and real inputs, whereas the unconditional loss produces a pointing gradient exclusively to the actual images.. As a result, we advocate adopting the One-Way Output for use in text-to-image translation. Figure 3(b) depicts how our discriminator combines the image feature and phrase vector into a single input before passing it through two convolution layers that generate a single adversarial loss.

3.6 Efficient Text-Image Fusion

We present a new Deep text-image Fusion Block that can effectively combine text and picture data (DF Block). Our DF Block takes the text-image fusion process farther than any prior modules, resulting in a complete text-image fusion. Our DF-GAN uses a 7-UPBlock generator, as illustrated in Figure 1. It is possible to find two Text-Image Fusion blocks inside of a UP Block. In order to forecast the language-conditioned channel-wise scaling parameters and shifting parameters from sentence vector e, we use two MLPs (Multilayer Perceptron), as illustrated in Figure 6(c):

$$\gamma = \mathrm{MLP1}(e), \qquad \theta = \mathrm{MLP2}(e) \qquad [3]$$

For a given input feature map $X \in \mathbb{R}^{B\times C\times H\times W}$, we first conduct the channel-wise scaling operation on X with the scaling parameter γ, then apply the channel-wise shifting operation with the shifting parameter θ. Such a process can be expressed as follows:

where AF F denotes the Affine Transformation; xi is the i th channel of visual feature maps; e is the sentence vector; γi and θi are scaling parameter and shifting parameter for the i th channel of visual feature maps.

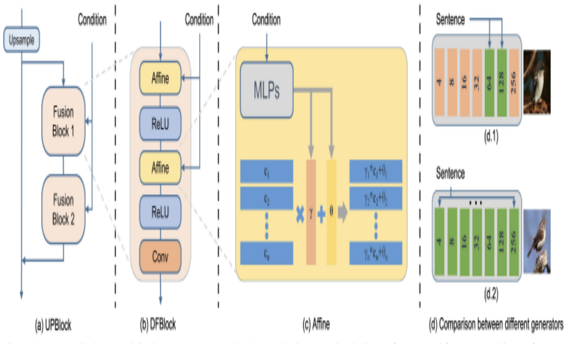


Figure 4: (a) An ordinary UPBlock on the power grid. The UPBlock employs two Fusion Blocks to combine text and picture information, and then upsamples the combined result. Our DFBlock generator (d.2) is compared to (d.1) the generator that uses cross-modal attention.

Portray a variety of graphical elements based on a variety of written explanations This is because most current text-to-picture GANs rely on the cross modal attention mechanism, which experiences exponentially rising computing costs as image sizes expand.

4 EXPERIMENTS

We provide the datasets, training details, and metrics we utilized to evaluate our trials below, and we conclude with quantitative and qualitative assessments of DF-GAN and its derivatives.

4.1 Datasets

We next follow the footsteps of prior work and test the proposed model on two difficult datasets, namely CUB bird and COCO (Lin, Maire, et al. , 2014). The 200 bird species included in the CUB dataset's 11,788 photos. For every picture of a bird, there are 10 words for it in several languages. Eighty thousand pictures are available for training purposes and forty thousand are available for testing purposes in the COCO dataset. There are five language explanations for each picture in this series.

4.2 Training Details

Adam (Kingma, Adam, et al. , 2015) is used to achieve optimal performance for our network, with parameters 1=0.0 and 2=0.9. Two Timescale Update Rule (TTUR) (Heusel, Ramsauer, et al. , 2017) specifies a learning rate of 0.0001 for the generator and 0.0004 for the discriminator.

