Applied Computing and Informatics

Image generation based on image description using artificial intelligence

Andrej Šimić and Marina Bagić Babac
Faculty of Electrical Engineering and Computing, University of Zagreb, Zagreb, Croatia

Received 18 May 2025
Revised 1 September 2025
Accepted 28 October 2025

Abstract

Purpose – The purpose of this study is to explore and evaluate advanced text-to-image synthesis methods that generate realistic and semantically aligned images from textual descriptions. By leveraging modern deep learning approaches, the research aims to improve image generation quality, diversity and textual coherence using artificial intelligence techniques.

Design/methodology/approach – The research focuses on designing, implementing and training four different text-to-image generator architectures based on generative adversarial networks (GANs) and transformer-based text embeddings. Two distinct text-to-image fusion strategies were applied: deep fusion (DF) via affine transformations and a semantic-spatial attention mechanism. The models were trained on three large datasets (CUB-200, MS-COCO and ImageNet), resulting in 12 unique generator-discriminator configurations. Performance was evaluated using the inception score and Fréchet inception distance (FID).

Findings – The proposed architecture, combining DF blocks and Semantic-Spatial Aware Convolution Network (SSACN) blocks, achieved competitive results, outperforming several existing models such as AttnGAN, MirrorGAN and DF-GAN in terms of FID. The best-performing model demonstrated its ability to generate diverse and high-quality images that are semantically consistent with the input captions. The use of semantic-spatial fusion further improved the focus and alignment of generated content to the relevant regions described in the text.

Originality/value – This work contributes to the field of text-to-image synthesis by introducing and experimentally validating a hybrid fusion approach that integrates global and spatially aware semantic conditioning. The developed models, supported by a systematic evaluation across multiple datasets, demonstrate improved performance over several state-of-the-art solutions, offering a valuable framework for future research in multimodal content generation.

Keywords Artificial intelligence (AI), Natural language processing (NLP), Text-to-image synthesis, Image generation, Image description, Generative adversarial networks (GANs), Computer vision, Deep learning

Paper type Research article

1. Introduction

Text-to-image synthesis, the task of generating realistic images from textual descriptions, has gained significant attention due to its applications in graphic design, marketing and gaming [1]. Advances in deep learning, particularly generative adversarial networks (GANs) [2] and transformers [3], have enabled models to produce high-quality, semantically consistent images [4]. However, challenges remain in effectively fusing textual and visual features to ensure alignment between generated images and captions.

This paper investigates GAN-based text-to-image synthesis architectures, focusing on novel techniques to integrate textual information into image generation. We propose models

Funding: This work was supported by the European Union's Horizon Europe research and innovation programme (No. 101086179). The funders had no role in the design of the study; nor in the collection, analyses or interpretation of data; nor in the writing of the manuscript; nor in the decision to publish the results.

Applied Computing and Informatics
Emerald Publishing Limited
e-ISSN: 2210-8327
p-ISSN: 2634-1964
DOI 10.1108/ACI-05-2025-0186

that leverage advanced fusion mechanisms to enhance semantic consistency, building on recent methods ([5, 6]). Our approach combines multiple generator configurations, trained and evaluated on large-scale datasets [7], including CUB-200 [8], MS-COCO [9] and ImageNet [10]. Using the inception score (IS) [11] and Fréchet inception distance (FID) [12], we assess the quality and diversity of generated images, comparing our results to state-of-the-art models. Our findings demonstrate improved performance in generating diverse, high-quality images aligned with textual inputs, with potential applications in automated content creation and visualization.

2. Related work

Early approaches to artificial intelligence (AI)-generated art focused on style transfer, from simple image analogies [13] to neural feature visualization like DeepDream [14]. Optimization-based methods [15] were later accelerated by feed-forward generation [16].

The introduction of GANs [2] enabled novel applications such as image-to-image translation [17], artistic style deviation [18] and improved realism through large-scale models [19]. In text-to-image synthesis, early models [20] introduced text-conditioned GANs, extended in StackGAN [21] via multi-stage generation and conditioning augmentation.

More recently, transformer and diffusion-based models such as DALL·E [22], Guided Language to Image Diffusion for Generation and Editing (GLIDE) [23], Imagen [24] and Stable Diffusion [25] have set new benchmarks in fidelity and text-image alignment [26]. However, these models require massive resources and often lack transparency.

Our work builds on GAN-based methods, introducing a hybrid architecture that integrates deep fusion (DF) and semantic-spatial attention. It achieves competitive results on public datasets (CUB-200, MS-COCO and ImageNet), outperforming baselines like AttnGAN, MirrorGAN and DF-GAN in FID, while maintaining interpretability and efficiency.

