IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 4, April 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304062 |

Personalized Visual Content Generation: Cutting-Edge AI Solutions

Chinmay Kamble, Nakul Kamatkar, Dr. Girija Chiddarwar, Ayasha Sikilkar, Neha Marathe

Department of Computer Engineering, Marathwada Mitra Mandal's College of Engineering Pune, Maharashtra, India

ABSTRACT: In the realms of computer vision and deep learning, the generation of images and videos from various data types, including text, scene graphs, and object layouts, represents a frontier of both challenges and breakthroughs. This paper offers, to the best of the authors' knowledge, the first comprehensive overview of existing methods in both image and video generation domains. We explore a range of deep learning-based strategies, emphasizing the nature of algorithms used, types of data employed, and the primary objectives of different techniques. In the image generation sector, we delve into the nuances of diverse methods, presenting a detailed analysis of each category, along with a discussion on existing image generation datasets. In the context of video generation, we focus on recent advancements in generating long-range, realistic video sequences from initial frames, considering various approaches such as Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and Transformer models. This review critically examines and compares the performance of these methods, employing suitable evaluation metrics for each category. By providing a comprehensive comparison, we aim to illuminate the state-of-the-art, pinpointing the limitations and strengths of current solutions. Additionally, the paper addresses current challenges and potential future directions in both image and video generation fields, offering insights into the trajectory of these burgeoning technologies and their potential applications in areas like autonomous robotics and trajectory prediction. This synthesis not only presents a clear picture of the current landscape but also sets the stage for future innovations in the dynamic field of automated image and video generation.

KEYWORDS: Image generation, Text-to-image generation, Sketch-to-image generation, Layout-to-image generation, Image-to-image translation, Panoramic image generation, Video Generation, Deep Learning, Generative Adversarial Networks

I. INTRODUCTION

With the development in various fields, the data analysis has become a real challenge especially the processing of the visual data [1]. In addition, with the development of analysis techniques such as machine and deep learning lead the researcher to find many solutions for some problems that were not doable using traditional and statistic methods. Also, this allows researcher to create many filed of studies, like image generation from different type of data, that represents solution for other existing tasks [2]. This leads to an opportunity of analysing captured or generated data with a high performance [3]. These fields of study especially computer vision tasks, have shot up in need due to the massive amounts of data generated from these equipment's and the Artificial Intelligence (AI) tools [4]. In the AI learning process, there is a strong need for properly labelled data where information can be extracted from the images for multiple inferences [5].

Generating new image from another is one of the challenging tasks in computer vision. With the introduction of deep learning techniques especially using CNN models this task has become doable which attracted the researcher to create new images not just based on other images but also from text, sketch, scene graph, layout or also from a set of scenes. Few years ago, the development of AI techniques including generative adversarial networks (GANs) [7] has revived images generation task to generate new images with a high performance in term of image quality and pertinent content. It has produced realistic images of human face, furniture, or scenes that are difficult to distinguish from real images [8]. Consequently, AI models have been used to create target images using different ways including Imageto- Image Translation, Sketch-to-Image Generation, Conditional Image Generation, Text-to- Image Generation, Video Generation, Panoramic Image Generation, and Scene graph Image Generation. For each category, many papers have been proposed exploiting various features and techniques to reach the best and effective results.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304062 |

Video generation has gained substantial traction as a research topic. Researchers all over the world are working on many different methods to generate long range, high resolution, indistinguishably natural videos. The popularity is due to many critical advantages attached to it. Prominent fields that have great relevance to video generation are reinforcement learning [1] and motion planning of autonomous systems [2]. Although it is an interesting task, it is a very difficult problem to model, owing to its multi-modal nature and the exponentially growing tree of possibilities after the passage of initial frames. Video generation has numerous real-world applications which include interpolation of full-sized videos from cropped videos and scene generation in video games. Further, it can also be utilized for generating videos which can serve as datasets for machine learning models. In robotics, video generation can play a vital role in handling occlusion, predicting object movement [3] and trajectory optimization. Multiple approaches have been proposed in the prior literature and improvement in realism of generated frames is commendable. These approaches are based on pixel transformations [4], modelling the stochastic nature by sampling latent variables from prior networks [5] and using Variational autoencoders (VAEs) [6] which are trained by optimizing the lower bound on likelihood of the data in a latent variable model. However, the main network still relies on pixel-wise MSE loss, which causes the results to be blurry as the latent representations do not sufficiently capture the

