Evaluating Image Synthesis: A Modest Review of Techniques and Metrics

Roney Nogueira de Sousa
Instituto Federal de Educação,
Ciência e Tecnologia do Ceará
Av. Treze de Maio, 2081 - Benfica,
Fortaleza - CE
Email: roney.nogueira.sousa08@aluno.ifce.edu.br

Saulo Anderson Freitas Oliveira
Instituto Federal de Educação,
Ciência e Tecnologia do Ceará
Av. Treze de Maio, 2081 - Benfica,
Fortaleza - CE
Email: saulo.oliveira@ifce.edu.br

Abstract - *This paper reviews various image synthesis methods, highlighting key techniques such as Convolutional Neural Networks (CNNs), Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and Diffusion Models. We analyze commonly used datasets and evaluation metrics, including SSIM, MS-SSIM, FID, IS, and LPIPS. Our findings show a preference for SSIM in structural quality assessment, while FID and IS are favored for overall quality and diversity. The growing use of LPIPS indicates a shift towards advanced perceptual metrics. This review emphasizes the necessity of combining multiple metrics for a comprehensive evaluation of image synthesis models, aiding future research in the field.*

I. Introduction

Image synthesis is a field of Artificial Intelligence (AI) developed with the intention of generating artificial images from various types of input data, such as text, audio, images, or sketches. This field has garnered increasing interest from the scientific community, spurred by innovations like the introduction of convolutional neural networks. These advancements have enabled the generation of images not only from other images but also from text, sketches, speech, and additional sources [1].

The advancement in the high-performance image synthesis process occurred with the introduction of generative adversarial Networks [2]. These models are composed of two networks: a generator, which tries to create realistic images, and a discriminator, which tries to distinguish between real and generated images. This competition between the networks results in a continuous improvement in the quality of the generated images.

Another important technique is the variational autoencoder [3], a type of generative model that consists of an encoder and a decoder trained to minimize the reconstruction error between the original data and the encoded-decoded data. Additionally, Diffusion Models [4] are also relevant in the field of image synthesis. These models start by slowly adding random noise to the input through direct diffusion steps, learning to reverse the diffusion process to reconstruct the input from the noise.

The relevance of studying image synthesis lies in its vast potential applications across multiple domains. Scientific research can benefit from synthesized images for simulation and analysis. The ability to generate realistic images from various input types expands the possibilities and utility of AI in these areas, driving further innovation and development.

Evaluating image synthesis models is crucial for several reasons. Firstly, the quality of generated images must meet specific standards to be useful in practical applications. Metrics such as Structural Similarity Index (SSIM), Fréchet Inception Distance (FID), and Inception Score (IS) help quantify the structural accuracy, overall quality, and diversity of the generated images, respectively. Secondly, understanding the strengths and weaknesses of different models allows researchers to make informed decisions about which models to use for specific tasks.

In summary, the field of image synthesis has advanced rapidly thanks to the development of various innovative techniques. These approaches allow for the generation of high-quality images from diverse data sources, expanding the possibilities for applications in various areas such as art, entertainment, healthcare, and scientific research. Rigorous evaluation of these models is essential to ensure their effectiveness and to drive further advancements in image synthesis technologies.

Despite these advances, there remains a critical need to continuously evaluate and improve these models. The relevance of this study lies in its aim to provide a comprehensive evaluation of the different models and metrics used in image synthesis. By assessing these models, we can identify their strengths and weaknesses, which is essential for guiding future research and development. This evaluation is particularly important given the diverse applications of image synthesis, where the quality and reliability of generated images can have significant impacts.

A. Article Structure

This article presents the following sections:

Section III: This section presents the benchmark datasets popularly used to train various models, with a brief review of their compositions.

Section IV: This section will present the evaluation metrics used to validate the models.

Section V: This section will present an analysis, and discussion, of the results obtained.

Section VI: This section provides the conclusion of the current study, summarizing the main findings and their implications. It will also outline future work to address the limitations identified in this research.

II. Research Methodology

Parsifal[1] is being used as a support tool for conducting the systematic literature review, guiding and implementing the process. Consequently, the following methodology is organized according to the software stages, encompassing: defining objectives; establishing PICOC criteria; formulating research questions; identifying sources or research databases; establishing selection criteria (inclusion or exclusion); data extraction; presentation of results and discussion.