3. Methodology

This study develops and evaluates text-to-image synthesis models based on GANs, incorporating advanced text-to-image fusion techniques. The proposed system consists of a generator and a discriminator, both conditioned on textual captions encoded using the Deep Attentional Multimodal Similarity Model (DAMSM) [27].

Generator architecture. The generator is a deep deconvolutional neural network that maps a random noise vector $z \in R^{100}$ and a text embedding vector $t \in R^{256}$ to an red green blue (RGB) image of dimensions $3 \times 256 \times 256$. The architecture comprises seven fusion blocks, each followed by an UpBlock that doubles the feature map size using nearest-neighbor interpolation, starting from ngf$\times 8 \times 4 \times 4$ to ngf$\times 256 \times 256$ (where ngf $= 64$). The final layers include batch normalization, a leaky ReLU (slope 0.2), a deconvolutional layer (kernel size 3×3, stride 1 and valid padding) and a hyperbolic tangent activation to produce the output image. Two types of fusion blocks are employed: DF Blocks [5] and Semantic-Spatial Aware Convolution Network (SSACN) Blocks [6].

(1) *DF Block*: Inputs feature maps and a text embedding, processed through two affine transformations (each with fully connected layers to predict scaling (γ) and shifting (β) parameters), followed by ReLU and a convolutional layer (kernel 3×3, stride 1, valid padding). This enables DF of textual and visual features [5].

(2) *SSACN Block*: Extends the DF Block by incorporating a mask predictor that generates a spatial mask to focus text fusion on relevant image subregions. The mask predictor consists of a convolutional layer (kernel 3×3, 100 output channels), batch normalization, ReLU, another convolutional layer (kernel 1×1, 1 output channel) and a sigmoid activation. The mask modulates the affine transformation parameters, enhancing semantic consistency [6].

Four generator configurations were designed by varying the fusion blocks: (1) 7 DF Blocks (7DF, 28.7M parameters), (2) 4 DF Blocks followed by 3 SSACN Blocks (4DF-3SSACN, 29.5M parameters), (3) 4 SSACN Blocks followed by 3 DF Blocks (4SSACN-3DF, 30.5M parameters) and (4) 7 SSACN Blocks (7SSACN, 31.4M parameters).

Discriminator architecture. The discriminator is a deep convolutional neural network that evaluates image-caption pairs, outputting a score in [0, 1] (0 for fake/mismatched, 1 for real/matched). It processes an RGB image ($3 \times 256 \times 256$) and a text embedding (R^{256}) through seven convolutional layers. The first layer uses a 3×3 kernel, stride 1 and valid padding, producing ndf feature maps (ndf = 64). The next five layers use 4×4 kernels, stride 2 and padding 1, doubling the channels and halving the spatial dimensions. The final layer (kernel 4×4, stride 2 and valid padding) outputs a scalar. Leaky ReLU (slope 0.2) follows all but the first layer. The total parameters are approximately 80.4M.

The models were evaluated using the IS and FID on 10,000 generated images per model, compared against real images from each dataset. The 4DF-3SSACN configuration was selected for comparison with baselines (AttnGAN, MirrorGAN, DF-GAN and Semantic-Spatial Aware Generative Adversarial Network (SSA-GAN) based on its balanced performance.

The IS evaluates the quality and diversity of generated images [11]. It is computed as:

$$\mathrm{IS} = \exp(\mathrm{E_x}[\mathrm{D_{KL}}(\mathrm{p(y|x)} \parallel \mathrm{p(y)})])$$

where (E_x) denotes the expectation over generated images (x), (D_{KL}) is the Kullback-Leibler divergence, $(p(y|x))$ is the conditional class probability distribution from a pretrained Inception v3 model [28] and $(p(y))$ is the marginal class distribution. The KL divergence is defined as:

$$\left[D_{KL}(p(y|x) \parallel p(y)) = \sum_y p(y|x) \log\left(\frac{p(y|x)}{p(y)}\right) \right]$$

Higher IS values $(range: (1 \le IS(G) \le 1000))$ indicate high-quality $(sharp(p(y|x)))$ and diverse $(even(p(y)))$ images [29].

The FID measures similarity between feature distributions of real and synthetic images [26]. It is computed as:

$$\left[FID = \left| m - m_r \right|_2^2 + \mathrm{Tr}\left(C + C_r - 2(CC_r)^{1/2}\right) \right]$$

where (m) and (C) are the mean and covariance matrix of features from synthetic images, extracted from the last pooling layer of Inception v3, (m_r) and (C_r) are the mean and covariance of real image features, $(|\cdot|_2^2)$ is the squared L2 norm and (Tr) is the matrix trace (sum of diagonal elements). Lower FID values indicate better quality and diversity.

