Democratic and Popular Republic of Algeria

Ministry of higher education and scientific research

University 20 August 1955 Skikda

Faculty of Science

Department of Computer Science

SIAA

Master's thesis

On the Theme

The Detection of AI-Generated Video Sequences

Realised by :

Kabrane Mohamed Achraf

Supervised by :

Dr. Boughamouza Fateh

Academic year

2024 / 2025

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Dr. Boughamouza Fateh, for their invaluable guidance, patience, and support throughout the course of this research. Their expertise and encouragement have been instrumental in helping me navigate the many challenges of this project, and I am truly thankful for the time and attention they dedicated to my work.

I also wish to thank the members of the jury for their insightful feedback and for taking the time to review my thesis. My sincere appreciation goes as well to the faculty and staff of University of 20 aout 1955 Skikda, whose academic and administrative support created a nurturing environment for learning and research. I am equally grateful to my colleagues and fellow students for their camaraderie and the many enriching discussions we shared.

Lastly, I want to thank my family and friends for their support throughout this journey.

To my **mother**, thank you for your constant encouragement and quiet strength. Your love helped me stay calm and steady, even when things felt overwhelming.

To my **father**, your trust in me, your advice, and your steady presence gave me the motivation to keep moving forward when I was tired or uncertain.

To my siblings , thank you for the laughs, the check-ins, and the moments of distraction that helped me breathe when I needed it most.

And to my friends, thank you for being there, whether through kind words, shared silence, or just making me smile. Your support truly mattered.

Dedication

To my parents, my siblings, and my dearest friends

This work is for you. Thank you for your love, your patience, and your constant support.

For believing in me, even when I had doubts.

For encouraging me when things felt heavy.

For being there, always.

I carry this achievement with you.

Abstract

In recent years, the advancement of artificial intelligence has enabled the creation of highly realistic digital content. Among the most notable developments is the rise of deepfake videos — synthetic videos generated by AI models that can closely imitate human faces, voices, and movements. While such technologies offer creative potential, they also raise serious concerns about misinformation, identity fraud, and digital manipulation.

This thesis addresses the challenge of detecting AI-generated videos by proposing a deep learning-based system capable of distinguishing real from synthetic content. The work combines theoretical research on generative models and detection techniques with a practical implementation of a video classification model.

The proposed system uses a hybrid architecture that captures both visual and temporal features in video sequences. The project includes data preparation, model design, training, and evaluation. Results show the model's ability to detect synthetic videos with promising performance, contributing to ongoing efforts in media forensics and digital content verification.

This research reflects the growing need for tools that ensure trust and authenticity in an increasingly digital world.

Keywords: Artificial intelligence, deepfake, detection, synthetic videos, deep learning, CNN, LSTM

Résumé

Ces dernières années, l'intelligence artificielle a connu une évolution rapide, donnant naissance à de nouveaux outils capables de générer des contenus numériques réalistes. Parmi eux, les vidéos synthétiques générées par des modèles d'IA — connues sous le nom de *deepfakes* — représentent à la fois une prouesse technologique et une source d'inquiétude croissante. Ces vidéos peuvent imiter des visages, des voix et des gestes humains de manière très convaincante, rendant leur détection de plus en plus difficile.

Ce mémoire s'inscrit dans ce contexte et propose une approche pour détecter les vidéos générées artificiellement. L'objectif principal est de concevoir et d'implémenter un système intelligent capable de distinguer les vidéos réelles des vidéos synthétiques, en s'appuyant sur des techniques d'apprentissage profond.

La première partie du travail présente les fondements théoriques liés à l'intelligence artificielle générative, aux vidéos deepfake, et aux méthodes existantes de détection. La seconde partie est dédiée à la réalisation pratique : préparation des données, conception du modèle, entraînement, et évaluation des performances. Le système proposé combine un réseau de neurones convolutif avec une architecture LSTM pour analyser les caractéristiques visuelles et temporelles des séquences vidéo.

Les résultats obtenus permettent de valider l'efficacité de cette approche et ouvrent des pistes pour des améliorations futures. Ce travail contribue ainsi à la lutte contre la désinformation numérique et à la protection de l'authenticité des contenus multimédias.

Mots-clés : Intelligence artificielle, deepfake, détection, vidéos générées, apprentissage profond, CNN, LSTM

ملخص

في السنوات الأخيرة، أتاح التقدم في مجال الذكاء الاصطناعي إمكانية إنشاء محتوى رقمي واقعي للغاية. ومن بين أبرز هذه التطورات ظهور فيديوهات التزييف العميق، وهي مقاطع مصطنعة يتم إنشاؤها بواسطة نماذج ذكاء اصطناعي، قادرة على تقليد الوجوه البشرية والأصوات والحركات بدقة عالية. وعلى الرغم من الإمكانات الإبداعية التي توفرها هذه التقنيات، فإنها تثير أيضًا مخاوف جدية تتعلق بنشر المعلومات المضللة، وانتحال الهوية، والتلاعب الرقمي.

تتناول هذه المذكرة تحدي الكشف عن الفيديوهات المُنشأة بواسطة الذكاء الاصطناعي، من خلال اقتراح نظام يعتمد على التعلم العميق، قادر على التمييز بين المحتوى الحقيقي والمحتوى الاصطناعي. ويجمع هذا العمل بين البحث النظري حول النماذج التوليدية وتقنيات الكشف، والتنفيذ العملي لنموذج تصنيف الفيديوهات. يعتمد النظام المقترح على بنية هجينة تُمكّنه من استخراج الخصائص البصرية والزمنية من تسلسلات الفيديو. يشمل المشروع مراحل إعداد البيانات، وتصميم النموذج، وتدريبه، وتقييمه. وتُظهر النتائج قدرة النموذج على كشف الفيديوهات الاصطناعية بأداء واعد، ويساهم هذا العمل في الجهود المستمرة في مجال الطب الشرعي الرقمي والتحقق من صحة المحتوى الرقمي.

ويعكس هذا البحث الحاجة المتزايدة إلى أدوات تضمن الثقة والمصداقية في عالم رقمي متسارع.

الكلمات المفتاحية : الذكاء الاصطناعي، التزييف العميق، الكشف، الفيديوهات الاصطناعية، التعلم العميق، الشبكات العصبية الالتفافية (CNN)، الشبكات العصبية المتكررة طويلة الذاكرة(LSTM)

Table of content

Table of figures

Table of tables

Table of Acronyms

Acronym	Full Term
AI	Artificial Intelligence
ML	Machine Learning
DL	Deep Learning
GAN	Generative Adversarial Network
CNN	Convolutional Neural Network
LSTM	Long Short-Term Memory
RNN	Recurrent Neural Network
MTCNN	Multi-task Cascaded Convolutional Networks
FF++	FaceForensics++
DFDC	Deepfake Detection Challenge
GPU	Graphics Processing Unit
API	Application Programming Interface
BCE	Binary Cross-Entropy
MSE	Mean Squared Error
XAI	Explainable Artificial Intelligence

General Introduction

In recent years, the rapid development of artificial intelligence has brought forward new possibilities in the creation of digital content. Among the most significant advancements is the emergence of generative models capable of producing highly realistic text, images, and increasingly, videos. One of the most notable examples of this progress is deepfake technology — AI-generated videos that can convincingly simulate human faces, voices, and movements. These videos are often indistinguishable from authentic ones, making them both a remarkable achievement and a growing concern in the digital age.

While generative AI has opened creative opportunities in fields like cinema, advertising, and virtual communication, it also raises serious ethical, technical, and societal questions. Deepfakes can be used to spread false information, manipulate public opinion, impersonate individuals, or damage reputations. As these tools become easier to access and more difficult to detect, the need for effective detection systems has become more urgent than ever.

The aim of this thesis is to design and implement an intelligent system capable of identifying AI-generated videos. This involves exploring the characteristics of synthetic video content and applying machine learning techniques to detect subtle differences that may not be visible to the human eye. The work combines theoretical understanding with practical development, with the goal of building a model that can distinguish real from artificial media with reliability.

This research is organized into two main parts: a theoretical exploration and a practical application. The first part provides the necessary background on generative technologies and their implications, offering a foundation for understanding the nature of the problem and the available approaches.

The second part focuses on the design and implementation of a detection system based on deep learning. It includes the preparation of the dataset, the choice of a suitable model architecture, and the training and evaluation process. The methods are tested through a series of experiments designed to assess the effectiveness of the proposed approach.

Through this work, the thesis contributes to the growing field of AI media forensics by offering a focused and applicable solution to the detection of deepfake videos. It reflects an effort to

combine technical innovation with ethical awareness in response to the challenges posed by synthetic media in today's digital landscape.

Theoretical part

Chapter 1: Generative AI in Video

Introduction

The rise of generative AI has opened exciting new possibilities in video creation, transforming how content is produced and experienced. This chapter explores the foundations of this technology, from the basic principles of artificial intelligence to the advanced systems that power modern video generation. We'll examine how these tools are reshaping creative workflows and discuss their potential to redefine storytelling, visual effects, and digital media in the years to come.

1. Intelligence

The concept of intelligence may be defined as the capacity to learn, process, and apply knowledge for the purpose of finding the best solutions and achieving the goals, more so when the situation is ambiguous and keeps changing. This context stresses not only the learning competence but also the knack to be flexible and use proper methods in the face of fresh difficulties. One example of this is a factory robot that has been coded to do the tasks with highest precision and the same level of quality every time. It is not, however, "intelligent" because it does not exhibit the characteristic of wise and flexible decision-making that is typical of true intelligence.[1]

2. Artificial Intelligence

Artificial Intelligence (AI) is the field of study dedicated to creating systems and programs capable of mimicking human cognitive functions such as learning, reasoning, and problem-solving. First introduced by **John McCarthy in 1955**, who defined it as "the science and engineering of making intelligent machines," AI has undergone significant evolution over the decades. Early AI systems were limited to performing specific, pre-programmed tasks. However, with advancements in **machine learning** and **deep learning**, modern AI systems can now process vast amounts of data, learn from it, and continuously improve their performance without explicit reprogramming.