4.3 Evaluation Details

Our network's efficacy is measured using the Inception Score (IS) and the Frechet Inception Distance (FID) (Heusel, Ramsauer, et al. , 2017), both of which have been used in earlier publications. More specifically, IS calculates the Kullback-Leibler (KL) divergence between a conditional distribution and a marginal distribution. Each produced picture clearly corresponds to a given class, and a higher IS indicates a greater quality of the created photos. In order to compare synthetic and real-world pictures in the feature space of a pre-trained Inception v3 network, FID (Heusel, Ramsauer, et al. , 2017) calculates the Frechet distance between the distributions of the two types of images. As opposed to IS, FID is lower in photographs that seem more natural. Each model creates 30,000 pictures (256256 resolution) using randomly chosen text descriptions from the test dataset in order to calculate IS and FID..

4.4 Quantitative Evaluation

We evaluate the proposed technique against various state-of-the-art algorithms that have also used stacked structures to achieve exceptional performance in text-to-image synthesis, such as StackGAN , StackGAN++, AttnGAN, MirrorGAN, SD-GAN, and DM-GAN. Newer models were also used for comparison. It's important to note that modern models always include outside information or oversight. XMC-GAN employs the additional pretrained VGG-19 and Bert; DAEGAN employs the extra NLTK POS tagging and manually constructs rules for various datasets; and TIME employs the extra 2-D positional encoding.

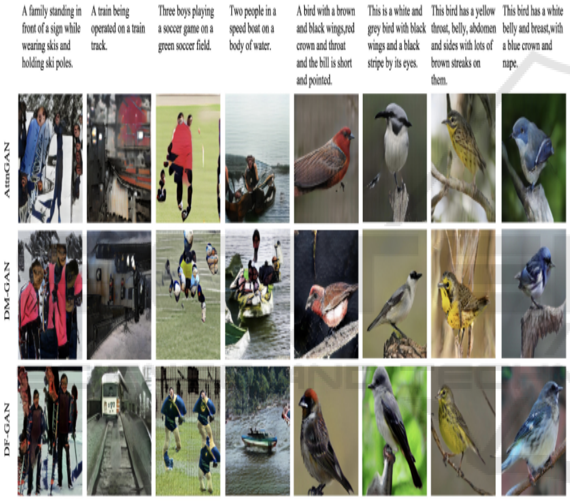

A family standing in front of a sign while wearing skis and holding ski poles.
A train being operated on a train track.
Three boys playing a soccer game on a green soccer field.
Two people in a speed boat on a body of water.
A bird with a brown and black wings,red crown and throat and the bill is short and pointed.
This is a white and grey bird with black wings and a black stripe by its eyes.
This bird has a yellow throat, belly, abdomen and sides with lots of brown streaks on them.
This bird has a white belly and breast,with a blue crown and nape.
AttnGAN
DM-GAN
DF-GAN

Figure 5: Using the test set of COCO and CUB datasets, we show examples of the pictures generated using AttnGAN, DM-GAN, and our new DF-GAN conditioned on text descriptions.

Table 1: Based on CUB and COCO datasets the results of FID,IS and NoP.

Model	COCO		CUB	
	FID↓	NoP↓	IS↑	FID↓
CPGAN	-	-	52.48	289M
TIME	3.91	13.3	29.14	111M
MirrorGAN	3.56	17.34	23.71	-
DAE-GAN	3.42	13.21	26.12	93M
AttnGAN	3.36	18.71	25.49	240M
SD-GAN	3.67	-		
XMC-GAN	-	-	8.3	156M
StackGAN	4.7	-	-	-
StackGAN++	4.84	-	-	
DM-GAN	3.75	14.19	21.64	36M
DF-GAN (Ours)	5.1	14.81	19.32	19M

As can be shown in Table 1, when compared to other state-of-the-art models, our DF-GAN has a much lower NoP while still producing respectable results. On the CUB dataset, our DF-GAN improves upon the IS metric (4.36 to 5.10) and reduces the FID metric (23.98 to 14.81) compared to AttnGAN which uses cross-modal attention to fuse text and image characteristics. FID is reduced from 35.49 to 19.32 using our DF-GAN on the COCO dataset.