4. Experiment setup and result analysis

The models were trained on three datasets: CUB-200 (11,788 images, 200 bird species) [8], MS-COCO (328,000 images, diverse scenes) [9] and ImageNet (1,200,000 images, 1,000 categories) [10]. Captions were encoded using the DAMSM [27], producing 256-dimensional sentence feature vectors. The training followed a minimax optimization, where the generator G minimizes $\log(1 - D(G(z,t)))$ and the discriminator D maximizes $E_{(x,t)\sim p_{\mathrm{dt}}}[\log D(x,t)] + wE_{(x,t')\sim p_{\mathrm{dt}}}[\log(1 - D(x,t'))] + (1-w)E_{z\sim p_z, t\sim p_{\mathrm{dt}}}[\log(1 - D(G(z,t),t))]$, with G generating images from noise z and text embedding t, D classifying image-caption pairs (x: real image, t': mismatched caption) and $w = 0.3$ balancing the loss contributions of mismatched and fake pairs. Manifold interpolation, blending text embeddings to create synthetic captions, was

applied to enhance data diversity [20]. The Adam optimizer (learning rate 0.0002, $\beta_1 = 0.5$, $\beta_2 = 0.999$) was used, with 500 epochs for CUB-200 and MS-COCO and 20 epochs for ImageNet, using a batch size of 32.

Table 1 shows the results of the FID and IS measures on the CUB-200 test images. The first column lists the generator models by their abbreviations, the second column holds the achieved FID scores and the third column lists the achieved IS scores. As noted before, a lower FID value means higher quality of images, and conversely, a lower IS value means higher quality and variety of images.

The 7SSACN generator model achieves the best FID score of 24.16, followed by the 4DF-3SSACN model with a slightly higher score of 24.77. The model 4SSACN-3DF yields the lowest FID score of 31.57, which is significantly worse than the other three models. The 4DF-3SSACN generator achieved the highest IS value of 4.45, while the 7SSACN model produced the lowest value of 3.94.

Table 2 shows the results of the different generator configurations on the MS-COCO dataset. The 7SSACN model achieves the lowest FID score of 21.18, and the 7DF and 4DF-3SSACN models follow with slightly lower scores. The 4SSACN-3DF performs worst, with a FID score of 22.62. As for the IS measure, the 4SSACN-3DF achieves the worst result of 18.82, while the 7DF performs best, with a score of 19.68.

Lastly, Table 3 shows the results of FID and IS measures on the ImageNet images. The model with the best performance was 7DF, with a FID score of 129.72, and an IS of 5.74. The worst model considering both measures was 4DF-3SSACN, with a FID of 138.17, and an IS value of 5.27. The FID scores of the ImageNet images are significantly higher than those of the other two datasets. The reason for this is the failure of the models to converge during the training process.

Table 1. The results of FID and IS on the CUB-200 images

Model	FID (↓)	IS (↑)
7DF	27.11	4.02
4DF-3SSACN	24.77	*4.45*
4SSACN-3DF	31.57	4.10
7SSACN	*24.16*	3.94

Table 2. The results of FID and IS on the MS-COCO images

Model	FID (↓)	IS (↑)
7DF	21.49	*19.68*
4DF-3SSACN	21.61	18.85
4SSACN-3DF	22.62	18.82
7SSACN	*21.18*	19.14

Table 3. The results of FID and IS on the ImageNet images

Model	FID (↓)	IS (↑)
7DF	*129.72*	*5.74*
4DF-3SSACN	138.17	5.27
4SSACN-3DF	132.46	5.60
7SSACN	136.11	5.33

Considering the results on all three datasets, the 7SSACN model proved to be the most effective. Its FID score was the lowest on both the CUB-200 and MS-COCO datasets, and its IS score was satisfactory on all three datasets. Although slightly worse than the 7SSACN model, the 4DF-3SSACN model also achieved significant results. It had the highest IS value on the CUB-200 dataset, and it also achieved a considerably lower FID score than the 4SSACN-3DF and 7DF models on the CUB-200 dataset.

The generator model that undoubtedly performed the worst was 4SSACN-3DF. It achieved the lowest FID score on the CUB-200 and MS-COCO datasets and the lowest IS score on the MS-COCO dataset. The reason for this could be the use of SSACN blocks in shallow layers of the generator architecture. In Table 4, we can see that the early layers of the generator network have smaller sizes of feature maps. Consequently, the spatial masks corresponding to these feature maps are also smaller. For example, the first layer outputs feature maps of size (4, 4), which means the mask predictor component produces a spatial mask of the same size. As the size of feature maps gradually increases throughout the network, the application of such a small mask affects a large part of the final image. Because of this, text-to-image fusion possibly isn't accurate in some text-relevant subregions of the final image, which can result in images that are less semantically consistent with the provided captions.