The shift from rule-based to data-driven AI enabled the development of Generative AI, which relies on learned data patterns to autonomously produce new content. Generative AI, as a result of this evolution, allows systems to create text, images, or audio based on complex data understanding. This progression marks a leap in AI capabilities, influencing areas like creativity, education, and healthcare.[2]

3. From Artificial Intelligence to Generative Artificial Intelligence

The evolution of AI has transitioned from symbolic, rule-based systems to sophisticated, data-driven architectures. In the 1950s and 1960s, AI research focused on symbolic AI, where systems operated based on predefined rules and logic. Programs like the Logic Theorist and ELIZA

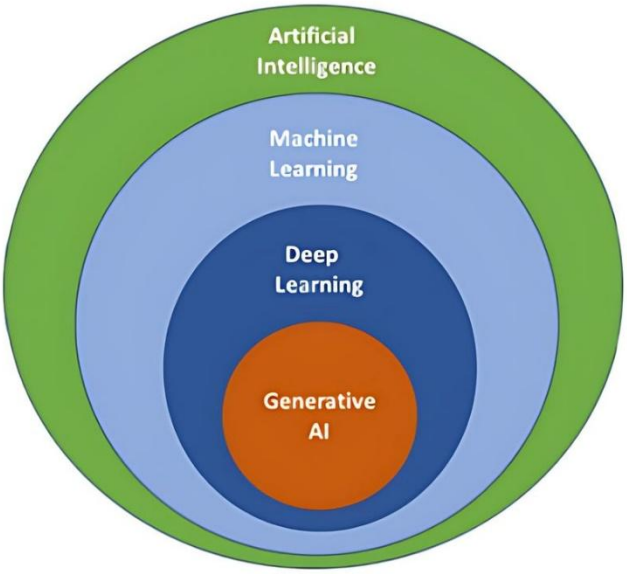

Figure 1 : Nested Domains of Artificial Intelligence [4]

exemplified this era, demonstrating the potential of machines to perform tasks such as theorem proving and natural language processing. However, these systems lacked the ability to learn from data, limiting their adaptability and scalability.[3]

The 1980s and 1990s marked a significant shift with the resurgence of neural networks and the advent of machine learning. The development of backpropagation algorithms enabled neural networks to learn from data, leading to the rise of Machine Learning (ML). This era also saw the emergence of expert systems, which used extensive databases of knowledge to make decisions in specific domains. The integration of statistical methods allowed AI systems to handle uncertainty and make predictions, paving the way for more dynamic and flexible applications.[5]

Deep Learning (DL) is a subset of machine learning that employs multilayered neural networks, known as deep neural networks, to simulate the complex decision-making processes of the human brain. These architectures enable computers to autonomously learn from vast amounts of unstructured data, uncovering intricate patterns and making informed decisions without explicit

programming. This approach has transformed the way machines understand and interact with complex data, leading to significant advancements in various fields.[6]

The 21st century ushered in the era of DL, characterized by the use of these sophisticated neural networks capable of learning complex patterns from vast datasets. Breakthroughs such as the development of convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have enabled significant advancements in image and speech recognition. More recently, transformer-based models like BERT and GPT have revolutionized natural language processing, allowing machines to understand and generate human-like text. These advancements have culminated in the rise of generative AI, where systems can create new content, such as text, images, and music, demonstrating a level of creativity and adaptability previously unattainable by machines.

4. Generative Artificial Intelligence

Generative Artificial Intelligence (Generative AI) refers to a subset of AI technologies designed to create new content—be it text, images, audio, or code—by learning patterns from existing data. Unlike traditional AI models that primarily focus on classification or prediction, generative models aim to produce novel outputs that resemble the data they were trained on. This capability has revolutionized various industries, enabling applications such as automated content creation, realistic image synthesis, and advanced language modeling.

4.1 Evolution of Generative Artificial Intelligence

The journey of Generative AI began in the 1960s with early models like ELIZA, a simple chatbot that mimicked human conversation using pattern matching and substitution methodologies. However, these early systems lacked the sophistication to generate truly novel content.[7]

A significant breakthrough occurred in 2014 with the introduction of Generative Adversarial Networks (GANs) by Ian Goodfellow and his colleagues. GANs consist of two neural networks—the generator and the discriminator—that work in tandem to produce increasingly realistic data outputs. This innovation paved the way for high-quality image and video generation.[8]

The evolution continued with the development of Transformer architectures in 2017, leading to models like BERT and GPT. These models excelled in understanding and generating human-like

text, marking a significant advancement in natural language processing. The 2020s witnessed an "AI boom," characterized by rapid advancements in transformer-based deep neural networks, particularly large language models (LLMs).

4.2 Architectures of Generative Artificial Intelligence Models

Generative AI encompasses a variety of model architectures, each with distinct mechanisms and applications. Below is an overview of the primary types:

4.2.1. Generative Adversarial Networks (GANs)

Introduced by Ian Goodfellow in 2014, GANs consist of two neural networks—the generator and the discriminator—that engage in a competitive process. The generator creates synthetic data, while the discriminator evaluates its authenticity. This adversarial training enables GANs to produce highly realistic images, videos, and audio. Variants like Deep Convolutional GANs (DCGANs) and Conditional GANs have expanded their applicability across various domains.[9]

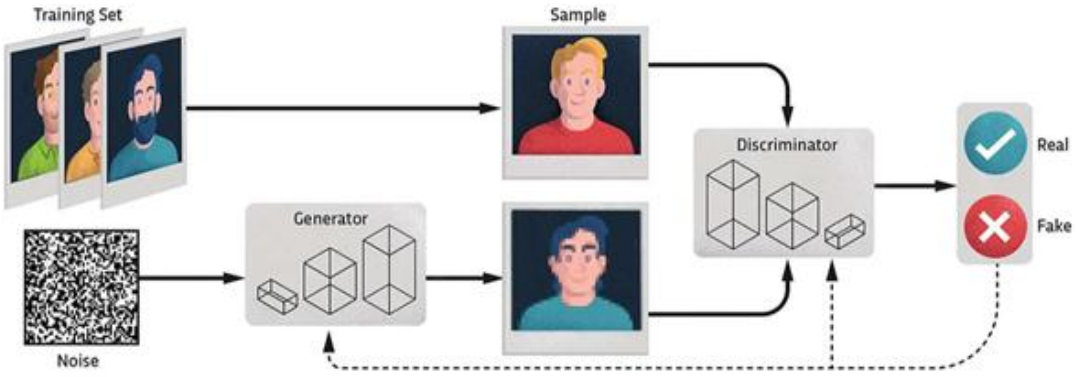

Figure 2 : overview of gan architecture [10]

4.2.2. Variational Autoencoders (VAEs)

VAEs are probabilistic generative models that encode input data into a latent space and then decode it to reconstruct the original data. This approach allows for the generation of new, similar data samples. VAEs are particularly useful in tasks requiring controlled data generation, such as image synthesis and anomaly detection.[11]

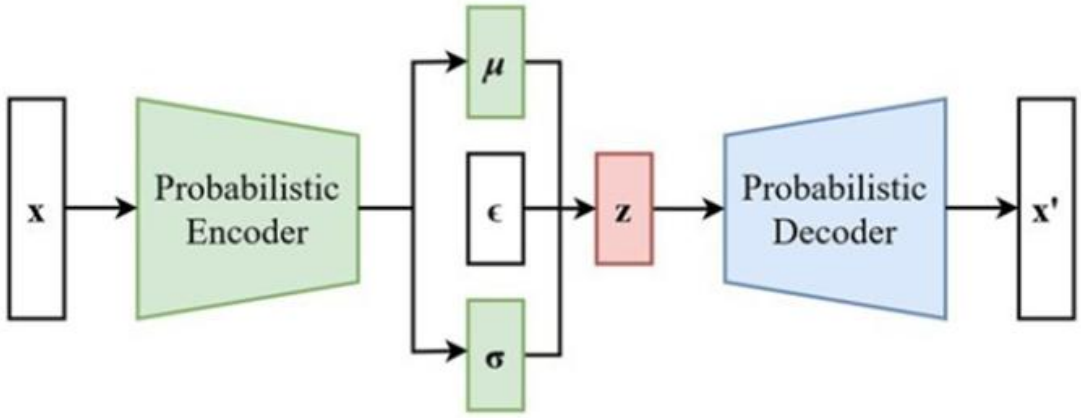

Figure 3 : : overview of VAEs architecture [12]

4.2.3. Autoregressive Models

Autoregressive models generate data sequentially, predicting each element based on preceding ones. Models like GPT (Generative Pre-trained Transformer) fall into this category, excelling in text generation tasks by producing coherent and contextually relevant content.[9]

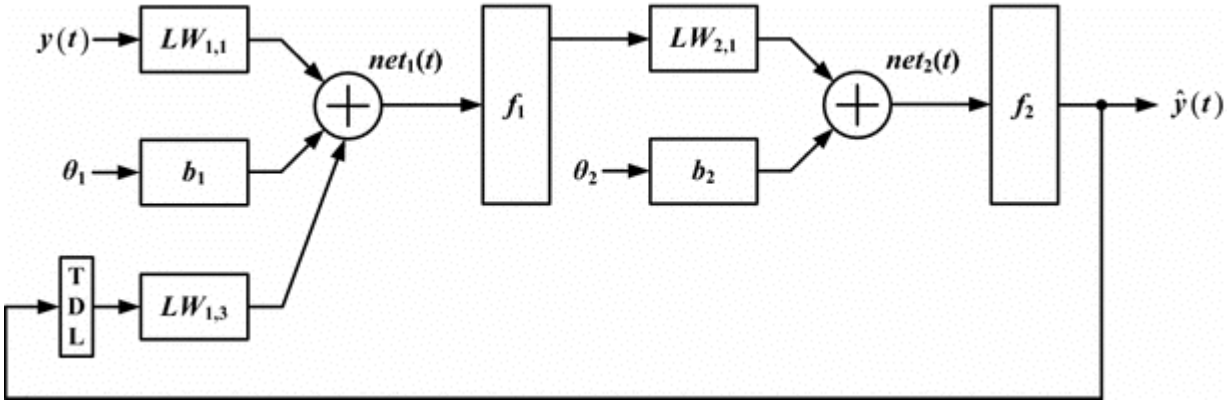

Figure 4 : overviez of Autoregressive Models architecture. [13]

4.2.4. Transformer-based Models

Transformers utilize self-attention mechanisms to process data, capturing long-range dependencies effectively. This architecture underpins models like BERT and GPT, which have achieved state-of-the-art performance in natural language processing tasks, including translation, summarization, and question-answering. [9]