4.5 Qualitative Evaluation

We additionally evaluate AttnGAN, DM-GAN, and our own suggested DF-GAN with regards to their respective visualisation outputs. Figure 6 demonstrates how AttnGAN and DM-GAN produced pictures are only a mixture of fuzzy shapes and a few visual elements (1st, 3rd , 5 th, 7th, and 8th columns). Both AttnGAN and DM-GAN provide incorrect bird forms, as seen in the fifth, seventh, and eighth columns, respectively. Our DF-GAN also produces synthetic pictures with more accurate item forms and fine-grained features (e.g., 1st, 3rd, 7th, and 8th columns). Our DF-GAN product also has a more realistic avian stance (e.g., 7th and 8th columns). We discover that our DF-GAN can capture more nuanced features in text descriptions compared to previous models by analysing the text-image semantic consistency. Figure 5 shows that the proposed DF-GAN is able to synthesis images that better match the textual descriptions.

5 COMPARISON WITH STATE-OF-THE-ART METHODS

We conduct qualitative comparisons of the proposed technique to many state-of-the-art methods on the MSCOCO, CUB-200, and Oxford-102 datasets. These methods include certain GAN-based methods, DALL-E, and CogView (Hong, Yang, et al. , 2018). In Table 2, we display the results of our FID (Liang, Pei, et al. , 2020) analysis between 30,000 synthetic and 30,000 real images. As compared to other GAN-based models with a similar amount of parameters, we find that our compact model, VQ-Diffusion-S, performs admirably on the CUB-200 and Oxford102 datasets. VQ-Diffusion-B, our starting model, enhances the efficiency even further. Further, on the MSCOCO dataset, our VQ-Diffusion-F model outperforms all other approaches by a wide margin, including those with ten times as many parameters as ours, such as DALL-E and CogView (Hong, Yang, et al. , 2018). Figure 2 displays some visual comparison results using DM-GAN and DF-GAN. Naturally, the synthetic images we produce are more faithful to the source text and have more realistic fine-grained features.

5.1 In the wild text-to-image synthesis

We train our model on three subsets of the LAION400M dataset, including the cartoon, icon, and human datasets, to show that it can generate images in the wild. Here in Figure 3 we show you the outcomes of our study. Despite the fact that our starting model is significantly less complex than that of DALL-E and CogView, we still managed to get impressive results. In contrast to the AR approach, which creates images in a sequential fashion (from top left to bottom right), our method generates images simultaneously from all directions. This means that our approach can be used for a wide variety of visual tasks, such as mask inpainting with irregular edges. It is not necessary to re-train a model for this purpose. After labelling the tokens in the out-of-shape area with the [MASK] token, we feed them into our model. Both unconditional and text-conditioned mask inpainting are supported by this method.

5.2 Ablations

Number of timesteps. We look into the training and inference time periods. In the experiment depicted in Table 2, we use the CUB-200 dataset. The results seem to plateau at around 200 training steps, but we found that increasing the number of steps from 10 to 100 yields the best results. So, in our tests we used a timestep size of 100 instead of the usual of 10. We test the generated images from 10, 25, 50, and 100 inference steps on five models with varying training steps to illustrate the quick inference technique. After eliminating half of the inference stages, we find the performance is still satisfactory. This might save another half of the inference time.

Mask-and-replace diffusion strategy. We investigate the performance benefits of the mask-and-replace technique on the Oxford-102 dataset. To test this, we varied the final mask rate (γT). Our mask-and-replace approach includes the exceptional situations of mask-only strategies γ(T = 1) and replace-only strategies (γT = 0). In Figure 4 we can see that the optimal performance occurs when M = 0.9. The error accumulation problem may arise when M is greater than 0.9, while determining which part of the network requires more focus may be challenging when M is less than 0.9.

Table 2: FID comparison of different text-to-image synthesis method on MSCOCO, CUB-200, and Oxford-102 datasets.

