AI-Generated Videos and Deepfakes: A Technical Primer

Abu Sufian

CNR-ISASI, Lecce, Italy.
&
University of Gour Banga, English Bazar, India.
e-mail: abu.sufian@isasi.cnr.it/sufian@ieee.org

Abstract

Artificial intelligence, specially deep learning (DL)-based computer vision algorithms has been revolutionizing video generation, enabling the creation of realistic videos through advanced algorithms specially through DL models. These AI-generated videos bring opportunities in many industries and individual content creators but also bring major threats to humanity for misuse as deepfakes. This tutorial paper overviews the methodologies driving AI-generated videos and its underlying key technologies. It briefly explores the foundational roles of Generative Adversarial Networks, Diffusion Model, and Autoencoders and their backbone DL algorithms in synthesizing video content. The paper also briefly discussed the opportunities of generated videos as well as potential threats in the era of deepfakes. It also discussed significant challenges, ethical considerations, and future directions for enhancing control and creativity in AI video generation. The content will be updated in successive versions of the paper from time to time as the state-of-the-art progresses.

Keywords: Auto-encoder, Deep Learning, Diffusion Model, GAN, Generative AI, LLM, SORA, Tutorial Paper.

1. Introduction

The AI-generated video technology largely improved through the advent of large language models (LLMs) [1, 2, 2, 3, 4], image generation models [5, 6, 7] and video geration models [8, 9, 10] specially through advancement of deep learning (DL) [11, 12, 13]. These advanced DL algorithms-based models now produce videos indistinguishable from real footage [14, 15]. This technology offers exciting opportunities in various industries, learner, content creators, etc.[16, 17, 18] but also presents a major challenge through the potential misuse as deepfakes. Deepfake videos are manipulated or generated digital contents that can convincingly depict real people saying or doing things they never did [19, 20, 21, 22, 14].

This tutorial paper provides a brief technical overview of AI-generated video technology and its core functionalities. It explores key approaches such as Generative Adversarial Networks (GAN) [23, 24], Diffusion Models [25, 26, 27], and Auto-encoder [28, 29]. The paper briefly mentioned how these techniques synthesize video content, highlighting their applications and potential threats. Moreover, the paper mentioned the technical challenges associated with AI-generated videos, and ethical considerations surrounding the technology's use, particularly in the era of deepfakes. Lastly, the paper explores future directions for AI video generation, focusing on enhancing control, fostering creative applications, future threats. The objective of this tutorial paper is to give an introductory awareness and recent progress of AI-generated video

it's technical background and ethical considerations to early researchers, users, and other stakeholders. So, the content will be updated in successive versions of the paper from time to time as the state-of-the-art progresses. This aims to help better manage risks and unlock the potential of AI-generated videos for positive uses and making plane to combat deepfakes.

The organization of the paper is as follows: Different AI-video generation tasks in Section 2, State-of-the-art with the power of DL in Section 3, potential applications of AI-generated video in Section 4, a brief discussion about ethical considerations in Section 5, possible challenges and future directions in Section 6, and finally, the conclusion in Section 7.

2. Different AI-Video Generation Tasks

AI algorithm generate video using different video generating tasks such as text-to-video, image-to-video, video-to-video translation, video prediction, video inpainting, deepfake video, etc. Here are provided brief discussion of the prominent task or approaches:

2.1. Text-to-Video

This approach takes textual descriptions as input and generates corresponding videos. It leverages advanced natural language processing (NLP) [30, 31] techniques to understand the textual prompt and translate it into a video frames. The idea is not new one [32, 33] but recently OpenAI shown their SORA'a ability in video generation as one snap depicted in Fig. 1. Other prominent tools like InVideo AI, and Phenaki allows users to create complex video scenes with detailed textual descriptions.

Figure 1: An AI-generated video by SORA using text prompt.

The process work with several steps:

- **Text Analysis**: NLP Frameworks such as GPT uses to analyze textual input, then semantic extraction is done through identification of entities, actions, and contextual attributes.

- **Scene and Script Generation:** Informing the sequence of scenes based on text analysis to construct storyboard. After that script is developed by detailing the actions and dialogue for each scene.

- **Visual Content Generation:** Utilizing software like Blender and Maya for creating and animating objects and characters. Then GAN such as StyleGAN [34] is employed for synthesizing high-fidelity video frames.

- **Audio Generation:** Uses of tools like Tacotron for synthesizing complementary audio components including integration of sound effects and background scores.

Sometimes post-production video editing software are used to create pseudo realistic video.

2.2. Image-to-Video

Here, the AI system analyzes a set of still images and generates a consecutive video frames that depicts a seamless continuation or animation of the scene [33, 35, 36]. This approach employs techniques like motion estimation and optical flow to create realistic motion patterns. The tools like Runway, Animoto, Simplified are very prominent for image-to-video generation.

The process involves several key steps and technologies to create a coherent and sequence of frames that visually narrate a story. Here is the key steps of the technical aspects:

- **Image Preparation and Analysis:** First organize images in a logical sequence to ensure a natural flow. Then extract and identify key features from each image using computer vision techniques.

- **Interpolation and Animation:** Designing algorithms to generate intermediate frames between two images to create smooth transitions. Techniques such as optical flow and DL-based interpolation such as DAIN are usually used. It also analyze the motion between images to generate in-between frames that depict gradual changes of motion.

- **Generative Models:** Implement GANs to synthesize realistic video frames from images. Models like StyleGAN, VQ-VAE, etc., can generate high-quality frames that ensure continuity and realism. RNNs, and LSTMs use to predict and generate subsequent frames based on previous ones ensuring temporal coherence.

- **Temporal Consistency:** Ensure temporal consistency across frames to avoid flickering and artifacts. Techniques like temporal coherence algorithms and post-processing filters help maintain smooth transitions. Video stabilization algorithms may also apply to reduce jitter and improve visual quality.

Here also post-production processing shall help to enhance the quality of the video.

2.3. Video-to-Video Translation

Video-to-video translation converts one video to another, such as transforming a day scene into a night scene. This is often accomplished using GANs, where the generator learns to map input frames to the desired output frames [37]. Here, scene transformation converts videos from one style to another. GANs application utilizes for frame mapping. Methods like Vid2Vid by NVIDIA, CycleGAN [38], etc., can effectively do this task. The diffusion models are also emerging as method generation for this task [27, 39].