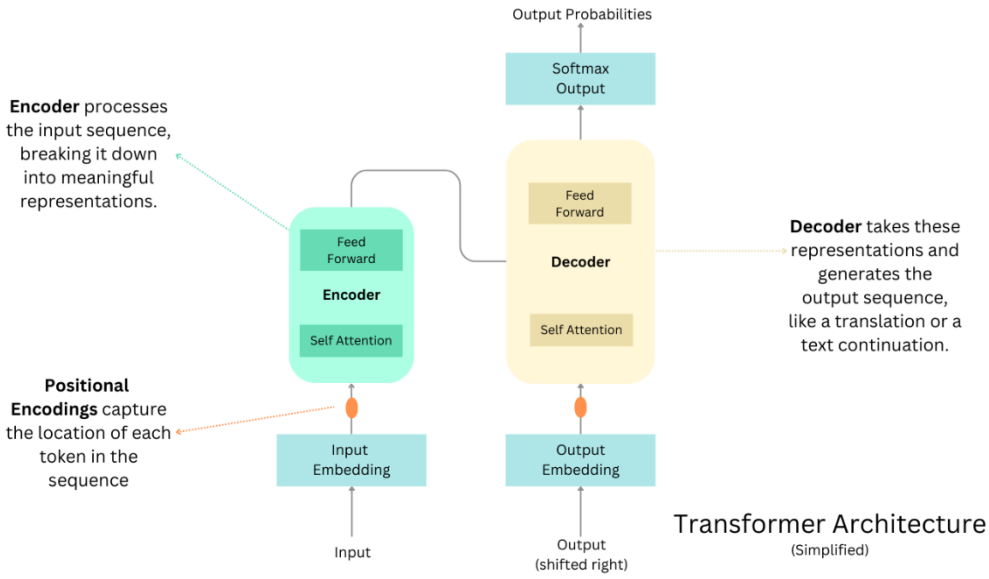

Figure 5 : Transformer-based Models architecture [14]

4.2.5. Diffusion Models

Diffusion models generate data by simulating a gradual denoising process, starting from random noise and refining it into coherent data. This technique has been employed in models like DALL·E 2 and Stable Diffusion, which are capable of producing high-quality images from textual descriptions.[9]

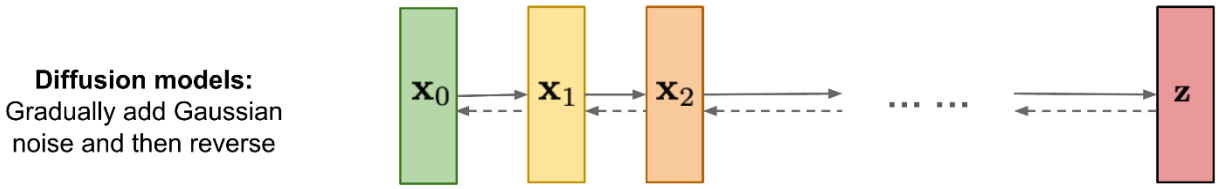

Figure 6 : Diffusion Models architecture [15]

4.2.6. Reinforcement Learning for Generative Tasks

Combining reinforcement learning with generative modeling allows systems to learn optimal generation strategies through trial and error. This integration enhances performance in complex tasks, enabling models to adapt and improve over time.[16]

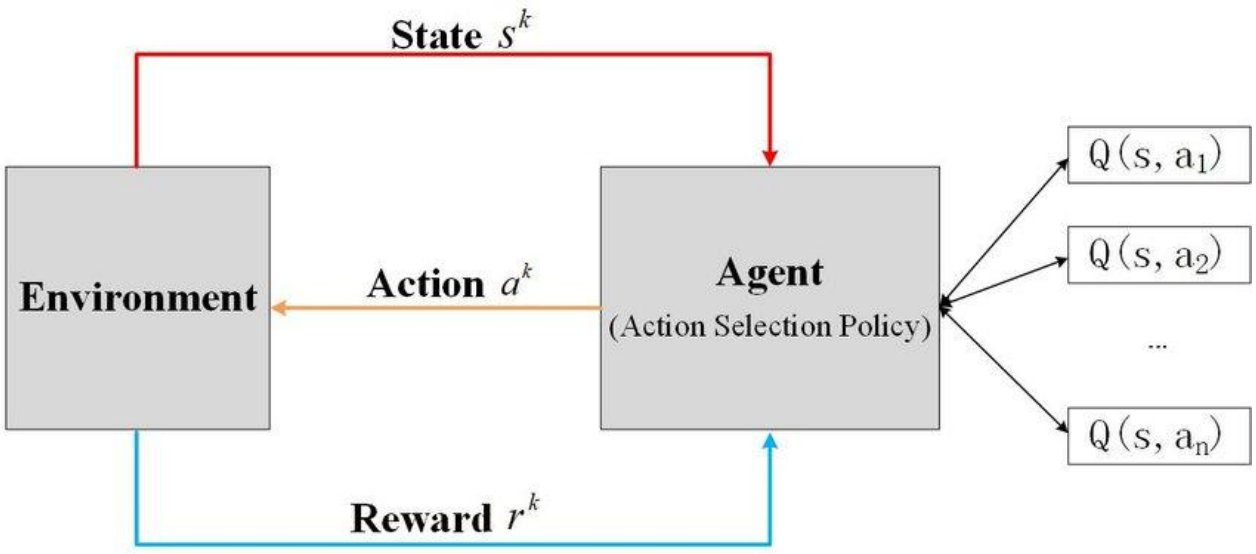

Figure 7 : Reinforcement Learning architecture [17]

4.3. Types of Generative Artificial Intelligence

Generative AI is versatile, with different models designed for specific tasks. Here are some types:

4.3.1. Text-to-Text

These models generate new textual content based on existing text inputs. They are widely used in applications such as content creation, summarization, translation, and conversational agents. Prominent examples include OpenAI's GPT series and Google's Gemini, which can produce coherent and contextually relevant text across various domains.[18]

4.3.2. Text-to-Image

Text-to-image models create visual representations from textual descriptions. They are instrumental in fields like design, advertising, and entertainment. Notable models include Google's Imagen, which utilizes diffusion techniques for high-fidelity image generation, and OpenAI's DALL·E, known for its ability to generate imaginative and diverse images from prompts.[19]

4.3.3. Image-to-Image

These models transform input images into modified versions, enabling tasks such as style transfer, super-resolution, and image restoration. For instance, they can convert sketches into photorealistic images or apply artistic styles to photographs. Techniques like conditional GANs are often employed in these transformations.[20]

4.3.4. Image-to-Text

Image-to-text models interpret visual content to generate descriptive textual outputs. Applications include image captioning, scene understanding, and assisting visually impaired individuals by describing images. Models like Google's Gemini can process images and provide detailed textual descriptions, enhancing accessibility and content understanding.[21]

4.3.5. Speech-to-Text

These models transcribe spoken language into written text, facilitating applications like voice assistants, transcription services, and real-time captioning. Advanced models can handle multiple languages and accents, improving communication and accessibility across diverse user groups.[22]

4.3.6. Text-to-Audio

Text-to-audio models generate spoken language or music from textual inputs. They are used in creating voiceovers, audiobooks, and virtual assistants. Models like Google's MusicLM can produce high-quality musical compositions from textual descriptions, expanding creative possibilities in music production.[23]

4.3.7. Text-to-Video

These models generate video content based on textual prompts, enabling the creation of animations, simulations, and visual storytelling. Google's Veo, for example, can produce high-definition videos from text inputs, supporting applications in filmmaking, education, and marketing.[24]

4.3.8. Multimodal AI

Multimodal AI models process and generate content across multiple data types, such as text, images, audio, and video. They enable more comprehensive understanding and generation capabilities, allowing for tasks like generating descriptive text from images or creating images based on textual and auditory inputs. Google's Gemini exemplifies this approach by integrating various modalities to perform complex tasks seamlessly.[25]

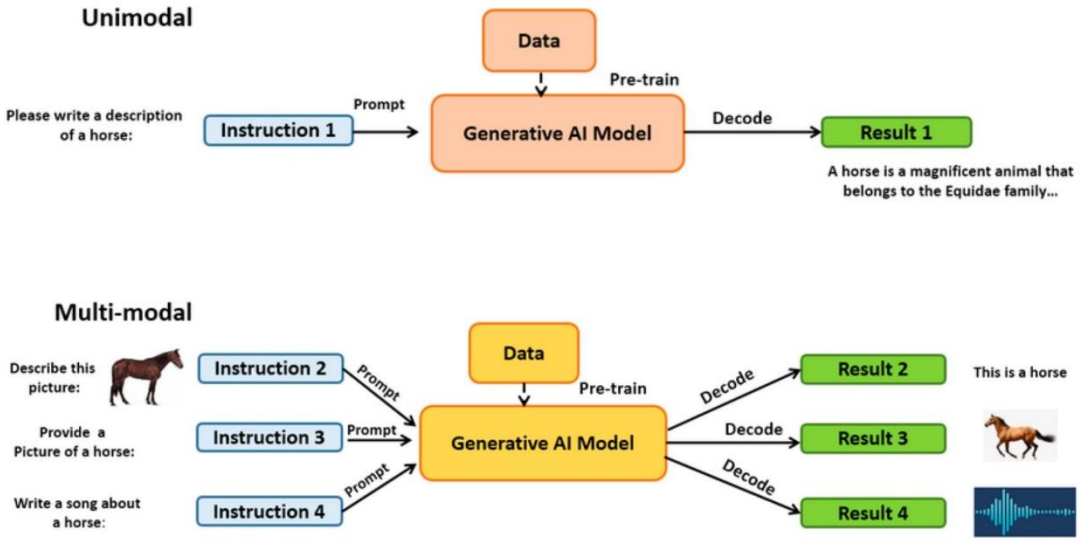

Figure 8 : Multi-modal AI VS Unimodal AI [26]

5. Generative Models for Video Synthesis

5.1 Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs) represent a true advancement in the field of artificial intelligence, particularly in generative modeling. GANs have since revolutionized the way machines can create realistic images, text, and even videos. The core idea behind GANs is the competition between two neural networks: the generator and the discriminator. The generator is responsible for producing synthetic data that mimics real-world examples, while the discriminator acts as a judge, trying to distinguish between real and generated data. These two networks are trained together in a process akin to a zero-sum game, where the generator continuously improves to fool the discriminator, and the discriminator enhances its ability to detect fakes. Over time, this dynamic training process leads to the creation of highly realistic outputs.

One of the key advantages of GANs is their ability to generate high-quality, diverse, and visually convincing content, making them particularly useful in applications such as image synthesis, style transfer, and even data augmentation. They have been widely adopted in various fields, from artificial image generation to medical imaging and video game development.

Despite their impressive capabilities, GANs come with significant challenges. One common issue is mode collapse, where the generator produces limited variations of data, leading to a lack of diversity. Another challenge is the difficulty in training, as the adversarial nature of GANs requires a delicate balance between the two networks to prevent instability or failure to converge.

The impact of GANs on modern AI research is profound, with ongoing advancements aimed at improving their efficiency, stability, and applicability in real-world problems. As research continues, GANs are expected to play an even more significant role in deep learning, pushing the boundaries of what AI can create.[27]

5.2 Diffusion-Based Models

Diffusion-based models have emerged as one of the most powerful approaches to generative modeling, offering a unique way to synthesize high-quality images, audio, and other complex data. These models operate by progressively introducing and then removing noise from data, allowing them to learn intricate structures within a dataset.