II. LITERATURE SURVEY

The realm of texture synthesis has witnessed significant advancements propelled by innovative techniques proposed by researchers. Among these, Markovian Generative Adversarial Networks (MGANs), introduced by Li and Wand [26], stand out for their remarkable effectiveness in generating realistic textures. This technique not only decodes images into lifelike textures but also demonstrates an exceptional capability in transforming brown noise into intricate paintings, thereby elevating the quality of texture synthesis. Complementing this, the Spatial GAN (SGAN) architecture, as suggested by Jetchev et al. [27], emerges as a fitting framework for texture synthesis tasks. Leveraging SGAN, complex textures can be synthesized by amalgamating diverse source images, leading to the creation of high-fidelity texture compositions. A notable aspect of GAN-based texture synthesis lies in its utilization of interpolation techniques to imbue generated textures with realistic details, ensuring a seamless transition in the texture generation process. However, despite these advancements, challenges such as "mode dropping" and convergence issues persist, necessitating further exploration and refinement of texture synthesis models.

A generative adversarial network (SRGAN) for image super-resolution (SR) was proposed by Ledig et al. [28] and has the potential to greatly enhance perceived quality.
In order to create an Enhanced SRGAN (ESRGAN), Wang et al. [29] refined SRGAN. This resulted in more realistic and natural textures as well as an improvement in the issue of artifacts in SRGAN and the visual quality of the generated images.
A novel Spatial Feature Transform (SFT) layer is included in Wang et al.'s [29] approach, SFTGAN, which recovers natural and realistic textures and can produce more realistic and aesthetically beautiful textures. The image super-resolution technique yields the best results by gradually generating artificial images by training generators and discriminators using low-resolution images and adding a higher-resolution network layer each time. It is capable of producing diversified, high-quality photos with high resolution. But it takes a long time to train, and more GPUs are needed.

In order to optimize the job of picture inpainting, Yang et al. [30] suggested a multi-scale neural patch synthesis method based on deep learning that makes use of texture restrictions and image content. The technique is able to retain the high-frequency details in addition to restoring images with semantically reasonable contents. Yeh et al.'s [30] novel method for semantic picture inpainting can produce satisfactory results and attain pixel-level photorealism. An efficient face completeness technique based on a deep generative model was presented by Li et al. [31]. It can recover photos with a significant pixel-missing region and produce a face completion result that is realistic.
These days, image inpainting techniques based on GANs can produce results that are more logical and semantically consistent than those of earlier techniques. At the moment, certain techniques use convolutions to repair pictures using free-form masks, which may repair pictures with several holes or use irregular shapes to fill in the spaces left by missing parts. Nonetheless, the size and placement of the masks have an impact on the inpainting quality of the image.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304062 |

III. APPROACHES

3.1. IMAGE GENERATION

Images generation can be done using different form as input including RGB images, videos, medical images, and text, etc. The output in general is an image or a video. The type of dataset used and the target data expected as output can be a 2D image or also a 3D video. Some researcher worked on 3D data to generate 3D images or objects. For example, to generate images of 3D objects, the authors in [7] proposed a Visual Object Networks (VON) for image generation. This method used silhouette and depth map features as input for the proposed GAN-based model.