This involves several key steps which includes:

- **Data Preparation:** Extract and preprocess frames from the input video, then normalize pixel values and provide annotations if needed.
- **Model Design and Training:** Use video-specific GAN algorithm like Vid2Vid for frame synthesis. Integration of RNNs, LSTMs, and optical flow networks is effective to capture and maintain temporal dependencies. Employing adversarial, content, temporal, and perceptual losses for ensuring frame realism and consistency.
- **Translation Process:** Translate frames sequentially, considering previous frames for context and use optical flow to guide frame generation and ensure motion consistency.
- **Post-Processing and Output Generation:** Need to reduce flickering and correct jitters also ensure consistent color across frames. Finally combine frames into a final video and render in the desired format.w

2.4. Video Prediction

Video prediction are generating future frames of imputed video based on a sequence of past frames of this video with temporal patterns [40]. This is primarily achieved using RNN or LSTM networks [41], which analyze the temporal patterns in the video data to predict subsequent possible future frames. That is pattern analysis based frame prediction is anticipates future frames from past data. Models are like Deep Video Prediction [42] and PredNet [42] are prominent for this kind of task.

This is a complex task requiring sophisticated ML algorithms to understand and predict temporal dynamics and visual consistency. Here is a brief overview of the technical aspects of the process:

- **Data Preparation and Preprocessing Frame Extraction:** Extract frames from the input video sequence then normalize pixel values to ensure consistent input data. After that identify and extract key features from each frame using computer vision algorithms.
- **Model Designing and Training:** Use RNNs and their variants, such as LSTM or Gated Recurrent Units (GRUs) to capture temporal dependencies among frames. After that apply CNNs to extract spatial features and patterns from each frame, the combined model like ConvLSTM, which combine CNNs and LSTMs are effective to simultaneously capture spatial and temporal information.

- **Temporal Dynamics and Sequence Modeling:** Training the model on sequences of frames to learn the dynamics of motion and appearance, then it can generate future frames by predicting the next frame based on previous frames.

- **Generative Models:** Variational Autoencoders (VAEs) shall be use to learn a latent representation of the frames and generate future frames from this latent space. Then implement GANs to generate realistic frames, where the generator creates frames and the discriminator evaluates their realism.

- **Training Process:** Here loss functions such as Mean Squared Error (MSE) for pixel-wise accuracy and adversarial loss for realism are effective. The back-propagation and optimization algorithms such as Adam minimize the loss and update model parameters.

- **Frame Synthesis and Refinement Frame Generation:** Generate future frames iteratively based on predicted sequences and apply post-processing techniques to enhance frame quality and remove artifacts.

- **Evaluation with Useful Metrics:** Use metrics such as Structural Similarity Index (SSIM), Peak Signal-to-Noise Ratio (PSNR), video-specific metrics, etc., to evaluate prediction accuracy and quality. The model also need be validate on unseen data to ensure generalization.

2.5. Video Inpainting

This technique focuses on filling the missing portions of existing videos [43, 44]. The AI analyzes the surrounding content and synthesizes realistic replacements for the missing parts, resulting in a complete video. Tools like DeepFlow-Guided Video Inpainting [45], DeepFill v1 & v2[1], etc. are useful for this task.

This complex task leverages ML with several key steps to achieve realistic results. The key steps includes:

- **Data Preparation:** Extract and preprocess frames from the input video and generate masks indicating missing or corrupted regions.

- **Model Selection and Training:** Designed models using DL architecture like temporal GANs or autoencoders for video inpainting. Incorporate RNNs, LSTMs, or ConvLSTMs to capture temporal dependencies and ensure smooth transitions. Leverage adversarial, reconstruction, perceptual, and temporal losses to guide the training the model for inpainting process.

- **Missing Region Generation:** Generate missing regions by considering spatial context within each frame and temporal context across frames. Then apply optical flow techniques to maintain motion consistency in the inpainted regions.

- **Output Generation:** Required combine inpainted frames into the final video sequence and render in the desired format. Use post-processing techniques to reduce artifacts and ensure temporal smoothness. Also required to ensure consistent color and lighting across inpainted and original regions.

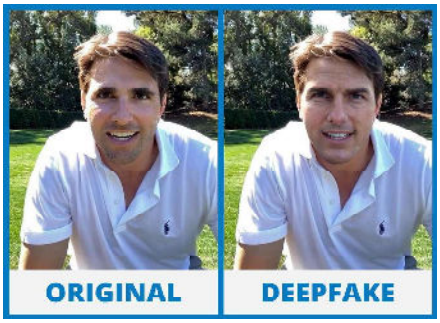

Figure 2: Deepfake example.

2.6. Deepfake Video

The deepfake consists two words: 'deep' here it mentioning 'deep learning' (DL) and 'fake' as usual not real [46]. The AI-generated video mainly created using DL, so, can be treat deepfake videos but videos that replace one person's likeness with a fake video, are usually called deepfake video as depicted in Fig.2. This technology uses GANs, diffusion model and autoencoders to create highly convincing fake videos [47, 48]. Deepfake raising both exciting possibilities [49] and big ethical concerns [50].

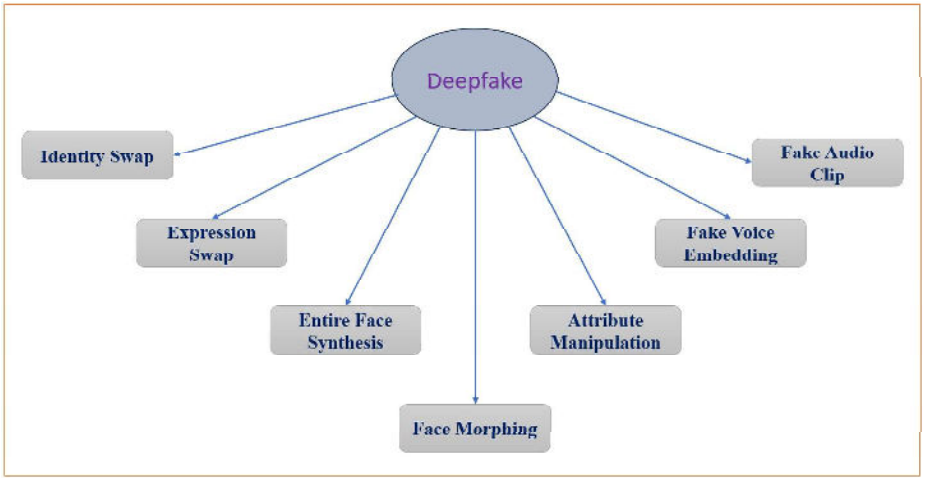

Figure 3: Common types of deepfake.

Deepfake could be different types as depicted in Fig.3. Models like DeepFaceLab [51], FaceSwap [52], Reface, Zao, DeepFake are largely use.