A. PICOC Criteria

The Parsifal tool incorporates the PICOC method, which is an approach to formulate and refine research questions by integrating five fundamental criteria: Population, Intervention, Comparison, Outcomes, and Context. The delineation of the PICOC criteria is organized as follows:

- Population: Articles relevant to the research topic, available in academic journals or presented at conferences.
- Intervention: Quantitative and qualitative methods used to evaluate the quality of images generated by the models.
- Comparison: Assessing the effectiveness and suitability of different metrics in evaluating image synthesis models.
- Outcomes: Determining which metrics are most accurate and representative of the quality of generated images, and how their selection influences the development and refinement of image synthesis models.
- Context: Applications in academic research, including areas such as computer vision.

B. Research Questions

The formulated research questions are directly related to evaluation metrics in image synthesis models. Table I shows the research questions along with their objectives.

C. Search Key and Research Databases

The search for articles was guided by a carefully crafted search key, aiming to specifically cover the relevant topics for this systematic review. The search key used was the following:

- **(“image synthesis” OR “image generation” OR “synthetic images”) AND (“evaluation metrics” OR “objective metrics” OR “automatic evaluation” OR “performance metrics” OR “automated evaluation metrics” OR “image quality metrics”)**

[1](https://parsif.al)

This key was crafted by combining terms relevant to the research scope. The use of logical operators like “AND” allowed the inclusion of multiple aspects, ensuring that the retrieved articles simultaneously addressed image synthesis, evaluation metrics, and other elements related to the generation and quality of synthetic images.

The search was conducted on the following platforms: IEEE Digital Library, Google Scholar, CAPES Journals, and Scopus. This multi-database approach aims to ensure broad coverage, encompassing relevant journals and conferences in the areas of interest.

D. Inclusion and Exclusion Criteria

After the initial search phase using the search key in the selected databases, the process of classifying the identified studies was carried out, following the previously established inclusion and exclusion criteria. These criteria were essential to ensure the selection of relevant studies and the exclusion of those that did not meet the specific requirements of the research scope.

The inclusion criteria (IC), presented in Table II, were defined to identify studies with specific characteristics relevant to the research scope.

On the other hand, the exclusion criteria (EC), detailed in Table III, were determined to eliminate studies that did not meet the desired requirements or presented specific limitations.

These inclusion and exclusion criteria were applied during the analysis of the search results, ensuring the relevance of the selected studies for the next phase of the systematic review.

E. Quality Assessment

To ensure the validity and relevance of the studies included in the review, a quality assessment checklist was employed using the Parsifal tool. This checklist provides a systematic framework for assessing the methodological quality of the selected studies, ensuring that only robust and reliable research is considered in the final analysis. Table IV shows how this checklist was developed.

The responses to the criteria were categorized as “Yes”, “Partially”, or “No”, with respective weights of 1.0, 0.5, and 0.0. The maximum quality score is 10.0, calculated based on the number of questions and the highest weighted response. For a study to be considered of sufficient quality for inclusion in the review, it must achieve a minimum score of 7.

III. Databases

This section presents the most commonly used datasets in the field of image generation.

ImageNet [2]: A large-scale database consisting of over 14 million images, including 1,034,908 human body images annotated with bounding boxes.

COCO val2014 dataset [3]: A dataset used for segmentation, object detection, keypoint detection, and captioning. The dataset has various features instantiated in 328,000 images.

[2]https://image-net.org/

[3]https://cocodataset.org/#home

TABLE I
RESEARCH QUESTIONS AND THEIR OBJECTIVES

Research Question	Objective
Q1	What are the most used metrics to evaluate the quality of images generated by artificial intelligence models?
Q2	How do different quality evaluation metrics compare in terms of accuracy and reliability when evaluating generated images?
Q3	Are there significant differences in the applicability of quality evaluation metrics between different types of images?
Q4	How have image evaluation metrics evolved over time in response to advances in image generation models?

TABLE II
INCLUSION CRITERIA

ID	Inclusion Criteria (IC)
IC1	Articles written in Portuguese or English
IC2	Studies discussing or using automatic evaluation metrics
IC3	Studies published in the last 5 years to ensure the review covers the most current technologies and methods.
IC4	Studies involving any AI model capable of generating images
IC5	Studies using automatic and objective metrics for evaluating the quality of images generated by AI models.
IC6	Original research articles published in peer-reviewed journals or conferences.