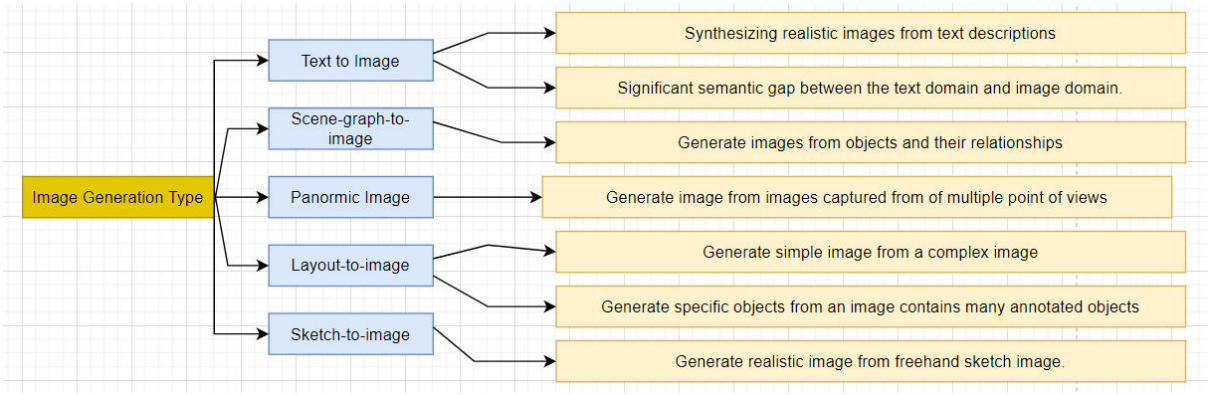

Fig.1 Ways for Image Generation

For another field, some researcher worked on medical images for generating synthetic images. In [8], the authors proposed a GAN-based method on Brain MR images inspired from Deep Convolutional GAN (DCGAN) and Wasserstein GAN (WGAN) architectures to generate multi-sequence brain Magnetic Resonance (MR) from the original ones. For that, Zhao et al. [12] proposed a novel image generation model termed (Vari GANs) to generate a multi-view image from a single view which can be more realistic looking images for commercial purposes.
For blurry images reconstruction, the authors in [21] exploited a multistage Variational Auto-
Encoders (VAE) based model to reconstruct the images

3.1.1. SKETCH-TO-IMAGE GENERATION is the operation of generating realistic images from sketch images. The application of this technique can be used for generate cartoon images, face images from their sketch images, etc. For the same purpose, Liu et al. [10] proposed a new model called auto-painter which can automatically generate compatible colours given a cartoon sketch image exploiting conditional generative adversarial networks (cGAN).

3.1.2. CONDITIONAL IMAGE GENERATION: most of images generation methods used image or video as input to generate or reconstruct new images. While some methods, for many purposes, exploited two or more inputs for the proposed models. These inputs presented as condition to the network to specify the image target.For example, the authors in [13] proposed a conditional image generation method that extract the structure of visual objects from the unlabelled images using appearance and geometry of the objects. In the same context, the authors in [14] exploited variational U-Net for generating synthetic images of objects based on the appearances of the object as well as the mask of the object or silhouette of the person for person generation.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304062 |