The creation of realistic deepfake is not a easy task, because it required large of number data and complex DL architecture. Here is a brief technical overview of the process:

- **Data Collection and Preparation:** Need to collect extensive datasets of images and videos of both the source and target individuals or objects. Then normalize, align, and crop faces or objects in the dataset for consistency using facial landmark detection for precise alignment.
- **Model Design and Training:** Employ GAN architectures such as StyleGAN or CycleGAN, tailored for face synthesis. Diffusion model such as stable diffusion, etc. Use autoencoders or variational autoencoders (VAEs) to encode and

[1] github'dot'com/JiahuiYu/generative_inpainting

decode facial features. Implement models like FaceSwap, which use encoder-decoder networks to capture and replicate facial features and expressions.

- **Training Process:** The adversarial training is effective to the generator to collect features to creates synthetic frames. The discriminator evaluates their realism, refining the model iteratively. Use a combination of adversarial loss, perceptual loss for high-level feature similarity), identity loss to preserve the individual's identity, and temporal loss to ensure smooth transitions between frames.

- **Fake Video Generation:** Encode the source face or object into a latent space representation and decode the latent representation onto the target video frames, ensuring and movements and facial expressions match the target. Combine the processed frames into a continuous video sequence and render in the desired format and resolution. Required to maintain temporal coherence across frames to avoid flickering and ensure smooth motion.

Sometimes post processing required such as blending which seamlessly blend the synthesized face with the target video using image processing techniques. Color correction to ensure consistent lighting and color matching between the synthesized face and the target video. Video Stabilization may be apply as stabilization techniques to correct any jitter introduced during synthesis.

3. State-of-the-Art with the Power of Deep Learning

Rise of AI-generated video primarily because of DL, a subfield of machine learning (ML) inspired by the function and structure of the human brain [11, 12, 13]. DL extracted features automatically during training where classical ML required human experts for features extraction as shown in Fig.4 (figure source [2], accessed on July 14, 2024). DL algorithms, specially GAN, Diffusion models, and Autoencoders play pivotal roles in video generation. These models enable the generation of high realistic videos by learning from vast amounts of training data.

GAN consist of a generator and a discriminator that compete against each other to produce and evaluate synthetic video content respectively. Diffusion mwodels employ a stochastic forward and backward process to iteratively enhance the quality of generated video frames, ensuring coherence and realism. Autoencoders encode and decode video frames to reconstruct input data, useful for tasks like video inpainting and compression. Here, each of the models is briefly described.

3.1. GAN

GAN is at the forefront of AI-generated video technology [53]. A GAN consists of two neural networks: the generator and the discriminator [23]. The generator creates synthetic data, while the discriminator evaluates its authenticity against real data. Through iterative fintuning, GANs produce highly realistic videos as mentioned in Fig. 5 (figure source [3], accessed on July 14, 2024).

i). **Generator:** The generator part creates video frames from random noise or latent space. It is designed to capture both spatial and temporal patterns to produce

[2]Medium'dot'com/@hassaanidrees7/demystifying-deep-learning-the-future-of-ai-a9a1b9476a98

[3]labellerr'dot'com/blog/what-is-gan-how-does-it-work/

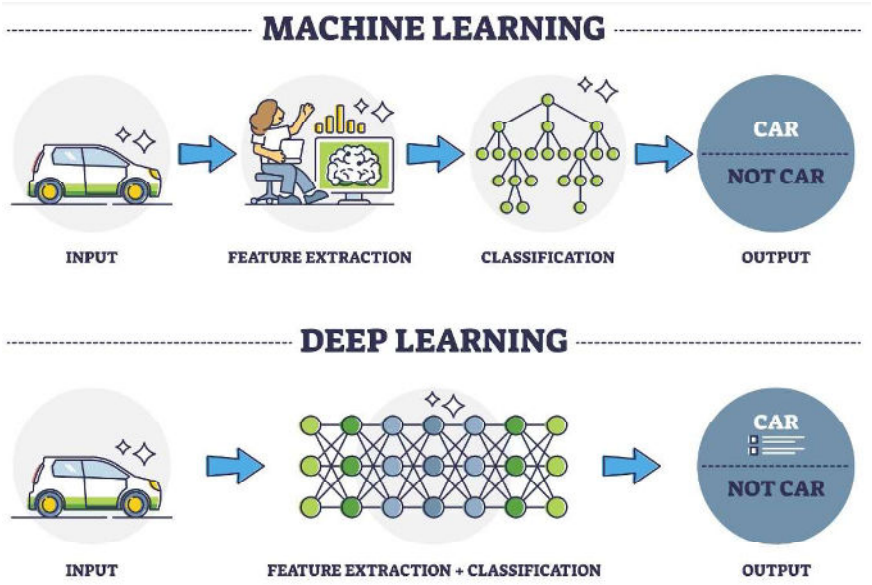

Figure 4: Workflow of ML and DL.[2]

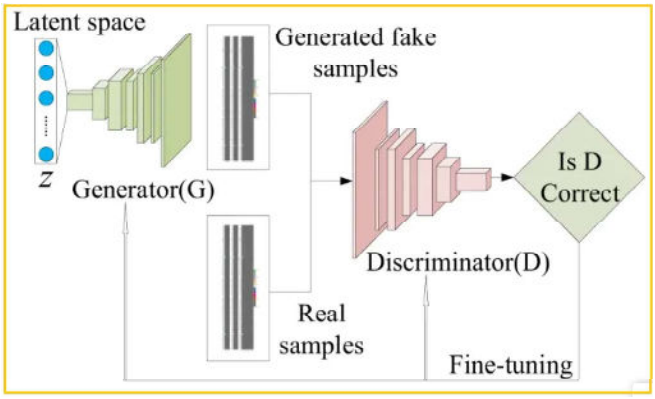

Figure 5: A typical working flow of a Diffusion Model.[3]

coherent video frames.

ii). **Discriminator:** The discriminator part evaluates the generated video frames to distinguishing between real and fake videos. It ensures temporal consistency by analyzing sequences of frames rather than single images.

Recent advancements in GANs for video generation mainly focus on improving temporal coherence and visual quality. Techniques like Temporal GANs (TGANs) [54, 55] and Video GANs [56] utilize temporal convolutions and spatiotemporal attention to ensure smooth frame transitions and capture complex motion patterns. Incorporating 3D convolutions and progressive growth strategies enhances resolution and detail [57]. Additionally, conditioning GANs with contextual information further refines video generation by aligning content with specific scenarios [58]. These innovations collectively enhance the realism and consistency of generated video sequences.