TABLE III
EXCLUSION CRITERIA

ID	Exclusion Criteria (EC)
EC1	Case studies with no applicability or generalization beyond the specific context studied.
EC2	Studies published more than 5 years ago unless they are of historical significance to the field.
EC3	Articles not subjected to peer review.
EC4	Articles for which the full text is not accessible or requires payment.
EC5	Studies focusing on applications unrelated to image generation.
EC6	Studies published in languages other than English or Portuguese.
EC7	Studies that do not clearly specify the methodologies used to apply or evaluate image quality metrics.
EC8	Articles focusing exclusively on subjective evaluations of image quality without objective analysis.

TABLE IV
QUALITY ASSESSMENT CRITERIA

ID	Question
Q1	Are the study objectives clearly defined and specific?
Q2	Are the metrics used to evaluate image quality clearly defined and justified?
Q3	Are details provided on how the metrics are calculated and interpreted?
Q4	Does the study discuss the validity and reliability of the metrics used?
Q5	Does the study specify the data sources used to train and test the models?
Q6	Are the limitations of the data, such as bias or sample size, mentioned?
Q7	Are the statistical analysis techniques used appropriate for the data type and study objective?
Q8	Does the discussion contextualize the results within the field of AI image generation?
Q9	Are the results presented clearly and in detail?
Q10	Does the study address the generalization of the results to different types of images or usage conditions?

Market-1501 [4]: A dataset for person identification, containing 32,668 annotated bounding boxes of 1501 individuals.

DeepFashion [5]: A large-scale clothing database containing over 800,000 images. Each image in this database is labeled with 50 categories and 1000 attributes.

CelebA [6]: A facial attribute database containing more than 200,000 images of celebrities, each with 40 attribute annotations.

CIFAR-10 [7]: This dataset contains more than 60,000 images organized into 10 classes: automobile, airplane, deer, bird, cat, dog, frog, truck, ship, and horse.

CUB 200 [8]: One of the most used datasets for fine-grained visual categorization tasks. This database contains 11,788 images of 200 bird subcategories.

Oxford 102 flower [9]: A collection of 102 flower categories commonly found in the United Kingdom, containing between 40 and 258 images per category.

MNIST [10]: Widely used to train various image processing systems. It contains over 70,000 images of handwritten digits.

Omniglot [11]: A database of handwritten characters, containing 1,623 different handwritten characters collected from 50 different alphabets.

VGG-Face [12]: A facial identity recognition database containing over 2,622 identities and consisting of more than 2.6

[4] https://paperswithcode.com/dataset/market-1501
[5] https://mmlab.ie.cuhk.edu.hk/projects/DeepFashion.html
[6] https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html
[7] https://www.cs.toronto.edu/~kriz/cifar.html
[8] https://www.tensorflow.org/datasets/catalog/caltech_birds2011?hl=en

[9] https://www.tensorflow.org/datasets/catalog/oxford_flowers102?hl=en
[10] https://www.tensorflow.org/datasets/catalog/mnist?hl=en
[11] https://github.com/brendenlake/omniglot
[12] https://paperswithcode.com/dataset/vgg-face-1

million images.

IV. Evaluation Metrics

To evaluate the performance of image synthesis models, both qualitative and quantitative metrics are used:

- **Qualitative Metrics**: Based on user observations and preferences, focusing on the quality and correspondence of generated images to human perception. These assessments can vary between individuals, be time-consuming, and challenging to find suitable participants.
- **Quantitative Metrics**: Using statistics to evaluate the model provides a more robust and reliable assessment, using numerical values to measure the quality and effectiveness of models, eliminating the subjectivity inherent in qualitative assessments.

Proposed by Salimans et al. in 2016, the IS [5] is a quantitative evaluation metric used to measure both the quality and diversity of generated images. A good model should be able to generate high-quality images with great variety. This metric is defined by Equation 1.

$$\text{IS} = \exp\left(\mathbb{E}_{\mathbf{x}\sim p_g}\left[D_{KL}\left(p(y|\mathbf{x}) \parallel p(y)\right)\right]\right) \tag{1}$$

In this formula, $p(y|\mathbf{x})$ is the distribution of the classification of the generated images $\mathbf{x}$, and $p(y)$ is the marginal distribution $p(y) = \int p(y|\mathbf{x})p_g(\mathbf{x})d\mathbf{x}$. The IS uses the Kullback-Leibler (KL) divergence to measure how much the distribution of the generated classes diverges from the marginal distribution, encouraging the production of images that are both distinct and realistic. As stated by Barratt and Sharma, the introduction of IS aims to capture two important qualities of a generative model [6]:

- **Image Clarity**: Generated images should present distinct and sharp objects, meaning the entropy of $p(y|x)$ should be low.
- **Image Diversity**: The generative model should produce a wide variety of images covering all classes of ImageNet, indicating that the entropy of $p(y)$ should be high.