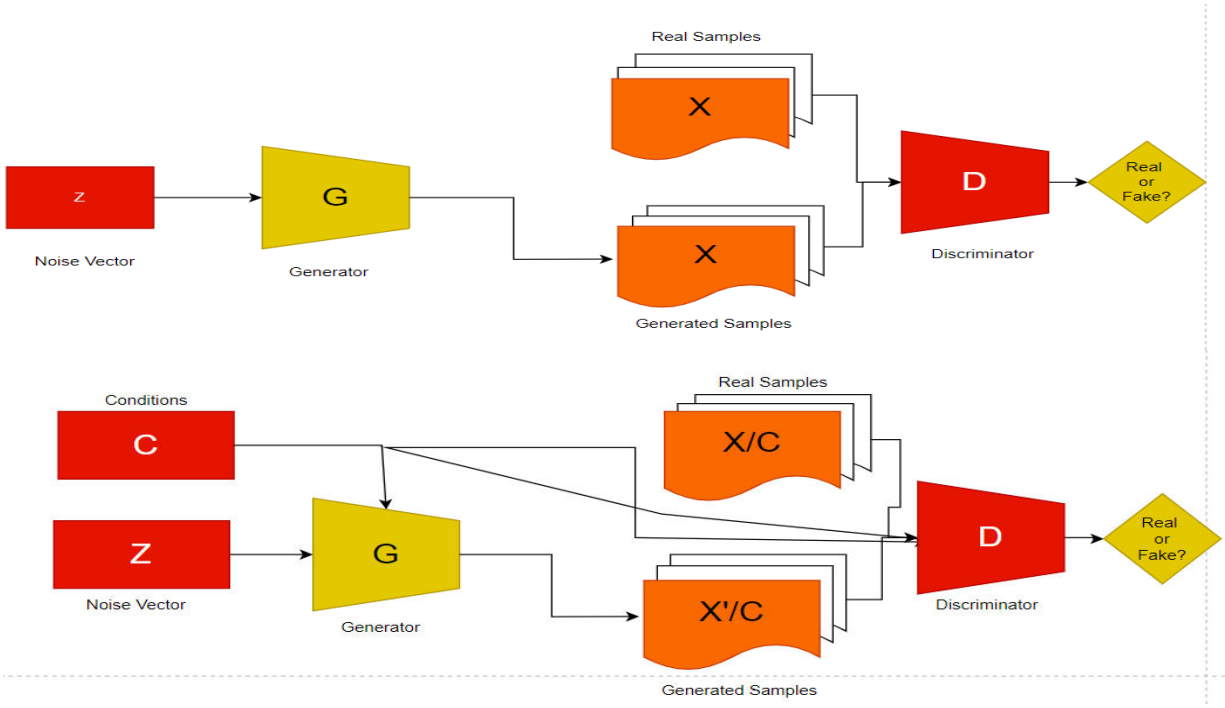

Fig 2 : Conditional Image Generation

3.1.3.TEXT-TO-IMAGE GENERATION is the task of generating images from text and captions. This type of methods is the reverse operation of images captioning. To do that many methods have been proposed. For example, Sharma et al. [11] proposed a new approach that utilizes Vis Dial dialogues along with text descriptions to generate images while the proposed architecture is inspired by Stack GAN. Another method in [18] while the authors proposed an Attentional Generative Adversarial Network (Attn GAN) that can synthesize fine-grained details at different sub-regions of the image by paying attentions to the relevant words in the natural language description.

3.1.4. STABLE DIFFUSION MODEL - The stable diffusion model operates by adding and then removing Gaussian noise from images. It enhances traditional diffusion models with an encoder-decoder and U-Net design, incorporating self-attention and cross-attention mechanisms in its Transformer layers. This model compresses images into a latent space and then decompresses them, with diffusion processes occurring in this latent space.

3.1.5. DATASETS

In this section we presented the most used datasets in image generation field.

MNIST database: is a large used dataset for training various image processing systems. The MNIST database contains more than 70K images of handwritten digits. This dataset used 60,000 images for training and 10,000 images for testing.

ImageNet2: is a large-scale database consists of more than 14,197,122 annotated images according to the WordNet hierarchy, the 1,034,908 images of human body are annotated with bounding box, the dataset is used for "ImageNet Large Scale Visual Recognition Challenge (ILSVRC)".

Microsoft COCO val2014 dataset: is a large-scale segmentation, object detection, key point detection, and captioning database. The dataset has various features instanced in 328K images.

CUB 200 dataset: the Caltech-UCSD Birds 200 (CUB 200) is one of the most used datasets for fine-grained visual categorization task. This database contains 11,788 images of 200 subcategories correlating to birds, 5,994 images for training and 5,794 images for testing.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304062 |

Market-1501 dataset: is a large dataset for person re-identification, containing 32,668 annotated bounding boxes of 1501 identities separate from 751 identities for training and 750 for testing.