3.2. Diffusion Models

Diffusion models are used to create high-quality, realistic video sequences by modeling the transformation of video data over time [25, 26]. As shown in Fig. 6 (figure

source [4], accessed on July 14, 2024), it work in two phases :-
i). Diffusion (Forward Process): The model starts with an existing video and gradually adds noise to each frame, progressively destroying the video's structure and temporal coherence. This process is repeated for multiple steps until the video frames are nearly indistinguishable from random noise.
ii). Denoising (Reverse Process): The denoising involves training a neural network to denoise the video frames step-by-step, effectively reversing the diffusion process. The network learns to predict and remove the added noise, gradually reconstructing the original video frames. This denoising process is extended to maintain temporal consistency between frames, ensuring smooth transitions and coherent motion.

Diffusion models for video generation have seen many variations and significant advancements. For generating high-quality images, adapted for video by ensuring temporal coherence and efficiency Stable Diffusion [59] is very effective. It utilizes advanced denoising techniques to maintain high spatial resolution and temporal consistency across frames [60]. To handle temporal dynamics in videos Denoising Diffusion Probabilistic Models (DDPM) [61] extends foundational diffusion models. These type of models apply a sequence of denoising steps to gradually reconstruct the video from Gaussian noise. Some diffusion model specifically tailored for video generation, using temporal convolutions or 3D convolutions such as VDM [27], VQ-VDM [62], VIDM [63] Some methods process on additional inputs like text, audio, or initial frames for contextually relevant videos where Conditional Diffusion Models work well [64]. It integrate conditional variables directly into the diffusion steps to guide the generation process. Recently transformers-based diffusion models are high highly effective for AI-generated video creation. This type of models excel to capture long-range temporal dependencies and ensuring coherent narrative flow. Here self-attention mechanisms within the transformer architecture to process and integrate both spatial and temporal information across the entire video sequence are effective [65, 66].

[4]turing'dot'com/kb/ultimate-battle-between-deep-learning-and-machine-learning

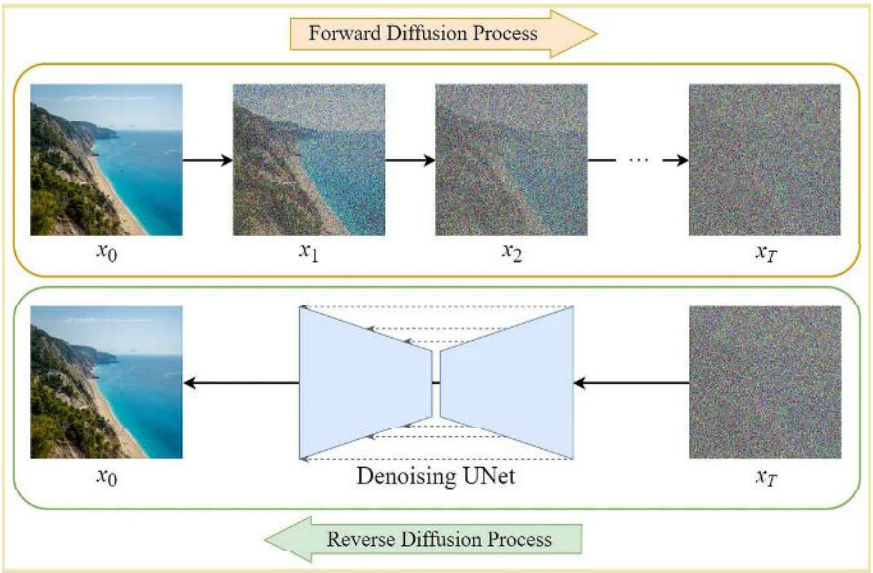

Figure 6: A typical working flow of a Diffusion Model.[4]

3.3. Autoencoder

Autoencoders can be adapted to learn the latent representations of video sequences and generate new videos [67]. It has three primary components:
i). Encoder: The encoder compresses the input video frames into a lower-dimensional latent representation. This involves convolutional layers to capture spatial features and recurrent layers to capture temporal dependencies.
ii). Latent Space: The latent space represents a compressed version of the input video including essential features.
iii). Decoder: The decoder reconstructs new video frames from the latent space that resemble the input video.

Recent advancements in autoencoder techniques have significantly enhanced video generation capabilities. Classical autoencoders have been adapted to handle temporal dynamics through innovations like convolutional LSTMs and Transformer-based models, which capture both spatial and temporal features more effectively. Recent techniques, such as variational autoencoders (VAEs) [68] combined with RNNs [69] or attention mechanisms [70, 71], allow to generate more coherent and contextually accurate video by modeling long-range dependencies and complex motion patterns.

The above three DL model are largely use in video generation. But their backbone primarily are either convolutional neural network (CNN) [72], RNN (or LSTM) [73], and Transformer [74] or combinations.

3.4. CNN

CNN, a type of DL, is primarily used for image recognition tasks and play a significant role in video generation [72, 75]. It consists of multiple layers such as convolution layers, activation, polling, and fully connected layers. It help in extracting spatial features from video frames, contributing to the creation of high-quality visual content. It work through feature extraction which identifies key elements in video frames and image synthesis which produces detailed and realistic visuals.

Recent advancements in CNN like ResNet's [76] residual connections and DenseNet's [77] dense connections have enabled the training deeper of more efficient networks. Innovations such as Inception modules [78] and EfficientNets [79] have optimized performance by balancing network depth, width, and resolution. Additionally, attention mechanisms [74] and self-supervised learning [80] have enhanced the robustness and generalization of CNNs, expanding their role as a backbone algorithm to video generation models.

3.5. RNN

RNNs, particularly LSTM networks are used for sequence prediction tasks. These are crucial in video generation for maintaining temporal coherence, ensuring that sequential frames in a video are consistent and logically connected. The LSTM Networks capture long-term dependencies in video sequences and temporal coherence which ensures smooth transitions between frames in a video [81].

Advances like LSTM and Gated Recurrent Units (GRUs) [82] have addressed the vanishing gradient problem and allowing for the effective modeling of long-term dependencies. Improvements through attention mechanisms have further enhanced the performance of RNNs by enabling the model to focus on relevant parts of the sequence. Additionally, the integration of Transformer architectures has significantly enhanced the capability of RNNs, making them robust and efficient for a wide range of applications such as NLP and video generation [83].

3.6. Transformer

A transformer model is a type of DL that revolutionized natural language processing (NLP) [84] by using self-attention mechanisms, allowing the model to weigh the importance of different words in a sentence regardless of their position. Unlike traditional RNNs, Transformers process entire sequences in parallel, significantly improving computational efficiency. Transformers have become the foundation for many state-of-the-art models, including BERT [85], GPT [86], and T5 [87]. Transformer are also very useful in video generation [8, 88].