Model	MSCOCO	CUB-200	Oxford-102
Cogview	37.1	-	-
DALLE]	37.5	46.1	-
DAE-GAN	38.12	16.19	-
EFF-T2I	-	19.17	14.47
SEGAN	28.28	20.17	-
DF-GAN	27.42	18.81	-
AttnGAN	28.49	24.98	-
StackGAN	67.05	54.89	52.28
StackGAN++	78.59	18.3	49.68
DM-GAN	28.64	18.09	-
VQ-D-S	29.17	10.97	16.95
VQ-D-B	18.75	12.94	13.88
VQ-D-F	12.86	11.32	13.14

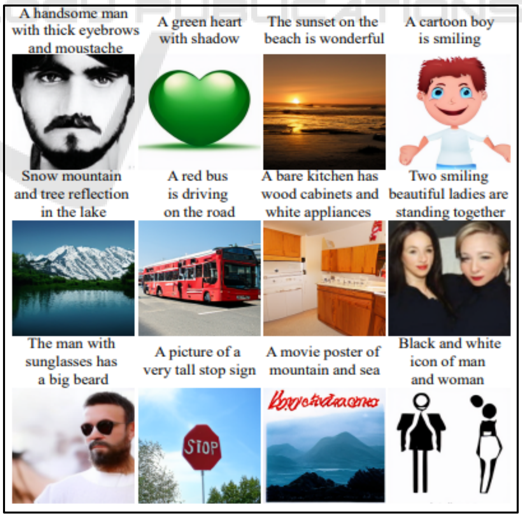

A handsome man with thick eyebrows and moustache
A green heart with shadow
The sunset on the beach is wonderful
A cartoon boy is smiling
Snow mountain and tree reflection in the lake
A red bus is driving on the road
A bare kitchen has wood cabinets and white appliances
Two smiling beautiful ladies are standing together
The man with sunglasses has a big beard
A picture of a very tall stop sign
A movie poster of mountain and sea
Black and white icon of man and woman
STOP

Figure 6: The text-to-image synthesis results.

Truncation. We also show the critical role that truncation sampling plays in our discrete diffusion-based approach. There's a chance that this would prevent the network from randomly picking tokens with low probabilities. The top r tokens of p(x 0|xt, y) are the only ones that will be kept in the inference phase. We test the outcomes on the CUB-200 dataset using varying truncation rates (r). Figure 4 shows that optimal performance is reached when the truncation rate is 0.86.

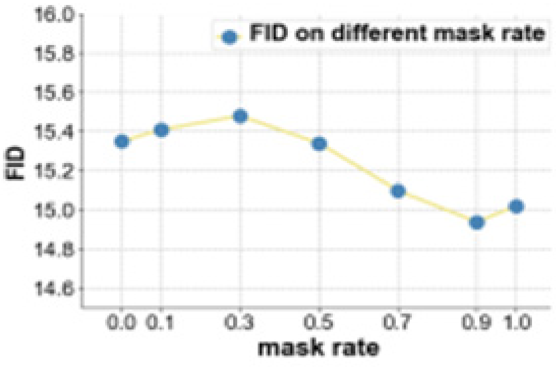


Figure 7: Ablation study on the mask rate.

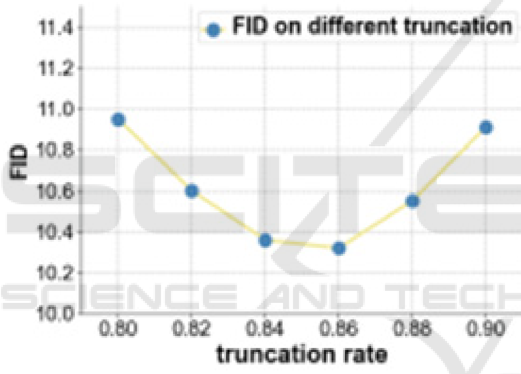


Figure 8: Ablation study on the truncation rate.