3.1.6. RESULT ANALYSIS AND DISCUSSION

In this section, we present the experimental results of the state-of-the-art methods evaluated on standard image generation benchmarks including DeepFashion, Market-1501, MNIST, CelebA, Visual Genome, COCO-Stuff, Cub, ImageNet, MS COCO, CIFAR-10, and Oxford- 102. The accuracy and the efficiency of each method are compared using the common metrics used for evaluating the performance of each method such as Inception Score (IS), Fréchet Inception Distance (FID)

1 EvaluationMetrics
There are several ways to evaluate the performance between real and generated images. In this section, we review some universally-agreed and popularly adopted measures for image generation model evaluation.

1.1 Inception Score (IS)
The Inception Score (IS) is an objective performance metric, used to evaluate the quality of generated images or synthetic images, generated by Generative Adversarial Networks (GANs). It measures how realistic and diverse the output images are.

1.2 Fréchet Inception Distance (FID)
The **Fréchet inception distance (FID)** is a metric used to assess the quality of images created by a generative model, like a generative adversarial network (GAN). Unlike the earlier inception score (IS), which evaluates only the distribution of generated images, the FID compares the distribution of generated images with the distribution of a set of real images ("ground truth")

3.1.7. DATASET EVALUATIONS SUMMARY:

1. **ImageNet**: The self-conditional GAN [7] stands out with the highest Inception Score (IS) of 14.962 and the lowest Fréchet Inception Distance (FID) of 41.76, indicating superior performance in conditional image generation.
2. **MS COCO**: In text-to-image generation, XMC-GAN [17] leads with the highest IS and the only reported FID (9.330), outperforming AttnGAN [18] and ControlGAN [7] in IS by a notable margin.
3. **Market-1501**: For multi-view and pose-guided image generation, PCGAN [4] shows the best results with the highest IS (3.657) and a significant lead in FID over HS+HA [9]. SSIM metric evaluations show close results among methods, with [14], [6], and [7] leading.
4. **MNIST**: In high-quality image generation, Fréchet-GAN [4] attains the lowest FID, with AAAE [4] scoring highest in IS. The SSIM metric shows a wider variance among methods compared to other datasets, with [2] leading.
5. **Cub**: For text-to-image generation, LeicaGAN [3] achieves the top IS (4.62), closely followed by ControlGAN [7] and MirrorGan [5]. IS is the primary metric used for evaluation in this dataset.

3.2. VIDEO GENERATION

APPROACHES

3.2.1. MODELLING PIXEL TRANSFORMATIONS

In the realm of unsupervised video generation, a noteworthy contribution is the approach introduced in [4]. Unlike previous methods that largely relied on predicting short-term changes, focused on small parts of images, or used synthetic data, this approach brings a novel perspective. It centres around the idea of directly manipulating the pixels across video frames, bypassing the need to understand or model the underlying internal state of the input data. This method uses an LSTM (Long Short-Term Memory) network [19] to forecast how pixels should be moved or transformed between successive frames. This technique involves generating a set of transformation kernels and a multi-channel mask, which together determine how pixels in one frame should be rearranged to create the next frame.
The paper details three specific models based on this concept: Dynamic Neural Advection (DNA), Convolutional Dynamic Neural Advection (CDNA), and Spatial Transformer Predictors (STP).Importantly, this approach doesn't

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304062 |

require the model to learn about specific objects in the video, which is a significant advantage for applications in reinforcement learning algorithms [1].

3.2.2. STOCHASTIC VIDEO GENERATION

Stochastic video generation is an intriguing area in video processing that focuses on predicting future frames of a video using random variables that change over time. This method is distinct because it incorporates elements of randomness or uncertainty into the video generation process. Early attempts in this field, like the one in [5], introduced a two-model system: a main network for generating videos and a prior network for adding randomness.