When a generative model meets both conditions, a high Kullback-Leibler (KL) divergence between the distributions $p(y)$ and $p(y|x)$ is expected, resulting in a high IS value.

According to Salimans et al. and Betzalel et al., although IS correlates with human evaluations of image quality, it has limitations [5], [7]. For example, since IS only considers the generated images and does not compare them with real ones, it does not adequately assess the generator's effectiveness. Additionally, IS does not indicate how well the generated images correspond to the provided input.

A commonly used quantitative measure for assessing image synthesis model quality is the FID [8]. This metric considers not only the generated images but also the real ones, calculating the distance between the distribution of features extracted from the generated images, $p_g(x)$, and from the real images, $p_{real}(x)$. The formula for FID is presented in Equation 2:

$$\text{FID} = \|\mu_{real} - \mu_g\|^2 + \text{Tr}(\Sigma_{real} + \Sigma_g - 2(\Sigma_{real}\Sigma_g)^{1/2}) \tag{2}$$

In this equation, μ_{real} and μ_g are the means of the features of the real and generated images, respectively. Σ_{real} and Σ_g are the covariance matrices of the features of the real and generated images, respectively. Tr denotes the trace operation of a matrix. However, as noted by Salimans et al. and Betzalel et al. [5], [7], since the distance between generated and real images depends on extracted features which can be affected by artifacts, the result can be impacted even by a small artifact in the feature space.

The Multi-Scale Structural Similarity (MS-SSIM) [9] is designed to evaluate the quality of generated images by comparing them with real images to measure their similarity. The basic principle of MS-SSIM is that the human visual system is effective at perceiving structural information in the environment, thus measuring the structural similarity between two images can be a way to assess their visual quality. The MS-SSIM metric value ranges from 0 to 1, with values closer to 1 indicating greater perceptual similarity between the compared images. Equation 3 shows the formula used for calculating MS-SSIM.

$$\text{MS-SSIM}(x, y) = [l_M(x, y)]^{\alpha_M} \prod_{j=1}^{M} [c_j(x, y)]^{\beta_j} [s_j(x, y)]^{\gamma_j} \tag{3}$$

In this equation, $l_M(x, y)$ is the luminance comparison at the highest scale, $c_j(x, y)$ and $s_j(x, y)$ are the contrast and structure comparisons at scale j. The α_M, β_j, and γ_j are the weights applied at each scale j and M represents the total number of scales.

MS-SSIM is an enhanced version of the SSIM [10], which measures the similarity between two images at multiple scales through successive downsampling steps. This process allows for the incorporation of details at different resolutions. Starting with the calculation of contrast and structural comparisons, iteratively, a low-pass filter is applied, and the image resolution is reduced by a factor of 2 after each application.

Like the other evaluation metrics mentioned, MS-SSIM and SSIM also have their limitations, such as being computationally more intensive than pixel-based metrics, and their performance can vary depending on the specific content of the image and application [11].

Another notable objective metric is the *Learned Perceptual Image Patch Similarity* (LPIPS) [12]. This metric aims to replicate human judgment on the similarity between two images, measuring the differences between the generated image and the corresponding real image. LPIPS calculates these differences in terms of visual features extracted from a pre-trained neural network. Regarding LPIPS values, higher values indicate greater similarity between the generated image and the real image.

V. Results Analysis

In conducting this literature review, several studies on image generation methods were selected. Table V summarizes these studies, providing a comprehensive view of the different approaches and resources used in image generation, allowing for a comparison between the approaches in the field.

TABLE V
SELECTED STUDIES FOR LITERATURE REVIEW

Author	Year	Input Data Type	Dataset
[13]	2019	Image	DeepFashion
[14]	2019	Image	MNIST, Flower
[15]	2020	Sketch	CelebA
[16]	2020	Sketch	ShoeV2, ChairV2
[17]	2020	Text	CUB 200, Oxford102
[18]	2020	Speech	CUB200, Oxford102
[19]	2020	Image	CelebA
[20]	2020	Sketch	Sketchy, ImageNet
[21]	2020	Text	COCO, MNIST
[22]	2020	Image	Market1501, DeepFashion
[23]	2021	Image	DeepFashion
[24]	2021	Text	CUB 200, Oxford102
[25]	2021	Image	MNIST, Omniglot, VGG-Face
[26]	2021	Speech	CUB200, Oxford102
[27]	2021	Text	COCO

Table VI provides an overview of the evaluation metrics applied in the reviewed articles. It shows that each study is evaluated using one or more of these metrics, reflecting the methodological diversity and different approaches adopted in current literature. The listed metrics contribute to the evaluation of the generated image quality in different ways, as discussed in Section IV.