The key idea behind diffusion models is a two-step process: forward diffusion and reverse denoising. In the forward process, noise is incrementally added to data over multiple time steps, gradually transforming it into pure randomness. The reverse process, which is learned by a neural network, attempts to remove this noise step by step, reconstructing realistic samples from what initially seemed like noise. This method allows the model to generate highly detailed and diverse outputs. Diffusion models have strong theoretical foundations, drawing connections to probabilistic modeling and stochastic differential equations. Compared to older generative models like GANs (Generative Adversarial Networks) and VAEs (Variational Autoencoders), diffusion models often achieve superior quality, stability, and diversity in generated samples. Additionally, they avoid some of the common issues found in GANs, such as mode collapse, where the model generates only a limited variety of outputs.

One of the most exciting aspects of diffusion models is their adaptability. They can be conditioned on specific inputs, allowing users to guide the generation process, such as by providing a text prompt to generate an image. Techniques like classifier-free guidance enhance this capability by allowing fine-tuned control over the outputs without needing separate classification models.

Despite their advantages, diffusion models come with computational challenges. Training and sampling can be computationally expensive, requiring multiple iterative steps to produce a single output. Researchers are actively working on improving their efficiency, with techniques like latent diffusion models and score-based generative modeling showing promise in reducing computational costs while maintaining quality.

With continuous advancements, diffusion models are expected to play a central role in the future of AI-driven content creation, offering unprecedented control, realism, and creativity in generative tasks.[28]

5.3 Other Generative Approaches

While Generative Adversarial Networks (GANs) and Diffusion Models dominate the current landscape of generative media, several alternative approaches have also shown promising results in synthesizing video content.

5.3.1 Variational Autoencoders (VAEs)

Starting with the basics, variational autoencoders (VAEs) are a class of generative models in machine learning that can generate new data samples resembling the input data they were trained on. They also perform tasks common to other autoencoders, such as denoising. Like all autoencoders, VAEs are deep learning models composed of an encoder that learns to isolate the important latent variables from training data and a decoder that then uses those latent variables to reconstruct the input data.

However, whereas most autoencoder architectures encode a discrete, fixed representation of latent variables, VAEs encode a continuous, probabilistic representation of that latent space. This enables a VAE to not only accurately reconstruct the exact original input but also use variational inference to generate new data samples that resemble the original input data.

The neural network architecture for the variational autoencoder was originally proposed in a 2013 paper by Diederik P. Kingma and Max Welling, titled "Auto-Encoding Variational Bayes." This paper also popularized what they called the reparameterization trick, an important machine learning technique that enables the use of randomness as a model input without compromising the model's differentiability—that is, the ability to optimize the model's parameters. While VAEs are frequently discussed in the context of image generation, they can be used for a diverse array of artificial intelligence (AI) applications, from anomaly detection to generating new drug molecules.[29]

5.3.2 Transformers-based models

Transformer models have revolutionized deep learning by introducing a self-attention mechanism that allows for the efficient processing of sequential data. Unlike traditional recurrent neural networks (RNNs) and convolutional neural networks (CNNs), transformers can handle entire sequences in parallel, capturing long-range dependencies more effectively. This architecture has become foundational in various AI applications, including natural language processing, computer vision, and time series forecasting. Notably, models like BERT and GPT-3, which are based on transformer architecture, have set new benchmarks in tasks such as text generation, summarization, and question answering. The ability of transformers to process data in parallel also enables faster training and inference, making them suitable for large-scale AI applications.[30]

5.3.3 Flow-based models

Flow-based deep generative models have emerged as a compelling alternative to traditional generative approaches like Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs). These models leverage a series of invertible transformations to map simple probability distributions (e.g., Gaussian) to complex data distributions, enabling exact likelihood computation and efficient sampling.

In their comprehensive report, Xu and Dong (2020) talked and discussed various flow-based architectures, including NICE, RealNVP, Glow, Masked Autoregressive Flow (MAF), and Inverse Autoregressive Flow (IAF). They highlight the advantages of these models, such as tractable density estimation and invertibility, which facilitate both data generation and inference. Through experiments on the MNIST dataset, the authors demonstrate the effectiveness of flow-based models in generating high-quality samples, underscoring their potential in unsupervised learning tasks.[31]

In summary, while GANs and diffusion models currently dominate the field, alternative generative methods continue to evolve, offering trade-offs between training stability, interpretability, and fidelity.

6. Definition and Characteristics of Video

Videos are a digital medium that captures and stores sequences of moving images, often accompanied by audio, for playback and analysis. According to Baeldung, videos are obtained using image-acquisition devices such as video cameras, smartphones, and camcorders. They can be categorized into two types: analog and digital.

1. **Analog Video :** Analog videos represent the earliest form of video technology. They are captured as continuous signals and stored on mediums like magnetic tapes. A sensitive plate captures a scene at an instance, and electrodes read line by line from left to right. A single reading from top to bottom of a photosensitive plate by an electrode is called a "frame." Consequently, a complete video consists of several frames displayed sequentially at a standard speed. Initially, these videos were monochrome, with electrodes representing only black and white. When the electrodes read the photosensitive plate, they output voltages, generating high signals for high voltage, creating analog signals.
2. **Digital Video :** The digitization of videos involves converting frames into a digital format. Each frame becomes a separate image, allowing for manipulation of parameters such as frame rate, depth, size, and resolution. Semiconductor-based sensors record the frames that make up digital movies. A frame structure is a matrix of elements holding pixel values, with the number of rows and columns indicating the frame size.

Characteristics of a Video :

Frames are technically independent images. Storing and playing these frames sequentially at a standard speed creates a video. The frame rate specifies the speed of the video; for example, 20 frames per second indicates reading and displaying 20 frames each second. The aspect ratio informs the ratio between width and height for displaying video, with the standard aspect ratio being 4:3. Color depth defines how visually appealing the video looks, with bits per pixel indicating the number of colors a pixel can display. The compression method used and the number of pixels utilized to represent the frames define the quality of the videos. [32]

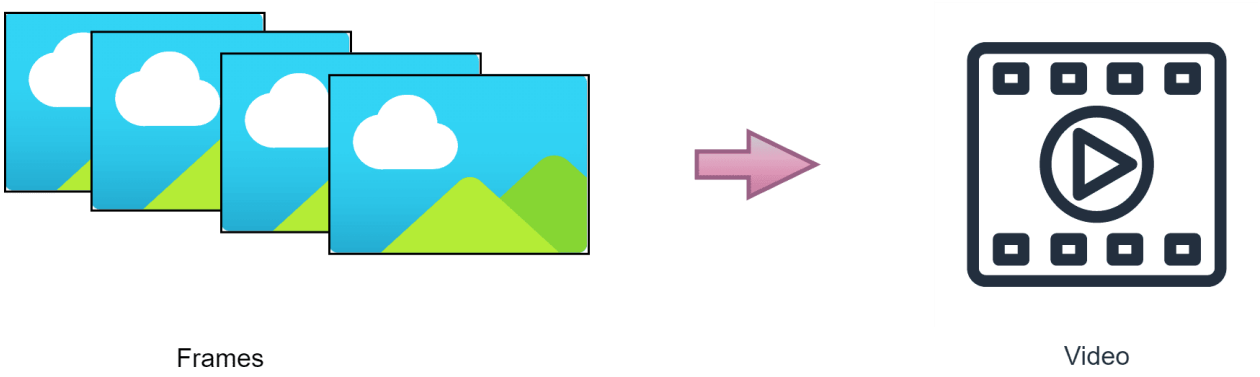

Figure 9 : Video components [33]

7. Applications of AI-Generated Videos

7.1 Cinema and Media Production

Artificial Intelligence has increasingly permeated the realm of cinema and media production, revolutionizing traditional filmmaking processes. In the study by Sun (2024), the application of AI in film production is examined across various stages.

Scriptwriting and Pre-Production

AI technologies have been employed in the scriptwriting phase, where natural language processing algorithms can generate story ideas, dialogues, and even entire scripts. This automation accelerates the creative process and offers novel narrative possibilities. In pre-production, AI assists in tasks such as casting by analyzing actors' past performances and predicting audience reception, as well as in location scouting through image recognition and data analysis.

Production

During the production phase, AI contributes to camera work and scene composition. For instance, intelligent cinematography systems can suggest optimal camera angles and movements based on scene analysis. AI-driven tools also facilitate real-time editing and special effects integration, enhancing efficiency on set.

Post-Production

In post-production, AI plays a significant role in editing, color grading, and visual effects. Machine learning algorithms can automate editing decisions by selecting the best takes and assembling sequences that align with the director's vision. AI also enables the enhancement of visual effects, allowing for more realistic and immersive experiences.

Advantages and Limitations

The integration of AI in film production offers numerous benefits, including :

- Increased efficiency
- Cost-effectiveness
- The ability to explore innovative storytelling techniques.

However, the study acknowledges limitations, such as : the potential loss of human touch and depth in AI-generated content. While AI can mimic human creativity to an extent, it may lack the nuanced understanding and emotional resonance that human creators bring to their work.[34]

7.2 Deepfakes and Image Manipulation

Artificial Intelligence has brought about a fundamental shift in the manipulation of visual and audio content, most notably through the development of deepfakes. These are synthetic media - often videos - produced using advanced machine learning techniques, such as deep learning, morphing, and warping, to fabricate realistic but entirely false audiovisual content. According to Boté-Vericad and Vállez (2022), deepfakes involve the deliberate manipulation of images, voices, and gestures in ways that can deceive the viewer by distorting time, place, and context. This makes them especially powerful, and potentially dangerous, tools of misinformation. At a technical level, creating a deepfake often involves multiple stages of manipulation. Methods such as morphing transform facial structures or fuse identities by gradually converting one image into another. Warping modifies image geometry for either enhancement or distortion. These techniques are enhanced through artificial intelligence, enabling more complex operations like facial reanimation, voice synthesis, and lip-syncing. While user-friendly mobile applications have made simple deepfake generation accessible to the general public, producing highly realistic and believable deepfakes still requires expertise across linguistics, video production, and animation.

Detection, however, remains a significant challenge. Tools such as recurrent neural networks are employed to identify frame-level inconsistencies, while other approaches focus on biometric anomalies - like unnatural eye-blinking patterns - or metadata inconsistencies. Emerging techniques include blockchain-based authentication and multimedia forensic analysis, though these remain costly and technically demanding.