The core architecture of these models includes an LSTM [19], a type of neural network good at understanding sequences (like video frames), situated between an encoder and a decoder. This LSTM generates a range of possible outcomes based on the given input, and the training involves choosing from these outcomes.

3.2.3 TRANSFORMER-BASED APPROACHES

Transformers have greatly impacted areas like language modeling and machine translation due to their proficiency in capturing long-range dependencies. However, their high computational cost is a significant drawback. The Axial Transformer [13] tackles this issue with "Axial Attention," efficiently processing multi-dimensional data along one axis at a time, reducing the need for high computational resources.

The Latent Video Transformer (LVT) [10] introduces a novel approach by working with discrete latent representations of video frames, encoded using the VQ-VAE2 autoencoder.

3.2.4 GENERATIVE ADVERSARIAL NETWORKS

Generative Adversarial Networks (GANs), comprising a Generator (G) and a Discriminator (D), are a key method in video generation. GANs excel in creating realistic videos by balancing texture, spatial consistency, and temporal dynamics.

A further advancement is TrIVD-GAN, which introduces a novel discriminator design for better performance and efficiency. It differs from DVD-GAN-FP by focusing on both per-frame global structure and local spatio-temporal dynamics. The performance of these models is effectively measured using Inception Score (IS) and Fréchet Video Distance (FVD), with TrIVD-GAN showing significant promise in video generation.

3.2.5 RESULTS

This section of the paper provides a comparative analysis of different approaches to video modeling, specifically focusing on their performance on the BAIR Robot Pushing dataset [4]. This dataset features videos of robot arm movements, comprising about 59,000 examples across both training and test sets.

The effectiveness of the various models is evaluated using two main metrics: bits/dim and Fréchet Video Distance (FVD) [9]. These metrics are crucial for assessing the quality of the video generation.

According to the data presented, the Video Transformer model stands out, delivering state-of-the-art results. This high level of performance is attributed to the efficacy of the FVD metric in evaluating video generation tasks. The Video Transformer model's success in this context suggests its superior capability in accurately modeling and generating complex video sequences, particularly in scenarios involving dynamic movements like those of a robot arm.

IV. CONCLUSION

In this review, we have merged the realms of image and video generation, providing a comprehensive overview of their methodologies, challenges, and advancements. For image generation, we explored various approaches categorized by input types such as images, sketches, layouts, and text, emphasizing the importance of large-scale datasets and conditioned generation techniques. Our analysis included a comparison of methods using standard benchmark datasets and evaluation metrics.

In the video generation segment, we focused on advanced models, particularly those evaluated using the BAIR Robot Pushing dataset, discussing their effectiveness, limitations, and the nuances of different generative approaches.

By integrating these two areas, our review offers a concise yet thorough perspective on the current state and future directions of image and video generation technologies. This synthesis is intended to serve as a valuable resource for researchers and practitioners in understanding and advancing these interrelated fields.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304062 |

REFERENCES

(For Image Generation)