4. Potential Applications of AI-generated Video

Potentiality of AI-generated video is enormous [89, 90]. Think about a million-dollar movie scene being generated using just a few prompts. In the same way, the advertising industry, content creation, and more industries are largely benefiting. So, the opportunity is vast, many industries are already befitting and here are few areas are mentioned below:

4.1. Content Creation and Entertainment

AI-generated videos are transforming the digital content creation and entertainment industry by enabling the creation of realistic special effects, virtual characters, and even entire scenes without the need for physical sets or actors [91, 92]. AI-based special effects can generates realistic effects and characters, same way virtual production creates scenes without physical presence.

AI model make automation in labor-intensive processes, such as scene rendering, special effects, and character animation. Techniques like GANs, diffusion models, and autoencoders generate realistic environments and characters, reducing the time and cost of traditional methods. Additionally, AI can personalize content at scale, tailoring videos to individual viewer preferences, increasing engagement. This technology also enables rapid prototyping and iterative creative processes, allowing creators to experiment and refine their work more smoothly.

4.2. Advertising

Advertisers can leverage AI-generated videos to create personalized and highly targeted advertisement, reaching audiences in innovative and attractive ways [93, 94]. Personalized tailors content to individual preferences and targeted advertising to reaches specific audience segments.

Models like GANs, diffusion models, and autoencoders can create realistic product images, engaging characters, and dynamic backgrounds, making production faster and cheaper. AI can personalize ads on a large scale, tailoring them to individual viewer preferences and behaviors, which boosts engagement and conversion rates. This technology also allows for quick testing and improvement, helping marketers optimize their campaigns more effectively.

4.3. Education

In education, AI-generated videos offer new ways of presenting information through interactive and engaging videos, enhancing the learning experience [95, 96]. It can produces engaging educational videos which enhance learning and improves educational outcomes.

Advanced models using GANs, diffusion models, and autoencoders can create realistic simulations and visualizations of educational contents, making complex topics

more understandable. The AI technologies can generate customized content tailored to individual learning paces and styles, providing a more personalized learning experience. Moreover, AI can facilitate rapid content generation and iterative improvement, enabling educators to quickly adapt materials based on student feedback and performance data. This leads to more dynamic and effective educational resources that can be continually refined for better learning outcomes.

5. Ethical Considerations

One major concern of AI-generated content is the misuse of deepfakes. This powerful AI tool can be used to create highly realistic synthetic videos, which can then be exploited to manipulate public opinion, spread misinformation, bias political campaigns, or harm individuals' reputations [19, 97, 98, 21]. Moreover, trust in digital contents will drastically reduce, and liars' dividend attacks will increase [99, 100, 101].

Therefore, AI generated video, specifically deepfakes raise serious ethical concerns, requiring strong solutions. They can show people doing or saying things they never did, which damages trust in the media and informed public discussions. This can seriously affect social relationships, politics, and personal reputations. Another issue is privacy and consent; people in these AI videos often don't know or control how their image is used. Moreover, deepfakes can be a security threat, as they can be used for identity theft, fraud, or blackmail.

To reduce these risks, it is crucial to develop effective methods to detect deepfakes accurately. Raising awareness about AI [102, 103] and improving the trustworthiness of AI models [104] can help rebuild confidence. We also need strong ethical guidelines and laws to manage how AI-generated videos are created and shared [105, 106]. It is important for AI researchers, ethicists, policymakers, and the public to work together to ensure AI use aligns with societal values and addresses potential dangers. Transparency in AI video creation is crucial, this includes clearly labeling AI-generated content using digital watermarks and detection techniques. Ethical AI development should focus on inclusivity and fairness, making sure that tools for detecting deepfakes are accessible to everyone and promote a fair and secure digital environment.

6. Possible Challenges and Future Directions

AI video generation is a rapidly evolving field. At this stage it faces many challenges from technical difficulties to misuse and ethical violation. The challenge like training data scarcity[107], generalization [108], explainablity [109], high computation resources requirement [110] are prominent. Misuse of deepfake has huge downside too. There are yet to develop rigorous ethical guidelines, creation and dissemination legal frameworks. Future developments will likely focus on:

Improved training methods: Techniques like transfer learning [111], meta and metric learning [112] and leveraging pre-trained models hold promise for reducing training data requirements and accelerating model development.

Innovative DL-based backbone algorithm: An effective backbone algorithm can offers resource efficiency, explainability and generalizability [113]. The backbone could be associated with state-of-the-art suitable optimization techniques to develop better video generation models.

Explainable and Interpretable Model: As discussed, deep learning (DL) is the major underlying algorithm for video generation model development. However, DL is a kind of black-box technique where explainability and interpretability are very low

[114]. Therefore, this is a potential research area in which one could invest.
Enhanced Control and Creativity: Advancements in intuitive tools and user interfaces will empower user to have greater control over the style and content of generated videos.
Mitigating Misuse: Research on deepfake detection methods and establishing ethical and legal guidelines will be crucial in ensuring the responsible use of this powerful technology [50, 115].

7. Conclusion

AI-generated video has seen remarkable rises, driven by various approaches and the power of deep learning (DL). Approaches like text-to-video, image-to-video, video prediction, video-to-video translation, video inpainting, and deepfake video showcase AI's abilities in transforming video content generation. Each method furnishes unique capabilities, from converting textual descriptions into dynamic visuals to predicting future frame sequence based on historical data.

DL models, such as Generative Adversarial Networks, diffusion models, autoencoders, and their backbone architecture like Convolutional Neural Networks, Recurrent Neural Networks, and Transformers, are very effective to improved the realism, coherence, and quality of AI-generated videos enabling more sophisticated contents.

Applications of AI-generated videos span content creation and entertainment, advertising, education, and many more. Despite many positive impacts, it has significant downsides and ethical concerns. As research progresses, understanding the technical underpinnings, the risk and ethics, AI-generated video promises to become more versatile and impactful things in the future. This tutorial article tried to given a primer on technical aspect of AI-generated video and potential threats as deepfake. The content will be updated in successive version of this tutorial.

References

[1] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, et al., A survey of large language models, arXiv preprint arXiv:2303.18223 (2023).

[2] M. U. Hadi, R. Qureshi, A. Shah, M. Irfan, A. Zafar, M. B. Shaikh, N. Akhtar, J. Wu, S. Mirjalili, et al., A survey on large language models: Applications, challenges, limitations, and practical usage, TechRxiv Preprints (2023).

[3] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, et al., A survey on evaluation of large language models, ACM Transactions on Intelligent Systems and Technology 15 (3) (2024) 1–45.