In Table 4, we show a comparison of our FID results with results of some of the existing research efforts in the field of text-to-image synthesis. Specifically, we compare it with the following GAN models – AttnGAN [27], MirrorGAN [30], DF-GAN [5] and SSA-GAN [6]. As shown in the table, the SSA-GAN achieved the lowest FID scores on both datasets, surpassing the other models by wide margins. Although our model performed worst on the CUB-200 images, it was second best on the MS-COCO dataset, significantly outperforming the other three models, along with the SSA-GAN model. The model with the worst FID score on the MS-COCO dataset is the AttnGAN, with a value of 35.49.

In Table 5, we show a comparison of our IS results with the results of the other GAN models. Because some of the IS scores on the MS-COCO dataset are unavailable, we only show the results on the CUB-200 dataset. As we can see from the table, the SSA-GAN has the highest IS score, with a value of 5.17. With all results being similar, our model slightly outperformed the AttnGAN, which achieved an IS of 4.36.

Upon closer analysis of the evaluation results of the models, we can notice that most of the achieved IS values are very low. Considering that the upper limit of the IS is 1,000, it is

Table 4. Comparison of FID results on the CUB-200 and MS-COCO images

Model	CUB-200	MS-COCO
AttnGAN	23.98	35.49
MirrorGAN	18.34	34.71
DF-GAN	19.24	28.92
SSA-GAN	*15.61*	*19.37*
4DF-3SSACN (ours)	24.77	21.180

Table 5. Comparison of IS results on the CUB-200 images

Model	IS (↑)
AttnGAN	4.36
MirrorGAN	4.56
DF-GAN	4.86
SSA-GAN	*5.17*
4DF-3SSACN (ours)	4.45

reasonable to assume that state-of-the-art models, such as SSA-GAN, would yield a significantly higher IS value than 5.17. An explanation for these unexpected results can be found in Ref. [29], where the authors call for researchers in the field of text-to-image to be cautious when using the IS metric to compare different generative models.

One of the limitations of the IS, which certainly disrupts the scores we achieved on the CUB-200 and MS-COCO datasets, is that it doesn't provide reliable results on datasets other than ImageNet, which was used to train its underlying Inception network. As the classes in the CUB-200 and MS-COCO aren't analogous to the ones in the ImageNet dataset, the predicted classes of the Inception network and the actual classes of images generated by the trained models are expected to mismatch to a certain degree.

More specifically, the dataset classes can be misaligned in two different ways – some classes can be present in the ImageNet dataset and not in the other two datasets, and conversely, there are classes that might be present in the CUB-200 and MS-COCO datasets but not in ImageNet. In the first scenario, the larger number of ImageNet classes reduces the calculated entropy of the marginal probability distribution $p(y)$, estimating the diversity of images and consequently decreases the computed IS. For instance, if a text-to-image synthesis model is trained on images of animals and is meant to generate images of this type only, the computed IS is going to be poor, no matter how effectively the model generates images of animals. In the second scenario, the lack of classes in the inception network causes it to incorrectly classify objects depicted in the generated images, lowering the entropy of the conditional probability distribution $\mathrm{p}(y|x)$, i.e. the estimated quality of images and the overall IS of the dataset. For instance, if the evaluated text-to-image model generates a large number of images showing a specific type of animal that the inception network is unfamiliar with, the resulting IS is likely going to be low, even if the images are of high quality and diverse.

This issue with the IS is especially prominent in the CUB-200 dataset, since its class labels consist of 200 different bird species, and the ImageNet dataset only contains 57 class labels related to bird species [31]. Because of this misalignment, the inception network is ideally only able to differentiate 57 different classes when calculating the IS on synthetic images generated by the model trained on the CUB-200 dataset. Evidently, the second mentioned scenario of misalignment in class labels is also true, as ImageNet contains a five times larger number of class labels when compared against the number of CUB-200 dataset classes, resulting in a low estimation of diversity in generated images. Lastly, we point out that out of all ImageNet images labeled as a type of bird, 7% are incorrectly annotated, decreasing the accuracy of the inception network and the overall IS [31]. Considering these observations, we conclude that the IS is not the preferred metric when dealing with the CUB-200 dataset, and a similar inference can be made in the case of the MS-COCO dataset.