Table 3: Different Result between VQ-Diffusion and VQ-Model w.r.t. steps and FID.

Model	Steps	FID	throughput (imgs/s)
VQ-AR-B	20	18.79	0.03
VQ-AR-S	35	17.32	0.08
VQ-D-S	25	16.56	1.25
VQ-D-S	50	14.82	0.67
VQ-D-S	100	13.67	0.37
VQ-D-B	25	15.13	0.47
VQ-D-B	50	13.75	0.24
VQ-D-B	100	12.84	0.13

Contrasting VQ-Diffusion and VQ-AR. To provide a level playing field, we swap out the diffusion image decoder for an autoregressive decoder using the same network architecture while leaving all other parameters, such as the image and text encoders, same. At the same time, we measure the efficiency of both strategies on a V100 GPU using a batch size of 32 samples. Fast inference technique VQ Diffusion is 15 times faster than the best FID-scoring VQ-AR model.

5.3 Unified generation model

Since our method is generic, it can be used for a variety of image synthesis tasks, including both unconditional and labelled synthesis. After stripping out the text encoder network and cross attention section from transformer blocks, we inject the class label via the AdaLN operator to produce images based on the label. In total, there are 24 512-by-512-pixel blocks of transformers across our network. The ImageNet dataset is used to train our model. Specifically, we use the VQ-GAN (Karras, Laine, et al. , 2019) model downsampled from 256 256 to 16 16 that has been publically published and trained on the ImageNet dataset to perform VQ-VAE. Table 4 displays our quantitative findings. In contrast to the higher FID scores reported by some task-specific GAN models, our method delivers a unified model that performs admirably on a wide variety of tasks.

Table 4: ImageNet and FFHQ based FID score comparison.

Model	ImageNet	FFHQ
StyleGAN2	-	3.9
BigGAN	6.59	11.9
BigGAN-deep	7.91	-
IDDPM	11.8	-
ADM-G	11.91	-
VQGAN	14.68	9.7
ImageBART	20.29	9.67
ADM-G (1.0guid)	05.00	-
VQGAN (acc0.05)	06.01	-
ImageBART (acc0.05)	8.04	-
Ours	11.89	6.33

The central idea is to create a model of the VQ-VAE latent space that is not autoregressive. To prevent the AR model's flaws from piling up, the authors propose a mask-and-replace diffusion technique. When compared to earlier GAN-based text-to-image approaches, our model's scene generation capability is superior. Both unconditional and conditioned image formation benefit greatly from our approach, and our results are quite promising.

6 LIMITATIONS

For future research, it is important to keep in mind the limits of DF-GAN, notwithstanding its advantage in text-to-image synthesis. It's important to note that our model's capacity to synthesize fine-grained visual features is constrained since we only provide sentence-level text information. Second, using pre-trained, big language models to provide more information may further enhance performance. In our further efforts, we hope to overcome these restrictions.

7 CONCLUSION AND FUTURE SCOPE

In this research work, we present a new deep feedforward (DF) GAN for text-to-image job. Here we provide a single-stage text-to-image Backbone capable of immediately synthesizing high-resolution pictures without intermediate stages or inter-generator dependencies. Furthermore, we provide a unique Target-Aware Discriminator that combines Matching-Aware Gradient Penalty (MAGP) with One-Way Output. The text-image semantic consistency may be improved further without the need for additional networks. Along with this, we provide a unique Deep text-image Fusion Block (DF Block) that completely fuses text and picture information more efficiently and profoundly. Extensive experiments show that our proposed DF-GAN far outperforms the state-of-the-art models on the CUB dataset and the even more difficult COCO dataset.

REFERENCES

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In International Conference on Learning Representations, 2019.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. arXiv preprint arXiv:2005.14165, 2020.

Jun Cheng, Fuxiang Wu, Yanling Tian, Lei Wang, and Dapeng Tao. Rifegan: Rich feature generation for text-toimage synthesis from prior knowledge. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10911–10920, 2020.