1. AkbariY, Almaadeed N, Al-maadeed S, ElharroussO (2021) Applications, databases and open computer vision research from drone videos and images: a survey. Artif Intell Rev 54(5):3887–3938
2. Elharrouss O, Almaadeed N, Al-Maadeed S (2021) A review of video surveillance systems. J Vis Commun Image Represent 77:103116
3. Elharrouss O, Al-Maadeed S, Subramanian N, Ottakath N, Almaadeed N, Himeur Y (2021) Panoptic segmentation: a review. arXiv preprint arXiv:2111.10250
4. MaL, SunQ, Georgoulis S, VanGool L, Schiele B, FritzM(2018) Disentangled person image generation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 99–108
5. ElharroussO, Moujahid D, Elkah S, Tahirih (2016) Moving object detection using a background modeling based on entropy theory and quad-tree decomposition. J Electron Imaging 25(6):061615
6. Maafiri A, Elharrouss O, Rfifi S, Al-Maadeed SA, Chougdali K (2021) DeepWTPCA-L1: a new deep face recognition model based on WTPCA-L1 norm features. IEEE Access 9:65091–65100
7. Zhu J-Y, Zhoutong Z, Chengkai Z, Jiajun W, Antonio T, Josh T, Bill F (2018) Visual object networks: image generation with disentangled 3D representations. Adv Neural Inform Process Syst, pp 118–129
8. Han C, Hayashi H, Rundo L, Araki R, Shimoda W, Muramatsu S, Furukawa Y, Mauri G, Nakayama H (2018) GAN-based synthetic brain MR image generation. In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018), pp 734–738. IEEE
9. Mao X, Wang S, Zheng L, Huang Q (2018) Semantic invariant cross-domain image generation with generative adversarial networks. Neurocomputing 293:55–63
10. Liu Y, Qin Z, Wan T, Luo Z (2018) Auto-painter: Cartoon image generation from sketch by using conditional Wasserstein generative adversarial networks. Neurocomputing 311:78–87
11. Sharma S, Suhubdy D, Michalski V, Kahou SE, Bengio Y (2018) Chatpainter: improving text to image generation using dialogue. arXiv preprint arXiv:1802.08216
12. Zhao B, Wu X, Cheng ZQ, Liu H, Jie Z, Feng J (2018) Multi-view image generation from a single-view. In: Proceedings of the 26th ACM international conference on multimedia, pp 383–391
13. Jakab T, Gupta A, Bilen H, Vedaldi A (2018) Conditional image generation for learning the structure of visual objects. Methods 43:44
14. Esser P, Sutter E, Ommer B (2018) A variational u-net for conditional appearance and shape generation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8857–8866
15. Johnson J, Gupta A, Fei-Fei L (2018) Image generation from scene graphs. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1219–1228
16. Bodla N, Hua G, Chellappa R (2018) Semi-supervised FusedGAN for conditional image generation. In: Proceedings of the European conference on computer vision (ECCV), pp 669–683
17. Siarohin A, Sangineto E, Lathuiliere S, Sebe N (2018) Deformable gans for pose-based human image generation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3408–3416
18. Xu T, Zhang P, Huang Q, Zhang H, Gan Z, Huang X, He X (2018) Attngan: fine-grained text to image generation with attentional generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1316–1324
19. Qian X, Fu Y, Xiang T, Wang W, Qiu J, Wu Y, Xue X (2018) Pose-normalized image generation for person re-identification. In: Proceedings of the European conference on computer vision (ECCV), pp 650–667
20. Lu Y, Wu S, Tai YW, Tang CK (2018) Image generation from sketch constraint using contextual gan. In: Proceedings of the European conference on computer vision (ECCV), pp 205–220 21. Cai L, Gao H, Ji S (2019) Multi-stage variational auto-encoders for coarse-to-fine image generation. In: Proceedings of the 2019 SIAM international conference on data mining. Society for Industrial and Applied Mathematics, pp 630–638
22. Chelsea F, Pieter A, Sergey L (2017) Modelagnostic meta-learning for fast adaptation of deep networks. CoRR, arXiv:1703.03400
23. Nichol A, Achiam J, Schulman J (2018) On first-order meta-learning algorithms. arXiv preprint arXiv:1803.02999
24. Clouâtre L, Demers M (2019) Figr: few-shot image generation with reptile. arXiv preprint arXiv:1901.02199
25. Tripathi S, Bhiwandiwalla A, Bastidas A, Tang H (2019) Using scene graph context to improve image generation. arXiv preprint arXiv:1901.03762