Another undesirable property of the IS worth considering, which might have also influenced the results obtained in the testing procedure of our models, is the sensitivity of the metric to the version of the Inception network used to calculate the scores. More specifically, minor changes in weights of the inception network have proved to result in extreme changes in the output IS for the identical input dataset. Although the inception networks with slightly altered weights produce almost exactly the same classification accuracies when run on a validation dataset, the difference in the computed IS values can reach up to 11.5% [29]. Even if all of the models being compared use the same version of the inception network in the testing procedure, which is in our case the Inception V3 architecture, the mere difference in the implementation, i.e. framework used to evaluate the generative model, can lead to a significant difference in the obtained results [32].

Next, we review the quality of images created by the generator models trained on the CUB-200 and MS-COCO datasets and asses their semantic consistency with the provided textual descriptions. As the generator models trained on the ImageNet dataset didn't converge, and the images created using these models seem distorted, we don't include them in the qualitative evaluation.

Figure 1 shows several images generated by the generator model trained on the CUB-200 dataset. More specifically, the images were created using the generator model, which consisted of four DF Blocks, followed by three SSACN Blocks. The images in each column correspond to the same textual description, which is displayed in above the column.

The generated images are mostly semantically consistent with the given description. All three birds depicted in the first column have a yellow breast and belly and a relatively small bill. The birds look realistic, and the fine-grained details, such as the eyes, wings and bill, are outlined accurately. The birds in the second column also match the description well – all three have a white and grey body and a black head. The shape of the birds is natural and realistic. Lastly, the birds in the third column are of the right color and have a pointy bill, which is consistent with the description.

Although the images seem authentic and the portrayed birds all have adequate proportions and shape, some imperfections can be noticed. One that is common to most of the displayed images is the inaccuracy in depicting the bird's feet and talons. These parts of the images are often blurry. An example of this can be seen in the third image of the second column, where the branch that the bird is standing on appears distorted.

Figure 2 shows images created by the generator trained on the MS-COCO dataset. Same as before, the generator model consisted of four DF Blocks, followed by three SSACN Blocks.

In the first column, the images of pizzas are all consistent with the description given. The depicted pizzas appear large and are covered in toppings. Also, apart from the slightly odd shape of the pizza in the second image, the images are of high quality and seem authentic. In the second column, the giraffes in the images also appear realistic. The long legs and neck, along with the skin patterns of dark brown spots and bright stripes, make the giraffes easily recognizable. In all three images, the trees are visible in the background, as the caption suggests. The images in the third column clearly show a person skiing. All three are of high quality and realistic, and fine-grained details such as the jacket of the person in the third image are depicted accurately. With a clear line between the snow and the sky, or forest in the case of

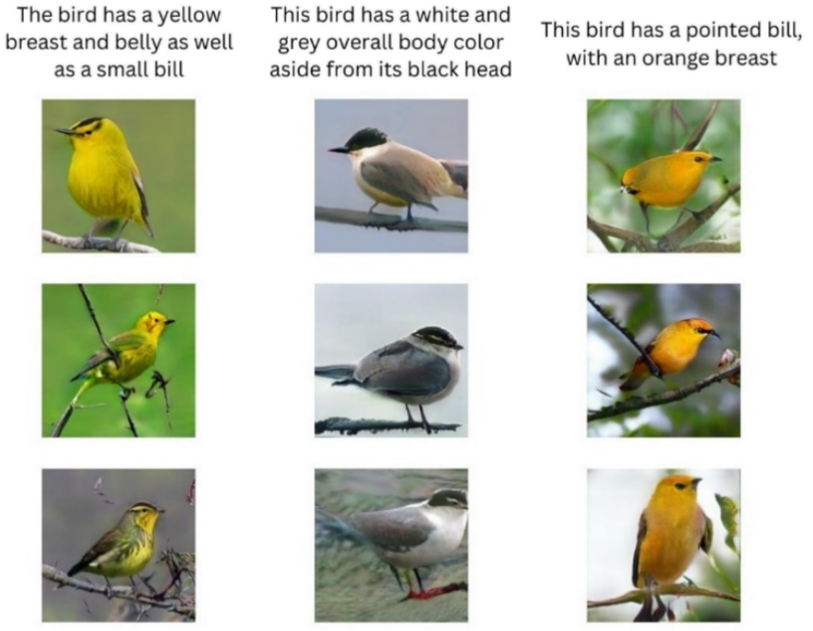

Figure 1. Images generated by the generator model developed in this study, trained on the CUB-200 dataset [8]