Wen-Huang Cheng, Sijie Song, Chieh-Yun Chen, Shintami Chusnul Hidayati, and Jiaying Liu. Fashion meets computer vision: A survey. ACM Computing Surveys (CSUR), 54(4):1–41, 2021.

Harm De Vries, Florian Strub, Jer´ emie Mary, Hugo ´ Larochelle, Olivier Pietquin, and Aaron C Courville. Modulating early visual processing by language. In Advances in Neural Information Processing Systems, pages 6594–6604, 2017. 5, 8

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cog view: Mastering text-to-image generation viatransformers. arXivpreprintarXiv:2105.13290, 2021. 2, 6

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Advances in Neural Information Processing Systems, pages 2672–2680, 2014. 1,

Yuchuan Gou, Qiancheng Wu, Minghao Li, Bo Gong, and Mei Han. Segattngan: Text to image generation with segmentation attention. arXiv preprint arXiv:2005.12444, 2020. 1

Caglar Gulcehre, Sarath Chandar, Kyunghyun Cho, and Yoshua Bengio. Dynamic neural turing machine with continuous and discrete addressing schemes. Neural computation, 30(4):857–884, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In Advances in neural information processing systems, pages 6626–6637, 2017.

Seunghoon Hong, Dingdong Yang, Jongwook Choi, and Honglak Lee. Inferring semantic layout for hierarchical textto-image synthesis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7986– 7994, 2018.

Xun Huang and Serge Belongie. Arbitrary style transfers in real-time with adaptive instance normalization. In Proceedings of the IEEE International Conference on Computer Vision, pages 1501–1510, 2017.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In International Conference on Machine Learning, 2015.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4401–4410, 2019.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8110–8119, 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In International Conference on Learning Representations, 2015.

Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip Torr. Controllable text-to-image generation. In Advances in Neural Information Processing Systems, pages 2065–2075, 2019.

Ruifan Li, Ning Wang, Fangxiang Feng, Guangwei Zhang, and Xiaojie Wang. Exploring global and local linguistic representation for text-to-image synthesis. IEEE Transactions on Multimedia, 2020.

Wenbo Li, Pengchuan Zhang, Lei Zhang, Qiuyuan Huang, Xiaodong He, Siwei Lyu, and Jianfeng Gao. Object-driven text-to-image synthesis via adversarial training. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 12174–12182, 2019.

Adong Liang, Wenjie Pei, and Feng Lu. Cpgan: Contentparsing generative adversarial networks for text-to-image synthesis. In European Conference on Computer Vision, pages 491–508. Springer, 2020.

Jae Hyun Lim and Jong Chul Ye. Geometric gan. arXiv preprint arXiv:1705.02894, 2017.

Junyang Lin, Rui Men, An Yang, Chang Zhou, Ming Ding, Yichang Zhang, Peng Wang, Ang Wang, Le Jiang, Xianyan Jia, et al. M6: A chinese multimodal pretrainer. arXiv preprint arXiv:2103.00823, 2021.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C Lawrence ´ Zitnick. Microsoft coco: Common objects in context. In European conference on computer vision, pages 740–755. Springer, 2014.

Bingchen Liu, Kunpeng Song, Yizhe Zhu, Gerard de Melo, and Ahmed Elgammal. Time: text and image mutual-translation adversarial networks. arXiv preprint arXiv:2005.13192, 2020.

Ming-Yu Liu, Xun Huang, Jiahui Yu, Ting-Chun Wang, and Arun Mallya. Generative adversarial networks for image and video synthesis: Algorithms and applications. Proceedings of the IEEE, 109(5):839–862, 2021.

Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In International Conference on Machine Learning, pages 3481–3490, 2018.

Takeru Miyato and Masanori Koyama. cgans with projection discriminator. arXiv preprint arXiv:1802.05637, 2018.