The societal implications of deepfakes are complex. While their use in satire and humor is often tolerated or even embraced - evident in television programs like *El Intermedio*, which regularly manipulate political videos for comedic effect - the same tolerance does not extend to contexts such as journalism or political discourse. In these arenas, deepfakes pose severe risks to public trust, media credibility, and democratic processes. Furthermore, manipulated videos have found application in sectors like healthcare, where they are used to simulate outcomes of cosmetic surgeries, raising further ethical questions about authenticity and informed consent. Privacy concerns also loom large. Many deepfake-related applications fail to transparently communicate how they collect, process, or share user data. Studies cited by the authors reveal that a significant number of apps either lack proper privacy policies or present them in language too complex for the average user to understand. This raises substantial concerns about user autonomy and data protection.

In sum, the rise of AI-generated deepfakes underscores both the creative potential and the ethical hazards of modern video manipulation technologies. As these tools continue to evolve, so too must our strategies for regulation, detection, and public education. Ensuring that digital literacy and ethical considerations keep pace with technological innovation is essential for safeguarding both individuals and democratic institutions.[35]

7.3 Ethical and Societal Implications

The integration of Artificial Intelligence has brought about significant changes, offering both opportunities and challenges. While AI tools enhance efficiency and open new creative avenues, they also raise important ethical and societal questions.

Redefining Creativity and Authorship

AI's ability to generate content blurs traditional notions of authorship. When AI systems produce music, scripts, or visual art, it's unclear who should be credited - the developer, the user, or the AI

itself. This ambiguity challenges existing frameworks for intellectual property and creative ownership.

Employment and the Creative Workforce

As AI automates tasks like video editing and content creation, there's concern about its impact on employment within the creative industries. While AI can handle repetitive tasks, potentially allowing human creators to focus on more complex aspects, there's a risk of job displacement if not managed carefully.

Bias and Representation

AI systems learn from existing data, which may contain societal biases. If not addressed, AI-generated content can perpetuate stereotypes or misrepresent certain groups. Ensuring diversity and fairness in AI outputs requires careful dataset selection and ongoing monitoring.

Intellectual Property Challenges

Current laws are not fully equipped to handle creations generated by AI. Determining ownership rights for AI-produced content is complex, especially when such content closely resembles existing works. This necessitates updates to legal frameworks to address the unique nature of AI-generated media.

Audience Perception and Trust

The use of AI in media can affect how audiences perceive content. Some viewers may question the authenticity or emotional depth of AI-generated works. Transparency about AI's role in content creation is crucial to maintain audience trust and engagement.

Balancing Innovation with Ethical Considerations

While AI offers tools that can democratize content creation and enhance productivity, it's essential to implement guidelines that ensure ethical use. This includes promoting transparency, preventing misuse, and safeguarding the interests of human creators.[36]

Conclusion

Generative AI is transforming the landscape of video creation, offering unprecedented tools for innovation in media, entertainment, and beyond. From its foundations in artificial intelligence to the cutting-edge models driving today's advancements, this technology is reshaping how we produce and experience visual content. As generative AI continues to evolve, it promises to push the boundaries of creativity, enabling new forms of storytelling and immersive experiences. The future of video lies at the intersection of human imagination and machine intelligence—ushering in a new era of limitless possibilities.

Chapter 2: State of the Art

Introduction

AI-generated video has evolved from simple predictive models to sophisticated systems capable of creating hyper-realistic content. This chapter explores today's cutting-edge techniques, from GANs to diffusion models, while examining the ongoing challenge of detecting synthetic media. We'll analyze key research, tools, and datasets driving progress—and the limitations we still face in this rapidly advancing field.

The race between generation and detection continues, with profound implications for media authenticity. Here's where the technology stands today.

1. Evolution of Video Generation Techniques

Over the years, the task of generating realistic videos has progressed from simple predictive models to powerful generative architectures. This evolution reflects the broader trajectory of deep generative modeling, blending advances in computer vision, natural language processing, and probabilistic modeling. In this section, we trace the major developments in video generation through four main phases: early predictive models, GAN-based generation, autoregressive and transformer models, and the recent surge in diffusion-based approaches.

1.1 Early Attempts: Frame Prediction and Autoencoders

The first wave of video generation research focused on predicting the next frames in a video sequence based on previous ones. These models - often built from convolutional LSTMs, RNNs, and variational autoencoders (VAEs) - were limited in their expressiveness. Their outputs typically lacked high-resolution details and often suffered from blur and temporal inconsistency.

These methods treated video generation as a frame-by-frame prediction task, ignoring higher-order structures like motion trajectories and global semantics. Although they laid the foundation for sequence modeling, their inability to model complex dynamics restricted their usefulness in high-quality synthesis.

1.2 GAN-Based Video Generation

The introduction of Generative Adversarial Networks (GANs) brought a paradigm shift in generative modeling. By pitting a generator against a discriminator, GANs enabled the creation of

sharper and more realistic frames. This innovation quickly translated into video generation via extensions such as:

- VideoGAN: One of the earliest models to extend GANs to the temporal domain.
- TGAN (Temporal GAN): Incorporated 3D convolutions to handle temporal information.
- MoCoGAN: Proposed disentangled modeling of content (what) and motion (how), allowing for better control and generalization.

Despite their success in generating short, high-fidelity video clips, GANs exhibited several limitations:

- Mode collapse: They tended to produce repetitive or low-diversity outputs.
- Temporal instability: Even when individual frames were sharp, coherence across time was often inconsistent.
- Difficult training: Adversarial learning was notoriously unstable, requiring careful tuning and heuristics.

1.3 Transformer-Based Models for Video

The transformer architecture, known for revolutionizing NLP, also began to influence video generation. Transformers provided two key advantages:

- Long-range dependency modeling: They captured relationships across distant frames.
- Flexible sequence length: They were better suited to varying-length sequences than RNNs.

Examples of transformer-based video generators include:

- VideoGPT: Used a VQ-VAE encoder to tokenize video frames, followed by an autoregressive transformer to model sequences.
- MaskGit for Video: Extended masked token prediction to temporal sequences.
- NUWA and CogVideo: Incorporated large language models and multi-modal inputs for text-to-video generation.

While transformer models improved temporal coherence and content diversity, they were resource-intensive, often requiring enormous training data and memory. Additionally, autoregressive generation remained slow and was prone to cascading errors over long sequences.

1.4 Diffusion Models: A New Paradigm

The most promising recent development is the diffusion-based approach to video generation. Inspired by statistical physics, diffusion models learn to reverse a gradual noise-adding process. They operate by generating random noise and progressively denoising it into coherent video frames.

Key advantages of diffusion models include:

- Stable training compared to GANs (no adversarial loss).
- High-quality outputs, both spatially and temporally.
- Flexible conditioning, supporting text, motion, and image inputs.
- Notable Diffusion-Based Architectures:

Video Diffusion Models (VDM): Extended DDPMs to the spatiotemporal domain.

Video-LDM: Introduced latent diffusion for videos, greatly reducing memory use while improving quality.

Align Your Latents (AYL): Focused on improving motion alignment and semantic consistency using latent constraints.

Diffusion models outperform GANs and autoregressive models in both fidelity and diversity, making them state-of-the-art in many benchmarks. They are also more robust to noise and require less heuristic tuning.

1.5 Conditional and Multimodal Video Generation

Another important trend is the shift toward controllable and conditional generation. Recent models can generate videos from:

- Text prompts (e.g., "**a dog running in snow**")
- Static images or keyframes
- Semantic maps or motion trajectories

These capabilities are powered by cross-modal transformers and CLIP-like encoders. As a result, video diffusion models are now capable of text-to-video generation, inpainting, and style transfer, broadening their utility in creative and industrial applications.

1.6 Challenges and Outlook

Despite the impressive progress, several challenges remain:

- Long video generation is still difficult due to computational limits and memory constraints.
- Semantic coherence across long sequences can break down, especially when generating dynamic scenes.
- Training cost is still a barrier for small labs or individual researchers.

Future work is likely to focus on efficiency, scalability, and better user control. Integration with 3D modeling, physics simulators, or real-time rendering systems may further extend capabilities.

The evolution of video generation techniques reflects a broader trajectory in deep learning: from simple, frame-level predictors to highly sophisticated, controllable generative systems. GANs introduced visual realism, transformers added temporal depth, and diffusion models now provide the best of both worlds—delivering high-quality, temporally consistent, and controllable video outputs.[37]

2. Review of Existing Research on AI-Generated Video Detection

The spread of AI-generated videos, particularly deepfakes, has prompted a significant shift in digital media forensics. The line between real and fabricated video content is increasingly blurred, thereby presenting critical threats to information integrity, individual privacy, and social trust. In this context, the academic community has developed a variety of detection strategies aimed at addressing the growing sophistication and accessibility of generative models. The goal of this section is to present a comprehensive overview of existing research efforts in AI-generated video detection, focusing specifically on the methodologies, benchmarks, and scientific contributions that have directly shaped the current state of the field. This literature review lays the groundwork for our own project, which aims to build upon the most promising approaches and propose an integrated application capable of real-time, generalizable detection of synthetic video content.

2.1 Benchmark Datasets: Foundation of Detection Research

A crucial driver behind progress in detection research has been the development of high-quality, large-scale datasets that capture the diversity of both real and AI-generated video content.

These datasets serve as the training and validation backbone for machine learning models and allow for standardized evaluation protocols.

One of the most comprehensive benchmarks introduced recently is **GenVideo**, a dataset curated by Chen et al. (2024) [38], encompassing over one million samples. It includes both real and synthetically generated videos across multiple domains and offers evaluation settings that test cross-generator generalization and degraded video robustness - two challenges particularly relevant for real-world deployment. This dataset is pivotal for training detectors that are not overfitted to a particular generation method but instead exhibit broad generalizability.

Similarly, **GenVidDet**, proposed by Ji et al. (2024) [39], focuses on large-scale learning by providing 2.66 million labeled instances with diverse resolutions, frame rates, and semantic categories. It supports temporal learning and motion-sensitive detection, a key requirement for distinguishing subtle patterns in deepfakes, such as mismatched lip-sync or inconsistent motion blur.

Other specialized datasets, such as **Chameleon** (Zeng et al., 2025) [40], are designed to capture scene transitions, perspective variation, and domain shifts, bringing detection efforts closer to real-world forensics. These datasets not only simulate realistic broadcast and social media scenarios but also introduce the concept of spatiotemporal complexity, which is often lost in synthetic video detection approaches limited to frame-by-frame analysis.

2.2 Key Detection Strategies and Contributions

Several detection architectures have emerged in response to the growing sophistication of video generation models. These approaches differ in their underlying methodology but converge on the common goal of identifying statistical, visual, or motion-based anomalies introduced during the synthesis process.