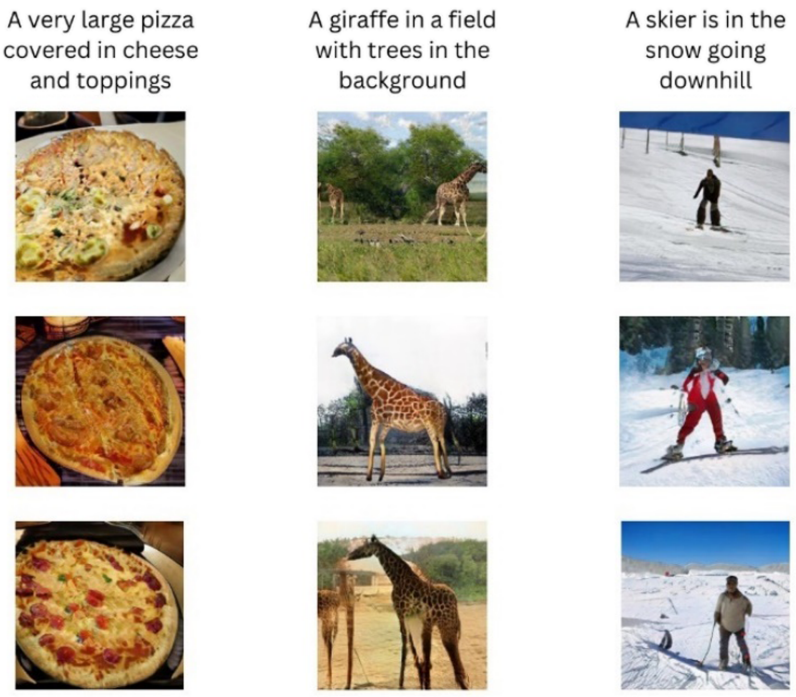

Figure 2. Images generated by the same model architecture, trained on the MS-COCO dataset [9]

the second image, the background of these images is also outlined very well. The images even show details such as ski traces and small bumps on the snow surface.

As for the imperfections in the images generated, a similar problem as in the case of CUB-200 images occurs. More specifically, the legs of the giraffes appear distorted, which is most visible in the third image of the second column. This flaw can also be seen in the first two images of the third column, where the arms and legs of the people skiing have odd shapes. The reason for this inability of our models to accurately portray arms and legs could be the complex structure of these body parts. The multiple joints in arms and legs enable a large variety of different positions and gestures, making it challenging for the generator model to correctly recognize a pattern. Even if the proportions of the arms and legs are portrayed correctly, their positioning seems unnatural and off-balance in most cases. This issue could possibly be resolved by expanding our datasets with diverse images of arms and legs in various positions [33].

5. Conclusion

In this work [1], we examined several different text-to-image architectures based on GANs and transformers. Using fully connected layers to compress the text embeddings into parameters of the standard affine transformation and applying channel-wise scaling and shifting operations, we achieved deep text-to-image fusion in the generator models. We also used a modified, semantic-spatial aware version of the affine transformation, which predicts a spatial mask in order to only fuse textual features into text-relevant subregions of the feature maps. By combining these two text-to-image fusion techniques, we constructed four different generator network architectures. The discriminator network, which is the same for all four generator network architectures, is a convolutional neural network that classifies input images as real or

fake. In the GAN training process, the generator and discriminator models compete in a minmax optimization problem, where the objective of the generator is to model the distribution of real images, and the objective of the discriminator is to learn to separate real images from fake ones.

We trained four different GAN configurations on three large text-to-image datasets – CUB-200, MS-COCO and ImageNet – resulting in a total of 12 different generator-discriminator model pairs. Using the IS and FID, we evaluated the quality and variety of images created with the trained generator models, determined the generator architecture that proved to be most efficient and compared the achieved results to some of the other existing research efforts in the field of text-to-image synthesis. Our generator model consisted of four DF Blocks, followed by three SSACN Blocks and achieved a decent IS score and a better FID score than several other existing GAN-based text-to-image synthesis systems. Outperforming models such as AttnGAN, MirrorGAN and DF-GAN, our model proved efficient in generating diverse and high-quality images that are semantically consistent with the provided captions.

Our methodology has promising applications in fields requiring high-fidelity, text-guided image synthesis. For example, in digital content creation, our model can streamline the production of tailored visuals for advertising [34]. In education, it can generate illustrative diagrams from textual descriptions [35], enhancing learning materials [36]. Additionally, in virtual reality, our approach supports the creation of immersive environments by generating contextually relevant visuals, leveraging the demonstrated semantic consistency.

Generative AI

Generative AI (ChatGPT developed by OpenAI) was used in the writing process to improve the readability and language of the manuscript.