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304062 |

26. C. Li and M. Wand, "Precomputed real-time texture synthesis with markovian generative adversarial networks," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2016, pp. 702–716.
27. N. Jetchev, U. Bergmann, and R. Vollgraf, "Texture synthesis with spatial generative adversarial networks," in Proc. Adv. Neural Inf. Process. Syst. (NIPS), 2017, pp. 1–11.
28. C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA, Jul. 2017, pp. 105–114.
29. X. Wang, "ESRGAN: Enhanced super-resolution generative adversarial networks," in Proc. Eur. Conf. Comput. Vis. Workshops (ECCVW), Sep. 2018, pp. 1–16
30. X. Wang, K. Yu, C. Dong, and C. Change Loy, "Recovering realistic texture in image super-resolution by deep spatial feature transform," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Salt Lake City, UT, USA, Jun. 2018, pp. 606–615.
31. C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, and H. Li, "Highresolution image inpainting using multi-scale neural patch synthesis," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 6721–6729.
32. Y. R. A. Yeh, C. Chen, T. Y. Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do, "Semantic image inpainting with deep generative models," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA, 2017, pp. 6882–6890.
33. Y. Li, S. Liu, J. Yang, and M. Yang, "Generative face completion," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA, Jul. 2017, pp. 5892–5900.
34. R. Mokady, S. Benaim, L. Wolf, and A. Bermano, "Mask based unsupervised content transfer," 2019, arXiv:1906.06558. [Online]. Available: http://arxiv.org/abs/1906.06558
35. W. Yin, Z. Liu, and C. Change Loy, "Instance-level facial attributes transfer with geometry-aware flow," in Proc. AAAI Conf. Artif. Intell., vol. 33, Jul. 2019, pp. 9111–9118.
36. X. Zhou, S. Huang, B. Li, Y. Li, J. Li, and Z. Zhang, "Text guided person image synthesis," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 3663–3672.

(For Video Generation)

[1] Yuxi Li. Deep reinforcement learning: An overview,2018.
[2] Tianyu Gu and John M. Dolan. On-road motion planning for autonomous vehicles, 2012.
[3] Khush Agrawal and Rohit Lal. Person following mobile robot using multiplexed detection and tracking. In Vilas R. Kalamkar and Katarina Monkova, editors, Advances in Mechanical Engineering, pages 815–822, Singapore, 2021. Springer Singapore. ISBN 978-981-15-3639-7.
[4] Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction, 2016.
[5] Emily Denton and Rob Fergus. Stochastic video generation with a learned prior, 2018.
[6] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2014.
[7] Manoj Kumar, Mohammad Babaeizadeh, Dumitru Erhan, Chelsea Finn, Sergey Levine, Laurent Dinh, and Durk Kingma. Videoflow: A conditional flow-based model for
stochastic video generation, 2020.
[8] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial
networks, 2014.
[9] Alex X. Lee, Richard Zhang, Frederik Ebert, Pieter Abbeel, Chelsea Finn, and Sergey Levine. Stochastic adversarial video prediction, 2018.
[10] Ruslan Rakhimov, Denis Volkhonskiy, Alexey Artemov, Denis Zorin, and Evgeny Burnaev. Latent video transformer, 2020.
[11] Aidan Clark, Jeff Donahue, and Karen Simonyan. Adversarial video generation on complex datasets, 2019.
[12] Pauline Luc, Aidan Clark, Sander Dieleman, Diego de Las Casas, Yotam Doron, Albin Cassirer, and Karen Simonyan. Transformation-based adversarial video prediction
on large-scale data.
[13] Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. Axial attention in multidimensional transformers, 2019.
[14] Dirk Weissenborn, Oscar T¨ackstr¨om, and Jakob Uszkoreit. Scaling autoregressive video models, 2020.
[15] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error, 2016.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304062 |

[16] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhutdinov. Unsupervised learning of video representations using lstms, 2016.
[17] Junhyuk Oh, Xiaoxiao Guo, Honglak Lee, Richard Lewis, and Satinder Singh. Action-conditional video prediction using deep networks in atari games, 2015.
[18] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Anticipating visual representations from unlabeled video, 2016.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details