DeMamba, proposed by Chen et al. (2024) [38], represents a state-of-the-art detection head designed to integrate into existing vision transformers. Rather than relying solely on spatial inconsistencies, DeMamba incorporates multi-scale spatiotemporal features, enabling the identification of fine-grained generation artifacts. Its plug-and-play nature makes it suitable for integration into our project's modular detection pipeline.

Another significant contribution is **DuB3D** (Dual-Branch 3D Transformer), developed by Ji et al. (2024) [30]. This architecture emphasizes the fusion of two complementary streams: raw RGB frames and computed optical flow. By incorporating motion dynamics, DuB3D addresses the common shortfall of spatial-only detectors and enhances resilience against subtle generation effects such as hyperrealistic skin textures or lighting coherence. In benchmarking scenarios, DuB3D surpassed traditional convolutional architectures by achieving 96.77% detection accuracy.

Likewise, the **AIGVDet** framework (Bai et al., 2024) [41] employs a dual-branch CNN architecture focused on anomaly detection within both the spatial and flow domains. The method aligns with the anomaly-based paradigm used in video surveillance and has demonstrated robust performance across unseen generation methods, suggesting strong generalization potential - a trait that we consider essential for our detection application.

These models collectively highlight a shift away from frame-level detection toward spatiotemporal modeling, which is more reflective of how videos are consumed and interpreted by humans. Our project draws directly from these findings to design a hybrid model capable of learning both static inconsistencies and motion-aware anomalies.

2.3 Relevance to Our Detection Application

In developing an intelligent application for AI-generated video detection, the reviewed research informs both the **architectural choices** and **evaluation methodologies** of our system. From a technical standpoint, the reviewed literature advocates for multi-stream, multi-modal architectures that combine spatial, temporal, and semantic features.

Furthermore, the reviewed benchmarks and datasets will serve as critical resources for our model training and validation pipeline.

Equally important, the reviewed literature underscores the necessity for cross-domain robustness. This aligns with our decision to test the system not only on high-resolution synthetic media but also on degraded formats typical of social media platforms (compressed, low-frame-rate, mobile-recorded content).

2.4 Limitations and Unexplored Gaps

Despite remarkable progress, existing detection approaches face several limitations. First, **real-time performance** is still a bottleneck; many state-of-the-art models are too computationally intensive for lightweight deployment. Second, **cross-modal synthesis** detection - for example, videos generated from text prompts using diffusion models - remains underexplored. These gaps are directly relevant to our project's second-phase goals, which include optimizing detection speed and extending detection capabilities to multi-modal generation techniques.

In addition, **explainability** is rarely addressed in current literature. Most models act as black boxes, providing little insight into why a video is flagged as fake. Given the ethical implications of false positives, especially in legal or journalistic contexts, our application includes a visualization module that highlights key regions and motion anomalies responsible for the model's decision - an approach informed by XAI (Explainable AI) frameworks.

3. Tools, Frameworks, and Available Datasets

The rapid advancement of generative models, particularly in the realm of deepfakes, has necessitated the development of robust tools and datasets to detect and mitigate potential misuse. This section delves into the prominent tools, frameworks, and datasets that have been instrumental in the field of AI-generated video detection, providing insights into their functionalities and applications.

3.1 Tools and Frameworks

3.1.1 DeepFaceLab

DeepFaceLab stands out as a comprehensive open-source tool designed for creating deepfakes. Developed by Perov, it offers a flexible and extensible framework for face-swapping tasks, enabling users to produce high-fidelity synthetic videos. Its modular architecture allows for customization, making it a valuable resource not only for generating deepfakes but also for understanding the underlying mechanisms, which is crucial for developing effective detection strategies.

3.1.2 FaceForensics++

FaceForensics++ is both a dataset and a framework aimed at facilitating research in facial manipulation detection. Introduced by Rössler et al. (2019) [42], it encompasses over 1,000 videos manipulated using techniques like Deepfakes, Face2Face, FaceSwap, and NeuralTextures. The dataset provides binary masks for each manipulated frame, supporting both classification and segmentation tasks. The accompanying codebase aids in training and evaluating detection models, making it a cornerstone resource in the field.

3.1.3 DFDC Deepfake Challenge Solution

The Deepfake Detection Challenge (DFDC), organized by Facebook AI, spurred the development of advanced detection algorithms. Among the top submissions was the solution by selimsef, which employs a frame-by-frame classification approach using EfficientNet architectures. The pipeline incorporates face detection, cropping, and heavy data augmentations to enhance model robustness. The publicly available codebase serves as a practical reference for implementing end-to-end detection systems.

3.1.4 Reality Defender

Reality Defender is a multi-model deepfake detection platform designed to analyze AI-generated content across video, images, audio, and text. Unlike traditional tools that rely on watermarks or prior authentication, Reality Defender uses probabilistic detection, allowing it to spot deepfake manipulation in real-world scenarios. This platform is widely used in government, media, and financial sectors to combat voice impersonation, document forgery, and AI-generated disinformation. It has also been adopted by public broadcasting companies in Asia and multinational banks, helping to prevent identity fraud and synthetic media threats. [43]

3.1.5 VastavX AI

VastavX AI is an artificial intelligence-based deepfake detection system developed by Zero Defend Security. Released in 2025, it is India's first deepfake detection technology, designed to analyze and verify digital media for authenticity. The system detects AI-generated videos, images, and audio with a reported accuracy of 99%, providing forensic insights into manipulated content. VastavX AI employs advanced machine learning techniques, forensic analysis, and metadata inspection to distinguish between real and AI-generated content. The platform is designed for use

by law enforcement agencies, media organizations, cybersecurity firms, and individuals seeking to verify digital content integrity.[44]

3.2 Available Datasets

3.2.1 FaceForensics++

As previously mentioned, FaceForensics++ offers a diverse set of manipulated videos, catering to various compression levels and manipulation techniques. Its extensive annotations and binary masks make it a standard benchmark for evaluating detection algorithms.

3.2.2 Deepfake Detection Challenge (DFDC) Dataset

The Deepfake Detection Challenge (DFDC) dataset, released by Facebook AI, is one of the most comprehensive collections available for deepfake detection research. It comprises over 100,000 videos, both real and manipulated, sourced from 3,426 paid actors. These videos were generated using a variety of deepfake techniques, including GAN-based face swapping methods. The dataset reflects real-world scenarios with variations in lighting, pose, and background, making it a valuable resource for training and evaluating deepfake detection models. The DFDC dataset has been instrumental in advancing the development of detection algorithms capable of generalizing to diverse and challenging conditions.[45]

3.2.3 Celeb-DF

Celeb-DF is a large-scale, high-quality deepfake video dataset introduced to address the limitations of previous datasets, which often suffered from low visual quality and lacked diversity. It contains 5,639 deepfake videos of celebrities, generated using an improved synthesis process that enhances visual fidelity. The dataset presents a challenging benchmark for detection algorithms, especially in distinguishing subtle manipulations in high-resolution videos. Celeb-DF has been widely adopted in the research community to evaluate and improve the robustness of deepfake detection methods.[46]

3.2.4 GenVidBench

GenVidBench is a challenging AI-generated video detection dataset designed to facilitate the development of robust detection models. It offers several key advantages:

1. **Cross Source and Cross Generator**: GenVidBench includes videos generated from multiple sources and generators, mitigating the interference of video content on detection and ensuring diversity in video attributes between the training and test sets.[47]
2. **State-of-the-Art Video Generators**: The dataset encompasses videos from eight state-of-the-art AI video generators, covering the latest advancements in video generation techniques.
3. **Rich Semantics**: Videos in GenVidBench are analyzed from multiple dimensions and classified into various semantic categories based on their content, enhancing the dataset's diversity and aiding in the development of more generalized detection models. [48]

3.2.5 DeeperForensics-1.0

DeeperForensics-1.0 is a large-scale, high-quality dataset introduced by Jiang et al. (2020) [49] to push the boundaries of generalizable deepfake detection. It comprises over **60,000 videos** generated from 100 real identities with different manipulation settings such as various lighting conditions, compression levels, and perturbations. One of the most significant contributions of this dataset is its effort to bridge the domain gap between synthetic and real-world conditions by introducing realistic distortions.

This helps researchers evaluate the robustness of their models in challenging environments. The dataset has been widely adopted in benchmarking tasks that focus on cross-manipulation generalization and real-world detection performance.

3.3 Benchmarking and Evaluation Protocols

Beyond tools and datasets, an essential component of research in this area is the **evaluation strategy**. Benchmarking models for AI-generated video detection requires consistent, fair protocols to measure robustness, generalizability, and performance across diverse manipulations and datasets.

Most academic evaluations use **AUC (Area Under the Curve)**, **accuracy**, and **F1 scores**. However, these metrics alone are insufficient when models are deployed in the wild. Researchers increasingly advocate for evaluations that simulate real-world conditions, such as **cross-dataset**

testing (training on one dataset and testing on another), and **adversarial robustness tests**, where the content is intentionally obfuscated or augmented to deceive detection algorithms (Dolhansky et al., 2020).[45]

Some challenges, like the **FaceForensics++ benchmark** and the **DFDC leaderboard**, already offer standardized protocols, but these often need adaptation to stay relevant with emerging manipulation techniques. For instance, new benchmarks like **GenVidBench** go further by organizing datasets semantically and by generation technique, encouraging a multidimensional evaluation of detection algorithms (Wang et al., 2024).[50]

To sum up, the field of AI-generated video detection is underpinned by a rich ecosystem of tools, frameworks, and datasets. Open-source detection pipelines like DFDC solutions, realistic and diverse datasets like FaceForensics++ and GenVidBench, and recent advances in adversarial robustness have all propelled this domain forward. As we continue developing our detection system, these resources not only serve as references but also as practical components to validate and enhance our approach.