Note

1. https://github.com/marinabagic/ImageGenerationAI

References

1. Šimić A, Babac MB. Artificial intelligence in classifying and creating art: a survey. Int J Stud Prj Rep. 2024; 2(1): 59-89. doi: 10.1504/ijspr.2024.137964.
2. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, *et al.* Generative adversarial nets. In Ghahramani Z, Welling M, Cortes C, Lawrence N, Weinberger KQ (Eds) Advances in Neural Information Processing Systems. [Internet]. Curran Associates; 2014. Available from: https://proceedings.neurips.cc/paper_files/paper/2014/file/f033ed80deb0234979a61f95710dbe25-Paper.pdf
3. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, *et al.* Attention is all you need. In Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S (Eds) *et al.*: Advances in Neural Information Processing Systems. [Internet]. Curran Associates; 2017. Available from: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
4. Bengio Y, Mesnil G, Dauphin Y, Rifai S. Better mixing via deep representations. [Internet], arXiv; 2012. Available from: http://arxiv.org/abs/1207.4404 [accessed 10 August 2025].
5. Tao M, Tang H, Wu F, Jing X, Bao BK, Xu C. DF-GAN: a simple and effective baseline for text-to-image synthesis. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). [Internet]. New Orleans, LA: IEEE; 2022. p. 16494–504. Available from: https://ieeexplore.ieee.org/document/9879122/ [accessed 10 August 2025].
6. Liao W, Hu K, Yang MY, Rosenhahn B. Text to image generation with semantic-spatial aware GAN. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022. p. 18187-96.

7. Chen X, Fang H, Lin TY, Vedantam R, Gupta S, Dollar P, *et al*. Microsoft COCO captions: data collection and evaluation server. [Internet]. arXiv; 2015. Available from: http://arxiv.org/abs/1504.00325 [accessed 10 August 2025].
8. Wah C, Branson S, Welinder P, Perona P, Belongie S. California institute of technology; 2011. Report No.: CNS-TR-2011-001.
9. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, *et al*. Microsoft COCO: common objects in context. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T (Eds) Computer vision – ECCV 2014. Lecture notes in computer science. [Internet]. Cham: Springer International Publishing 2014. p. 740–55; Vol. 8693. Available from: http://link.springer.com/10.1007/978-3-319-10602-1_48 [accessed 10 August 2025].
10. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. [Internet]. Miami, FL: IEEE; 2009. p. 248–55. Available from: https://ieeexplore.ieee.org/document/5206848/ [accessed 10 August 2025].
11. Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X. Improved techniques for training GANs. [Internet]. arXiv; 2016. Available from: https://arxiv.org/abs/1606.03498 [accessed 10 August 2025].
12. Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. GANs trained by a two time-scale update rule converge to a local nash equilibrium; 2017. Available from: https://arxiv.org/abs/1706.08500 [accessed 10 August 2025].
13. Hertzmann A, Jacobs CE, Oliver N, Curless B, Salesin DH. Image analogies. In: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques. [Internet]. ACM; 2001. p. 327–40. Available from: https://dl.acm.org/doi/10.1145/383259.383295 [accessed 10 August 2025].
14. Mordvintsev A, Olah C, Tyka M. Deepdream-a code example for visualizing neural networks. Google Research. 2015; 2(5): 67.
15. Gatys LA, Ecker AS, Bethge M. Image style transfer using convolutional neural networks. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). [Internet]. Las Vegas, NV: IEEE; 2016. p. 2414–23. Available from: http://ieeexplore.ieee.org/document/7780634/ [accessed 10 August 2025].
16. Ulyanov D, Lebedev V, Andrea, LV. Texture networks: feed-forward synthesis of textures and stylized images. In: Balcan MF, Weinberger KQ (Eds) Proceedings of the 33rd International Conference on Machine Learning. [Internet]. New York, NY: PMLR; 2016. p. 1349–57. Proceedings of Machine Learning Research; Vol. 48. Available from: https://proceedings.mlr.press/v48/ulyanov16.html
17. Zhu JY, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: 2017 IEEE International Conference on Computer Vision (ICCV). [Internet]. Venice: IEEE; 2017. p. 2242–51. Available from: http://ieeexplore.ieee.org/document/8237506/ [accessed 10 August 2025].
18. Elgammal A, Liu B, Elhoseiny M, Mazzone M. CAN: creative adversarial networks, generating 'art' by learning about styles and deviating from style norms. [Internet]. arXiv; 2017. Available from: http://arxiv.org/abs/1706.07068 [accessed 10 August 2025].
19. Brock A, Donahue J, Simonyan K. Large scale GAN training for high fidelity natural image synthesis. [Internet]. arXiv; 2019. Available from: http://arxiv.org/abs/1809.11096 [accessed 10 August 2025].
20. Reed S, Akata Z, Yan X, Logeswaran L, Schiele B, Lee H. Generative adversarial text to image synthesis. In: Balcan MF, Weinberger KQ (Eds). Proceedings of the 33rd International Conference on Machine Learning. [Internet]. New York, NY: PMLR; 2016. p. 1060–9. Proceedings of Machine Learning Research; Vol. 48. Available from: https://proceedings.mlr.press/v48/reed16.html
21. Zhang H, Xu T, Li H, Zhang S, Wang X, Huang X, Metaxas DN. StackGAN++: realistic image synthesis with stacked generative adversarial networks. IEEE Trans Pattern Anal Mach Intell. 2019; 41(8): 1947-62. doi: 10.1109/tpami.2018.2856256.