[4] M. A. K. Raiaan, M. S. H. Mukta, K. Fatema, N. M. Fahad, S. Sakib, M. M. J. Mim, J. Ahmad, M. E. Ali, S. Azam, A review on large language models: Architectures, applications, taxonomies, open issues and challenges, IEEE Access (2024).

[5] C. Zhang, C. Zhang, M. Zhang, I. S. Kweon, Text-to-image diffusion models in generative ai: A survey, arXiv preprint arXiv:2303.07909 (2023).

[6] J. Y. Koh, D. Fried, R. R. Salakhutdinov, Generating images with multimodal language models, Advances in Neural Information Processing Systems 36 (2024).

[7] J. Xu, X. Liu, Y. Wu, Y. Tong, Q. Li, M. Ding, J. Tang, Y. Dong, Imagereward: Learning and evaluating human preferences for text-to-image generation, Advances in Neural Information Processing Systems 36 (2024).

[8] W. Yan, Y. Zhang, P. Abbeel, A. Srinivas, Videogpt: Video generation using vq-vae and transformers, arXiv preprint arXiv:2104.10157 (2021).

[9] Y. Liu, K. Zhang, Y. Li, Z. Yan, C. Gao, R. Chen, Z. Yuan, Y. Huang, H. Sun, J. Gao, et al., Sora: A review on background, technology, limitations, and opportunities of large vision models, arXiv preprint arXiv:2402.17177 (2024).

[10] Y. Liu, X. Cun, X. Liu, X. Wang, Y. Zhang, H. Chen, Y. Liu, T. Zeng, R. Chan, Y. Shan, Evalcrafter: Benchmarking and evaluating large video generation models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 22139–22149.

[11] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, nature 521 (7553) (2015) 436–444.

[12] S. J. Prince, Understanding deep learning, MIT press, 2023.

[13] M. Hassanin, S. Anwar, I. Radwan, F. S. Khan, A. Mian, Visual attention methods in deep learning: An in-depth survey, Information Fusion 108 (2024) 102417.

[14] D. S. Vahdati, T. D. Nguyen, A. Azizpour, M. C. Stamm, Beyond deepfake images: Detecting ai-generated videos, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 4397–4408.

[15] K. Vayadande, M. Bohri, M. Chawala, A. M. Kulkarni, A. Mursal, The rise of ai-generated news videos: A detailed review, How Machine Learning is Innovating Today's World: A Concise Technical Guide (2024) 423–451.

[16] P. Pataranutaporn, V. Danry, J. Leong, P. Punpongsanon, D. Novy, P. Maes, M. Sra, Ai-generated characters for supporting personalized learning and well-being, Nature Machine Intelligence 3 (12) (2021) 1013–1022.

[17] V. Danry, J. Leong, P. Pataranutaporn, P. Tandon, Y. Liu, R. Shilkrot, P. Punpongsanon, T. Weissman, P. Maes, M. Sra, Ai-generated characters: putting deepfakes to good use, in: CHI Conference on Human Factors in Computing Systems Extended Abstracts, 2022, pp. 1–5.

[18] C. Liu, H. Yu, Ai-empowered persuasive video generation: A survey, ACM Computing Surveys 55 (13s) (2023) 1–31.

[19] M. Westerlund, The emergence of deepfake technology: A review, Technology innovation management review 9 (11) (2019).

[20] T. Zhang, Deepfake generation and detection, a survey, Multimedia Tools and Applications 81 (5) (2022) 6259–6276.

[21] B. Cho, B. M. Le, J. Kim, S. Woo, S. Tariq, A. Abuadbba, K. Moore, Towards understanding of deepfake videos in the wild, in: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, 2023, pp. 4530–4537.

[22] M. Mustak, J. Salminen, M. Mäntymäki, A. Rahman, Y. K. Dwivedi, Deepfakes: Deceptions, mitigations, and opportunities, Journal of Business Research 154 (2023) 113368.

[23] I. Goodfellow, J. P Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, Communications of the ACM 63 (11) (2020) 139–144.

[24] T. Chakraborty, U. R. KS, S. M. Naik, M. Panja, B. Manvitha, Ten years of generative adversarial nets (gans): a survey of the state-of-the-art, Machine Learning: Science and Technology 5 (1) (2024) 011001.

[25] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, M.-H. Yang, Diffusion models: A comprehensive survey of methods and applications, ACM Computing Surveys 56 (4) (2023) 1–39.

[26] F.-A. Croitoru, V. Hondru, R. T. Ionescu, M. Shah, Diffusion models in vision: A survey, IEEE Transactions on Pattern Analysis and Machine Intelligence 45 (9) (2023) 10850–10869.

[27] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, D. J. Fleet, Video diffusion models, Advances in Neural Information Processing Systems 35 (2022) 8633–8646.

[28] J. Zhai, S. Zhang, J. Chen, Q. He, Autoencoder and its various variants, in: 2018 IEEE international conference on systems, man, and cybernetics (SMC), IEEE, 2018, pp. 415–419.

[29] K. Berahmand, F. Daneshfar, E. S. Salehi, Y. Li, Y. Xu, Autoencoders and their applications in machine learning: a survey, Artificial Intelligence Review 57 (2) (2024) 28.

[30] S. C. Fanni, M. Febi, G. Aghakhanyan, E. Neri, Natural language processing, in: Introduction to Artificial Intelligence, Springer, 2023, pp. 87–99.

[31] L. Qin, Q. Chen, X. Feng, Y. Wu, Y. Zhang, Y. Li, M. Li, W. Che, P. S. Yu, Large language models meet nlp: A survey, arXiv preprint arXiv:2405.12819 (2024).

[32] G. Mittal, T. Marwah, V. N. Balasubramanian, Sync-draw: Automatic video generation using deep recurrent attentive architectures, in: Proceedings of the 25th ACM international conference on Multimedia, 2017, pp. 1096–1104.

[33] Y. Li, M. Min, D. Shen, D. Carlson, L. Carin, Video generation from text, in: Proceedings of the AAAI conference on artificial intelligence, Vol. 32, 2018.

[34] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 4401–4410.

[35] Y. Hu, C. Luo, Z. Chen, Make it move: controllable image-to-video generation with text descriptions, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 18219–18228.

[36] H. Ni, C. Shi, K. Li, S. X. Huang, M. R. Min, Conditional image-to-video generation with latent flow diffusion models, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 18444–18455.

[37] Y. Borole, R. Raut, Generative adversarial networks for video-to-video translation, in: Generative Adversarial Networks and Deep Learning, Chapman and Hall/CRC, 2023, pp. 53–66.