TABLE VI
EVALUATION METRICS APPLIED IN THE STUDIES

Author	SSIM	MS-SSIM	FID	IS	LPIPS
[13]	x	x		x	x
[14]					x
[15]	x		x	x	
[16]			x		x
[17]				x	
[18]			x	x	
[19]	x		x	x	
[20]			x	x	
[21]	x				
[22]	x			x	
[23]			x		x
[24]	x				
[25]			x	x	x
[26]			x	x	
[27]			x		x

Table VI highlights how different studies have prioritized various aspects of image quality. For example, some studies focused on structural similarity (SSIM and MS-SSIM), while others worked with global quality and element diversity assessments (FID and IS).

A. *Discussion of Results*

In this section, we discuss the implications of the results presented in tables V and VI, evaluating the approaches adopted by the different studies and their evaluation metrics.

The studies analyzed indicate a significant diversity in image generation techniques, ranging from sketches and text to images and speech as input data. The variety of input data reflects the flexibility and comprehensiveness of contemporary image synthesis methods, which seek to simulate the human ability to create images from various forms of representation. We note that the most widely used databases, such as DeepFashion, MNIST, CelebA, CUB200 and Oxford102, provide a wide spectrum of challenges for image synthesis models, contributing to the robustness of the developed methods.

The analysis of the evaluation metrics reveals a considerable preference for SSIM for the evaluation of the structural quality of the generated images, present in seven of the fifteen studies analyzed. The choice of SSIM can be attributed to its ability to capture important information about the luminance, contrast, and structure of the images, crucial elements for the human perception of visual quality.

On the other hand, the MS-SSIM metric, an extension of SSIM that incorporates evaluation at multiple scales, was used only once. The low adoption of MS-SSIM may be due to its additional complexity and greater difficulty of interpretation, despite its potential superiority in providing a more detailed and comprehensive analysis of structural quality at different levels of detail.

The FID metric is present in ten of the fifteen studies. This metric is particularly effective in identifying subtle discrepancies that may not be captured by direct similarity-based metrics such as SSIM.

Furthermore, IS was used in nine of the studies, often in conjunction with FID, thus reflecting a complementary approach, where researchers seek an assessment considering both the quality of the individual images and the diversity of the generated set.

The use of LPIPS in seven of the studies analyzed indicates a trend toward adopting more sophisticated perceptual metrics. LPIPS, unlike traditional metrics, learns directly from human perception, providing an assessment more aligned with how humans perceive image quality. Its inclusion suggests that researchers are increasingly interested in understanding how image synthesis methods perform in terms of perceptual quality, beyond purely technical assessments.

A critical aspect to consider is the combination of different metrics to obtain a more robust assessment. The analysis of the studies suggests that no single metric is sufficient to assess the quality of the generated images. For example, while SSIM and FID provide data on structural similarity and global quality, IS and LPIPS offer insights into diversity and human perception.

However, there are important limitations to be acknowledged. For example, variability in the datasets used can significantly influence evaluation results. Databases such as MNIST and CelebA have very different characteristics, and the effectiveness of a model can vary dramatically depending on the dataset used. These differences can lead to variations in how models perform across these datasets, highlighting the need to consider dataset-specific factors when interpreting validation metrics. As a result, conclusions drawn from one dataset may not be directly applicable to another without a consideration of these underlying differences.

VI. Conclusion and Future Work

In this paper, a brief literature review on image synthesis methods was conducted, examining various approaches, evaluation metrics, and commonly used datasets in this field.

The analysis of the use of metrics to evaluate model quality revealed a significant preference for SSIM in assessing the structural quality of images, while FID and IS were used to measure the overall quality and diversity of generated images. The adoption of LPIPS highlighted a growing trend towards using perceptual metrics, aligned with human perception.

The results of this review suggest that an approach combining multiple evaluation metrics is essential for understanding the quality of generated images. Allowing researchers to evaluate images from multiple perspectives, providing a more comprehensive view of the effectiveness of the methods.