22. Ramesh A, Dhariwal P, Nichol A, Chu C, Chen M. Hierarchical text-conditional image generation with CLIP latents. [Internet]. arXiv; 2022. Available from: http://arxiv.org/abs/2204.06125 [accessed 10 August 2025].
23. Nichol A, Dhariwal P, Ramesh A, Shyam P, Mishkin P, McGrew B, *et al.* GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. [Internet]. arXiv; 2021. Available from: https://arxiv.org/abs/2112.10741 [accessed 10 August 2025].
24. Saharia C, Chan W, Saxena S, Li L, Whang J, Denton E, *et al.* Photorealistic text-to-image diffusion models with deep language understanding. [Internet]. arXiv; 2022. Available from: http://arxiv.org/abs/2205.11487 [accessed 10 August 2025].
25. Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-resolution image synthesis with latent diffusion models. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). [Internet]. New Orleans, LA: IEEE; 2022. p. 10674–85. Available from: https://ieeexplore.ieee.org/document/9878449/ [accessed 10 August 2025].
26. Borji A. Pros and cons of GAN evaluation measures: new developments. Comp Vis Image Understanding. 2022; 215: 103329. doi: 10.1016/j.cviu.2021.103329.
27. Xu T, Zhang P, Huang Q, Zhang H, Gan Z, Huang X, *et al.* AttnGAN: fine-grained text to image generation with attentional generative adversarial networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. [Internet]. Salt Lake City, UT: IEEE; 2018. p. 1316–24. Available from: https://ieeexplore.ieee.org/document/8578241/ [accessed 10 August 2025].
28. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2015. p. 2818-26. doi: 10.1109/cvpr.2016.308.
29. Barratt S, Sharma R. A note on the inception score. [Internet]. arXiv; 2018. Available from: http://arxiv.org/abs/1801.01973 [accessed 10 August 2025].
30. Qiao T, Zhang J, Xu D, Tao D. MirrorGAN: learning text-to-image generation by redescription. [Internet]. arXiv; 2019. p. 1505-14. doi: 10.1109/cvpr.2019.00160. Available from: http://arxiv.org/abs/1903.05854 [accessed 10 August 2025].
31. Luccioni AS, Rolnick D. Bugs in the data: how ImageNet misrepresents biodiversity. [Internet]. arXiv; 2022. Available from: http://arxiv.org/abs/2208.11695 [accessed 10 August 2025].
32. Grgurević A, Bagić Babac M. Transformer-based approach for solving mathematical problems using automatic speech recognition. IEEE Access. 2025; 13: 79845-59. doi: 10.1109/access.2025.3564121.
33. Gezici AHB, Sefer E. Deep transformer-based asset price and direction prediction. IEEE Access. 2024; 12: 24164-78. doi: 10.1109/access.2024.3358452.
34. Tuncer T, Kaya U, Sefer E, Alacam O, Hoser T. Asset price and direction prediction via deep 2D transformer and convolutional neural networks. In: Proceedings of the third ACM international conference on AI in finance. [Internet]. New York NY: ACM; 2022. p. 79–86. Available from: https://dl.acm.org/doi/10.1145/3533271.3561738 [accessed 10 August 2025].
35. Mohanrasu SS, Phan LT, Rajan R, Manavalan B. Cost-sensitive feature selection for multi-label classification: multi-criteria decision-making approach. Appl Comput Inform. [Internet]. 2025. Available from: https://doi.org/10.1108/ACI-09-2024-0353 [accessed 10 August 2025].
36. Ivezić D, Babac MB. Trends and challenges of text-to-image generation: sustainability perspective. Croatian Regional Dev J. 2023; 4(1): 56-77. doi: 10.2478/crdj-2023-0004.

Corresponding author
Marina Bagić Babac can be contacted at: marina.bagic@fer.hr