4. Limitations and Challenges of Current Methods

4.1 Generalization to Unseen Manipulations

A significant challenge in deepfake detection is the limited ability of models to generalize to unseen manipulations. Many detectors are trained on specific datasets, leading to overfitting and poor performance when encountering new types of deepfakes. This issue arises because models often learn dataset-specific artifacts rather than universal features of manipulated content. For instance, detectors trained on the FaceForensics++ dataset may struggle with deepfakes from the Celeb-DF dataset due to differences in generation techniques and visual artifacts. To address this, researchers have proposed frameworks like GM-DF, which aim to enhance generalization across multiple scenarios by leveraging domain-aware meta-learning strategies .[51]

4.2 Vulnerability to Adversarial Attacks

Deepfake detection models are susceptible to adversarial attacks, where slight perturbations in the input can lead to misclassification. Attackers can exploit this vulnerability to bypass detection systems, raising concerns about the robustness of current methods. Studies have demonstrated that

even minor modifications can significantly degrade the performance of state-of-the-art detectors. For example, the 2D-Malafide attack introduces subtle perturbations that can mislead detection systems without noticeable changes to the human eye .[52]

4.3 Lack of Explainability

Many deep learning-based detectors function as "black boxes," providing little insight into their decision-making processes. This lack of transparency hinders trust and makes it difficult to understand why a particular video is classified as fake or real. Developing explainable AI models is crucial for user trust and for forensic analysis. Recent research has focused on integrating explainable AI techniques, such as network dissection algorithms, to interpret the internal workings of convolutional neural networks used in deepfake detection.[53]

4.4 Real-Time Detection Constraints

Implementing deepfake detection in real-time applications, such as live video streams, poses significant challenges. High computational requirements and latency issues make it difficult to deploy detection systems in scenarios where immediate verification is necessary. Optimizing models for speed without compromising accuracy remains an ongoing research area. Additionally, the rapid evolution of deepfake generation techniques outpaces the development of efficient detection algorithms, further complicating real-time detection efforts .[54]

4.5 Ethical and Privacy Concerns

The deployment of deepfake detection systems raises ethical and privacy issues. For instance, scanning user-generated content for deepfakes could infringe on privacy rights. Additionally, the potential for false positives may lead to unwarranted censorship or reputational damage. Balancing detection efforts with ethical considerations is essential. The recent passage of the "Take It Down Act" in the U.S. highlights the growing concern over non-consensual deepfake content and the need for legal frameworks to address such issues .[55]

Conclusion

AI video generation has made huge strides, from basic predictive models to today's ultra-realistic diffusion systems. But as the tech improves, so do the challenges around detection and ethics. While current models still struggle with long-term consistency and computing costs, they're

advancing fast. The future of video is being rewritten—we'll need equally smart solutions to handle what's coming next.

Chapter 3: Detection Techniques

Introduction

The battle to detect AI-generated videos unfolds across two fronts. On one side, classical forensic methods meticulously analyze visual and temporal artifacts—those subtle imperfections in blinking patterns, lip movements, and facial lighting that betray synthetic origins. These techniques rely on the telltale signs that generative models still struggle to perfect. Meanwhile, deep learning approaches bring automated pattern recognition to the fight, with convolutional neural networks scanning individual frames for anomalies while recurrent architectures track inconsistencies across time. This chapter examines both paradigms, revealing how each contributes to the evolving science of deepfake detection and why hybrid systems are increasingly becoming the standard.

1. Classical Detection Methods

1.1 Analysis of Visual Artifacts

Early deepfake detection techniques primarily focused on identifying visual inconsistencies introduced during the synthesis process. These artifacts often arise due to limitations in generative models and can manifest in various forms:

- **Facial Inconsistencies**: Deepfake algorithms sometimes produce unnatural facial features, such as asymmetrical expressions, inconsistent lighting, etc. These anomalies can be detected through careful analysis of facial regions (Concas et al., 2024).[56]
- **Eye Blinking Patterns**: Research has shown that deepfake videos often exhibit abnormal eye blinking rates, either blinking too frequently or not at all, due to the training data lacking sufficient closed-eye images (Lyu, 2019).[57]
- **Lip-Sync Discrepancies**: Mismatches between lip movements and spoken words are common in deepfakes, as the synchronization between audio and visual components is challenging to perfect (Astrid et al., 2025).[58]
- **Hand and Finger Anomalies**: Generative models may struggle to accurately render hands and fingers, leading to unnatural movements or incorrect numbers of fingers.

Figure 10 : Example of Hand and Finger Anomalies [59]

These visual artifacts serve as crucial indicators for detecting manipulated media, especially when combined with forensic analysis techniques.

1.2 Temporal Inconsistencies

Beyond static visual artifacts, temporal inconsistencies across video frames can reveal the presence of deepfakes. These inconsistencies arise due to the frame-by-frame generation approach of many deepfake algorithms, leading to unnatural transitions and movements:

- **Motion Discontinuities**: Real videos exhibit smooth and coherent motion, while deepfakes may display abrupt or unnatural movements, particularly in facial expressions and head movements (Gu et al., 2021).[60]
- **Identity Inconsistencies**: The Temporal Identity Inconsistency Network (TI²Net) focuses on detecting variations in identity features over time, which are often present in deepfake videos due to inconsistent rendering of facial attributes (Liu et al., 2023).[61]
- **Audio-Visual Mismatches**: Discrepancies between audio cues and visual elements, such as lip movements not aligning with speech, can indicate manipulated content (Astrid et al., 2025).

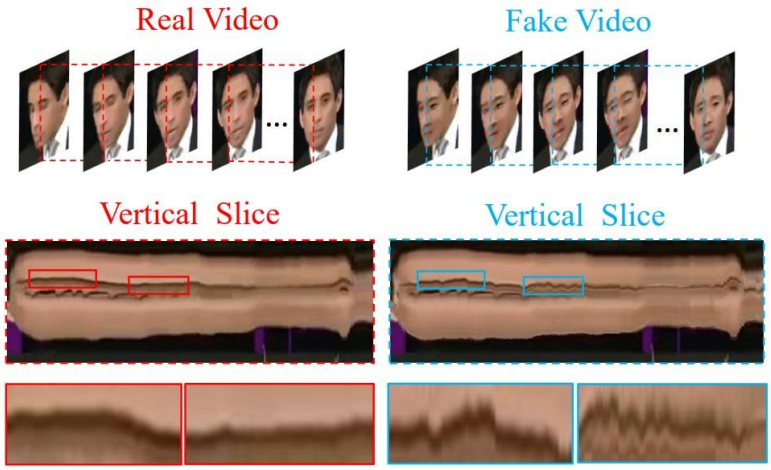

Figure 11 : Illustration of the temporal inconsistency between real and fake video [62]

Analyzing these temporal aspects enhances the robustness of deepfake detection systems by capturing inconsistencies that are not apparent in individual frames.

2. Deep Learning Approaches

2.1 Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) are a specialized class of deep learning models designed to process data with a grid-like topology, such as images. They have become fundamental tools in deepfake detection due to their ability to learn and extract hierarchical patterns from visual data - from low-level textures to high-level semantic features. CNNs operate through layers of learnable filters (or kernels), which convolve across the input to generate feature maps that emphasize relevant spatial characteristics. This ability to perform localized processing enables CNNs to detect subtle artifacts typically introduced during deepfake generation, such as inconsistencies in skin texture, unnatural lighting transitions, or boundary mismatches (Bonettini et al., 2020).[63]

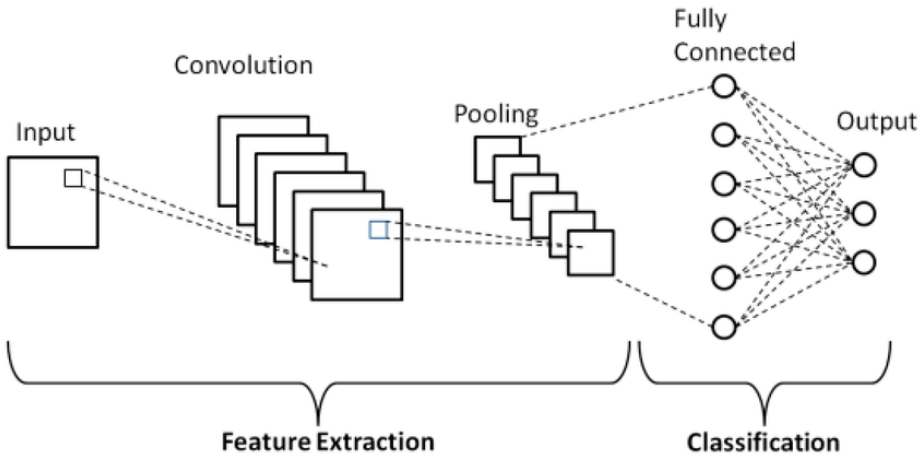

Figure 12 : basic convolutional neural network (CNN) architecture [64]

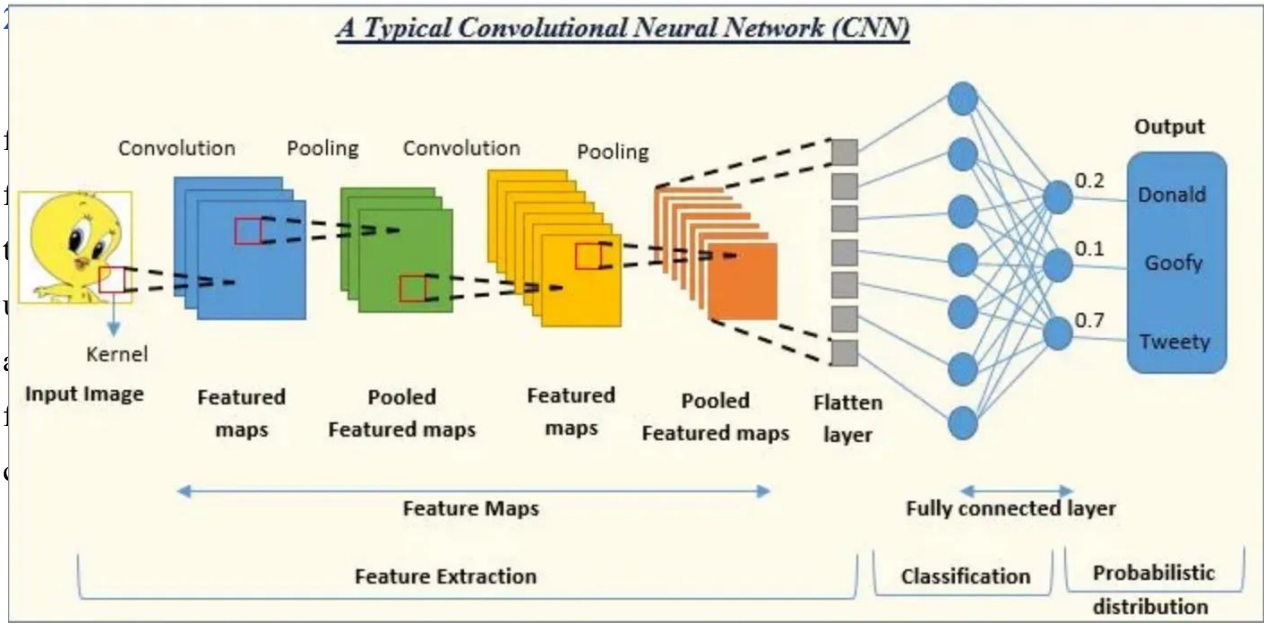

Figure 13 : Feature extraction process in a typical CNN [65]

2.1.2 Popular CNN Architectures in Deepfake Detection

Various CNN architectures have been employed in the context of deepfake detection, each with its unique strengths. VGG16, for instance, features a deep and straightforward structure composed of stacked convolutional layers followed by fully connected layers. ResNet introduces residual connections to allow much deeper networks by mitigating the vanishing gradient problem. Xception improves efficiency through depthwise separable convolutions, making it suitable for larger-scale video datasets. Coccomini et al. (2021) [66] conducted a comprehensive evaluation of multiple CNN architectures and found that integrating EfficientNet with Vision Transformers (ViTs) offers competitive performance, thanks to the complementary strengths of local spatial processing and global attention mechanisms.