While this study has provided insight into the current state of the image synthesis field and its evaluation metrics, other areas remain under investigation. Future work will focus on:

1) **Expanding the Dataset**. We plan to include a wider variety of datasets to better understand how different models perform across various types of input data.
2) **Exploring New Metrics**. Exploration of additional perceptual metrics to complement SSIM, FID, and IS, providing a more holistic evaluation of image quality.
3) **Longitudinal Studies**. Conducting longitudinal studies to observe how model performance evolves over time with continuous training and adaptation to new data.

Acknowledgments

The authors would like to acknowledge the financial support from the Coordination for the Improvement of Higher Education Personnel (CAPES — Funding Code 001) and the Federal Institute of Education, Science and Technology of Ceará (IFCE).

References

[1] M. Elasri, O. Elharrouss, S. Al-Maadeed, and H. Tairi, "Image generation: A review," *Neural Processing Letters*, vol. 54, no. 5, pp. 4609–4646, 2022.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[3] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[4] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International conference on machine learning*. PMLR, 2015, pp. 2256–2265.

[5] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," *Advances in neural information processing systems*, vol. 29, 2016.

[6] S. Barratt and R. Sharma, "A note on the inception score," *arXiv preprint arXiv:1801.01973*, 2018.

[7] E. Betzalel, C. Penso, A. Navon, and E. Fetaya, "A study on the evaluation of generative models," *arXiv preprint arXiv:2206.10935*, 2022.

[8] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[9] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2. Ieee, 2003, pp. 1398–1402.

[10] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[11] M. Prodan, G. V. Vlăsceanu, and C.-A. Boiangiu, "Comprehensive evaluation of metrics for image resemblance," *Journal of Information Systems & Operations Management*, vol. 17, no. 1, pp. 161–185, 2023.

[12] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[13] A. Grigorev, A. Sevastopolsky, A. Vakhitov, and V. Lempitsky, "Coordinate-based texture inpainting for pose-guided human image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 135–12 144.

[14] M. Zhai, L. Chen, F. Tung, J. He, M. Nawhal, and G. Mori, "Lifelong gan: Continual learning for conditional image generation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 2759–2768.

[15] U. Osahor, H. Kazemi, A. Dabouei, and N. Nasrabadi, "Quality guided sketch-to-photo image synthesis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 820–821.

[16] R. Liu, Q. Yu, and S. X. Yu, "Unsupervised sketch to photo synthesis," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*. Springer, 2020, pp. 36–52.

[17] L. Li, Y. Sun, F. Hu, T. Zhou, X. Xi, and J. Ren, "Text to realistic image generation with attentional concatenation generative adversarial networks," *Discrete Dynamics in Nature and Society*, vol. 2020, no. 1, p. 6452536, 2020.

[18] J. Li, X. Zhang, C. Jia, J. Xu, L. Zhang, Y. Wang, S. Ma, and W. Gao, "Direct speech-to-image translation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 3, pp. 517–529, 2020.

[19] H. Li and J. Tang, "Dairy goat image generation based on improved-self-attention generative adversarial networks," *IEEE Access*, vol. 8, pp. 62 448–62 457, 2020.

[20] Z. Li, C. Deng, E. Yang, and D. Tao, "Staged sketch-to-image synthesis via semi-supervised generative adversarial networks," *IEEE Transactions on Multimedia*, vol. 23, pp. 2694–2705, 2020.

[21] T. Zia, S. Arif, S. Murtaza, and M. A. Ullah, "Text-to-image generation with attention based recurrent neural networks," *arXiv preprint arXiv:2001.06658*, 2020.

[22] H. Tang, S. Bai, L. Zhang, P. H. Torr, and N. Sebe, "Xinggan for person image generation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*. Springer, 2020, pp. 717–734.

[23] J. Zhang, K. Li, Y.-K. Lai, and J. Yang, "Pise: Person image synthesis and editing with decoupled gan," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7982–7990.

[24] L. Gao, D. Chen, Z. Zhao, J. Shao, and H. T. Shen, "Lightweight dynamic conditional gan with pyramid attention for text-to-image synthesis," *Pattern Recognition*, vol. 110, p. 107384, 2021.

[25] A. Phaphuangwittayakul, Y. Guo, and F. Ying, "Fast adaptive meta-learning for few-shot image generation," *IEEE Transactions on Multimedia*, vol. 24, pp. 2205–2217, 2021.

[26] X. Wang, T. Qiao, J. Zhu, A. Hanjalic, and O. Scharenborg, "Generating images from spoken descriptions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 850–865, 2021.

[27] H. Zhang, J. Y. Koh, J. Baldridge, H. Lee, and Y. Yang, "Cross-modal contrastive learning for text-to-image generation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 833–842.