[38] J.-Y. Zhu, T. Park, P. Isola, A. A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Computer Vision (ICCV), 2017 IEEE International Conference on, 2017.

[39] W. Weng, R. Feng, Y. Wang, Q. Dai, C. Wang, D. Yin, Z. Zhao, K. Qiu, J. Bao, Y. Yuan, et al., Art-v: Auto-regressive text-to-video generation with diffusion models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 7395–7405.

[40] e. Escontrela, Video prediction models as rewards for reinforcement learning, Advances in Neural Information Processing Systems 36 (2024).

[41] Y. Yu, X. Si, C. Hu, J. Zhang, A review of recurrent neural networks: Lstm cells and network architectures, Neural computation 31 (7) (2019) 1235–1270.

[42] W. Lotter, G. Kreiman, D. Cox, Deep predictive coding networks for video prediction and unsupervised learning, arXiv preprint arXiv:1605.08104 (2016).

[43] W. Quan, J. Chen, Y. Liu, D.-M. Yan, P. Wonka, Deep learning-based image and video inpainting: A survey, International Journal of Computer Vision 132 (7) (2024) 2367–2400.

[44] J. Wu, X. Li, C. Si, S. Zhou, J. Yang, J. Zhang, Y. Li, K. Chen, Y. Tong, Z. Liu, et al., Towards language-driven video inpainting via multimodal large language models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 12501–12511.

[45] R. Xu, X. Li, B. Zhou, C. C. Loy, Deep flow-guided video inpainting, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3723–3732.

[46] T. T. Nguyen, Q. V. H. Nguyen, D. T. Nguyen, D. T. Nguyen, T. Huynh-The, S. Nahavandi, T. T. Nguyen, Q.-V. Pham, C. M. Nguyen, Deep learning for deepfakes creation and detection: A survey, Computer Vision and Image Understanding 223 (2022) 103525.

[47] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, H. Malik, Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward, Applied intelligence 53 (4) (2023) 3974–4026.

[48] A. AV, S. Das, A. Das, et al., Latent flow diffusion for deepfake video generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 3781–3790.

[49] C.-J. Chang, W.-C. Chien, Towards a positive thinking about deepfakes: Evaluating the experience of deepfake voices in the emotional and rational scenarios, in: International Conference on Human-Computer Interaction, Springer, 2024, pp. 311–325.

[50] E. Meskys, J. Kalpokiene, P. Jurcys, A. Liaudanskas, Regulating deep fakes: legal and ethical considerations, Journal of Intellectual Property Law & Practice 15 (1) (2020) 24–31.

[51] K. Liu, I. Perov, D. Gao, N. Chervoniy, W. Zhou, W. Zhang, Deepfacelab: Integrated, flexible and extensible face-swapping framework, Pattern Recognition 141 (2023) 109628.

[52] I. Korshunova, W. Shi, J. Dambre, L. Theis, Fast face-swap using convolutional neural networks, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 3677–3685.

[53] Y. Cao, S. Li, Y. Liu, Z. Yan, Y. Dai, P. S. Yu, L. Sun, A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt, arXiv preprint arXiv:2303.04226 (2023).

[54] M. Saito, E. Matsumoto, S. Saito, Temporal generative adversarial nets with singular value clipping, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2830–2839.

[55] A. Munoz, M. Zolfaghari, M. Argus, T. Brox, Temporal shift gan for large scale video generation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 3179–3188.

[56] N. Aldausari, A. Sowmya, N. Marcus, G. Mohammadi, Video generative adversarial networks: a review, ACM Computing Surveys (CSUR) 55 (2) (2022) 1–25.

[57] S. Vallecorsa, F. Carminati, G. Khattak, 3d convolutional gan for fast simulation, in: EPJ Web of Conferences, Vol. 214, EDP Sciences, 2019, p. 02010.

[58] M. Shahbazi, M. Danelljan, D. P. Paudel, L. Van Gool, Collapse by conditioning: Training class-conditional gans with limited data, arXiv preprint arXiv:2201.06578 (2022).

[59] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 10684–10695.

[60] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts, et al., Stable video diffusion: Scaling latent video diffusion models to large datasets, arXiv preprint arXiv:2311.15127 (2023).

[61] J. Anderson, N. Akram, Denoising diffusion probabilistic models (ddpm) dynamics: Unraveling change detection in evolving environments, Innovative Computer Sciences Journal 10 (1) (2024) 1–10.

[62] R. Kaji, K. Yanai, Vq-vdm: Video diffusion models with 3d vqgan, in: Proceedings of the 5th ACM International Conference on Multimedia in Asia, 2023, pp. 1–5.

[63] K. Mei, V. Patel, Vidm: Video implicit diffusion models, in: Proceedings of the AAAI conference on artificial intelligence, Vol. 37, 2023, pp. 9117–9125.

[64] Y. Tashiro, J. Song, Y. Song, S. Ermon, Csdi: Conditional score-based diffusion models for probabilistic time series imputation, Advances in Neural Information Processing Systems 34 (2021) 24804–24816.

[65] M. F. Sikder, R. Ramachandranpillai, F. Heintz, Transfusion: generating long, high fidelity time series using diffusion models with transformers, arXiv preprint arXiv:2307.12667 (2023).

[66] Z. Fei, M. Fan, C. Yu, D. Li, Y. Zhang, J. Huang, Dimba: Transformer-mamba diffusion models, arXiv preprint arXiv:2406.01159 (2024).

[67] Z. Lai, S. Liu, A. A. Efros, X. Wang, Video autoencoder: self-supervised disentanglement of static 3d structure and motion, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 9730–9740.

[68] L. Girin, S. Leglaive, X. Bie, J. Diard, T. Hueber, X. Alameda-Pineda, Dynamical variational autoencoders: A comprehensive review, arXiv preprint arXiv:2008.12595 (2020).

[69] M. Jang, S. Seo, P. Kang, Recurrent neural network-based semantic variational autoencoder for sequence-to-sequence learning, Information Sciences 490 (2019) 59–73.

[70] P. Shamsolmoali, M. Zareapoor, H. Zhou, D. Tao, X. Li, Vtae: Variational transformer autoencoder with manifolds learning, IEEE Transactions on Image Processing (2023).

[71] M. B. Rocha, R. A. Krohling, Vae-gna: a variational autoencoder with gaussian neurons in the latent space and attention mechanisms, Knowledge and Information Systems (2024) 1–23.

[72] A. Ghosh, A. Sufian, F. Sultana, A. Chakrabarti, D. De, Fundamental concepts of convolutional neural network, Recent trends and advances in artificial intelligence and Internet of Things (2020) 519–567.

[73] A. Sherstinsky, Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network, Physica D: Nonlinear Phenomena 404 (2020) 132306.

[74] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[75] H. Tian, P. Gao, R. Wei, M. Paul, Dilated convolutional neural network-based deep reference picture generation for video compression, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2022, pp. 2824–2828.

[76] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[77] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.

[78] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.

[79] M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: International conference on machine learning, PMLR, 2019, pp. 6105–6114.

[80] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, J. Tang, Self-supervised learning: Generative or contrastive, IEEE transactions on knowledge and data engineering 35 (1) (2021) 857–876.

[81] S. Gupta, A. Keshari, S. Das, Rv-gan: Recurrent gan for unconditional video generation, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 2024–2033.

[82] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, arXiv preprint arXiv:1412.3555 (2014).

[83] D. Neimark, O. Bar, M. Zohar, D. Asselmann, Video transformer network, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 3163–3172.

[84] K. Chowdhary, K. Chowdhary, Natural language processing, Fundamentals of artificial intelligence (2020) 603–649.

[85] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[86] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., Gpt-4 technical report, arXiv preprint arXiv:2303.08774 (2023).

[87] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of machine learning research 21 (140) (2020) 1–67.

[88] S. R. Dubey, S. K. Singh, Transformer-based generative adversarial networks in computer vision: A comprehensive survey, IEEE Transactions on Artificial Intelligence (2024).

[89] Y. Wang, Synthetic realities in the digital age: Navigating the opportunities and challenges of ai-generated content, TechRxiv Preprints (2023).

[90] K. Dunnell, G. Agarwal, P. Pataranutaporn, A. Lippman, P. Maes, Ai-generated media for exploring alternate realities, in: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, 2024, pp. 1–8.

[91] N. Anantrasirichai, D. Bull, Artificial intelligence in the creative industries: a review, Artificial intelligence review 55 (1) (2022) 589–656.

[92] T. H. Davenport, R. Bean, The impact of generative ai on hollywood and entertainment, MIT Sloan Management Review 19 (2023).

[93] M. Kumar, A. Kapoor, Generative ai and personalized video advertisements, Available at SSRN 4614118 (2023).

[94] M. Danesi, Ai in marketing and advertising, in: AI-Generated Popular Culture: A Semiotic Perspective, Springer, 2024, pp. 127–142.

[95] D. Leiker, A. R. Gyllen, I. Eldesouky, M. Cukurova, Generative ai for learning: investigating the potential of learning videos with synthetic virtual instructors, in: International conference on artificial intelligence in education, Springer, 2023, pp. 523–529.

[96] J. Lim, The potential of learning with ai-generated pedagogical agents in instructional videos, in: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, 2024, pp. 1–6.

[97] J. T. Hancock, J. N. Bailenson, The social impact of deepfakes (2021).

[98] M. Pawelec, Deepfakes and democracy (theory): How synthetic audio-visual media for disinformation and hate speech threaten core democratic functions, Digital society 1 (2) (2022) 19.

[99] C. Vaccari, A. Chadwick, Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news, Social media+ society 6 (1) (2020) 2056305120903408.

[100] H. Etienne, The future of online trust (and why deepfake is advancing it), AI and Ethics 1 (4) (2021) 553–562.

[101] K. J. Schiff, D. S. Schiff, N. Bueno, The liar's dividend: The impact of deepfakes and fake news on trust in political discourse (2023).

[102] M. Kandlhofer, P. Weixelbraun, M. Menzinger, G. Steinbauer-Wagner, Á. Kemenesi, Education and awareness for artificial intelligence, in: International Conference on Informatics in Schools: Situation, Evolution, and Perspectives, Springer Nature Switzerland Cham, 2023, pp. 3–12.

[103] A. Pant, R. Hoda, S. V. Spiegler, C. Tantithamthavorn, B. Turhan, Ethics in the age of ai: an analysis of ai practitioners' awareness and challenges, ACM Transactions on Software Engineering and Methodology 33 (3) (2024) 1–35.

[104] B. Chander, C. John, L. Warrier, K. Gopalakrishnan, Toward trustworthy artificial intelligence (tai) in the context of explainability and robustness, ACM Computing Surveys (2024).

[105] N. Díaz-Rodríguez, J. Del Ser, M. Coeckelbergh, M. L. de Prado, E. Herrera-Viedma, F. Herrera, Connecting the dots in trustworthy artificial intelligence: From ai principles, ethics, and key requirements to responsible ai systems and regulation, Information Fusion 99 (2023) 101896.

[106] M. N. Weldon, G. Thomas, L. Skidmore, Establishing a future-proof framework for ai regulation: Balancing ethics, transparency, and innovation, Transactions: The Tennessee Journal of Business Law 25 (2) (2024) 2.

[107] L. Alzubaidi, J. Bai, A. Al-Sabaawi, J. Santamaría, A. S. Albahri, B. S. N. Al-dabbagh, M. A. Fadhel, M. Manoufali, J. Zhang, A. H. Al-Timemy, et al., A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications, Journal of Big Data 10 (1) (2023) 46.

[108] C. Zhang, S. Bengio, M. Hardt, B. Recht, O. Vinyals, Understanding deep learning (still) requires rethinking generalization, Communications of the ACM 64 (3) (2021) 107–115.

[109] M. B. Fazi, Beyond human: deep learning, explainability and representation, Theory, Culture & Society 38 (7-8) (2021) 55–77.

[110] N. C. Thompson, K. Greenewald, K. Lee, G. F. Manso, The computational limits of deep learning, arXiv preprint arXiv:2007.05558 10 (2020).

[111] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, Q. He, A comprehensive survey on transfer learning, Proceedings of the IEEE 109 (1) (2020) 43–76.

[112] X. Li, X. Yang, Z. Ma, J.-H. Xue, Deep metric learning for few-shot image classification: A review of recent developments, Pattern Recognition 138 (2023) 109381.

[113] A. Sufian, A. Ghosh, D. Barman, M. Leo, C. Distante, B. Li, Fewfacenet: A lightweight few-shot learning-based incremental face authentication for edge cameras, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 2018–2027.

[114] V. Hassija, V. Chamola, A. Mahapatra, A. Singal, D. Goel, K. Huang, S. Scardapane, I. Spinelli, M. Mahmud, A. Hussain, Interpreting black-box models: a review on explainable artificial intelligence, Cognitive Computation 16 (1) (2024) 45–74.

[115] Y. Hariprasad, S. Iyengar, N. Subramanian, Deepfake video detection using lip region analysis with advanced artificial intelligence based anomaly detection technique, TechRxiv Preprints (2024).