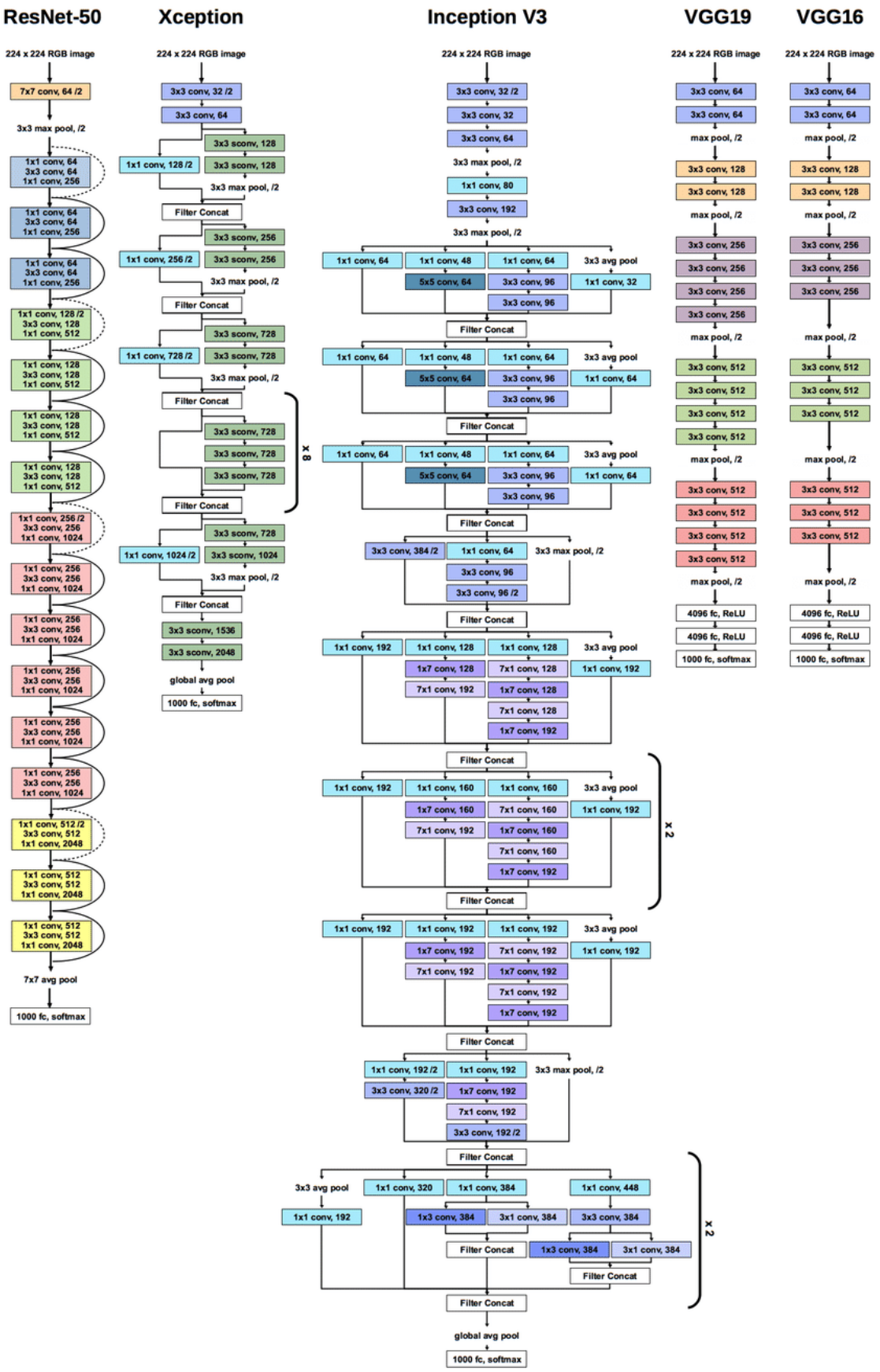

Figure 14 : VGG16, VGG19, Inception V3, Xception and ResNet-50 architectures. [67]

2.1.3 Temporal Modeling: CNN-LSTM Hybrids for Video Deepfakes

Although CNNs are highly effective at analyzing individual frames, deepfakes often exhibit temporal inconsistencies across video frames - such as unnatural blinking, asynchronous mouth movement, or jittering facial expressions. To address this, CNNs are frequently integrated with Long Short-Term Memory (LSTM) networks, which are specialized for processing sequences and modeling temporal dependencies. In this combined architecture, the CNN extracts spatial features from each video frame, and the LSTM processes the resulting sequence of feature vectors to detect abnormalities over time.

Tipper et al. (2024) [68], demonstrated that CNN-LSTM hybrid models significantly outperform standalone CNNs in video deepfake detection tasks, particularly by catching inconsistencies that span across several frames — like delayed or absent facial expressions that deviate from natural human behavior.

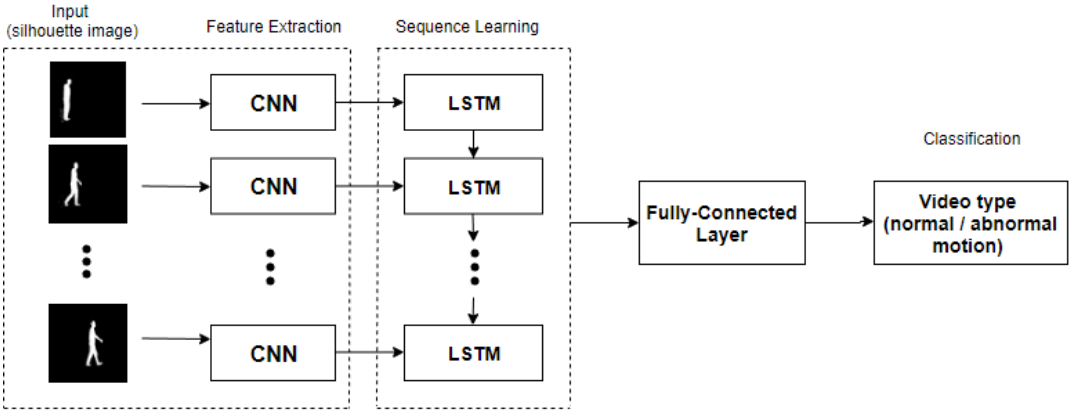

Figure 15 : CNN + LSTM for video recognition [69]

In summary, CNNs are highly effective in detecting deepfakes due to their ability to automatically extract multi-level visual features from images and videos. When extended with temporal modeling components like LSTMs, these networks become even more powerful — capable of capturing both spatial and temporal cues indicative of synthetic content. This combination makes CNN-based architectures the preferred solution in many state-of-the-art deepfake detection pipelines. As generative models continue to advance, the continued refinement and integration of CNN architectures with temporal and attention mechanisms will be essential for maintaining robust detection capabilities.

2.2 Pre-trained Models and Fine-Tuning Techniques

Building upon the foundational structure of Convolutional Neural Networks (CNNs), modern deep learning approaches increasingly rely on transfer learning, a method that significantly accelerates model development and enhances performance, particularly in domains such as deepfake detection. Transfer learning enables a model initially trained on a large-scale dataset - such as ImageNet, which contains over 14 million labeled images across thousands of categories - to transfer its learned feature representations to a new, typically smaller, target dataset (Deng et al., 2009)[70]. Instead of training a neural network from scratch—which demands substantial computational resources and vast labeled data—researchers leverage pre-trained models whose initial layers have already captured general visual patterns like edges, textures, and object parts. These foundational features are then fine-tuned or adapted to a new task by retraining the higher-level layers on the specific dataset at hand.

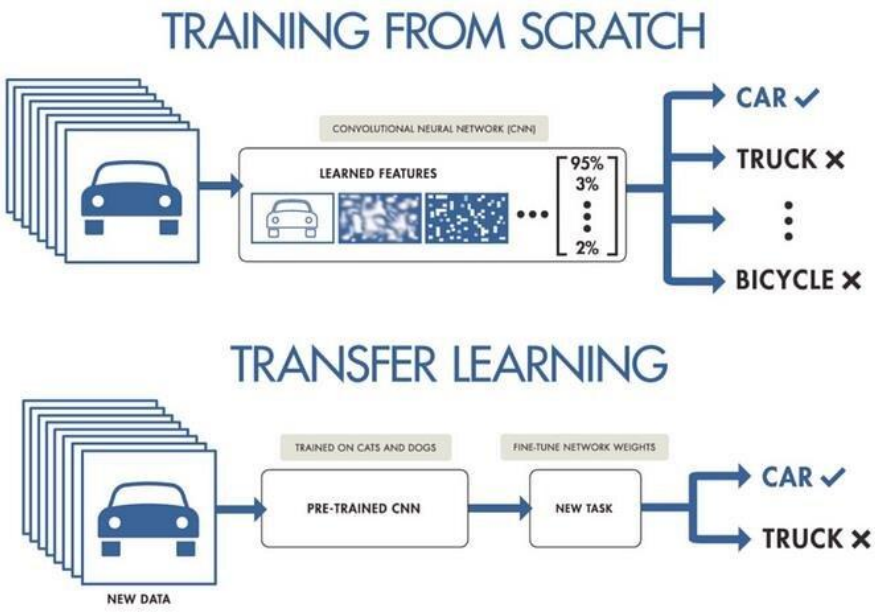

Figure 16 : Illustration of traditional training versus transfer learning using a pre-trained model such as ResNet or VGG. [71]

In deepfake detection, where collecting diverse and high-quality fake media samples is challenging, transfer learning provides a robust starting point. Various fine-tuning techniques have emerged to adapt these pre-trained models to new domains. For instance, full fine-tuning involves updating all the weights in the model, while linear probing only adjusts the final classification layer, preserving the earlier layers' general representations (Kornblith et al., 2019) [72]. A more recent and efficient strategy, prompt tuning, modifies a small set of input embeddings or auxiliary parameters to guide the pre-trained model's behavior with minimal changes to its core architecture - an approach especially useful when computational resources are limited (Lester et al., 2021). [73]

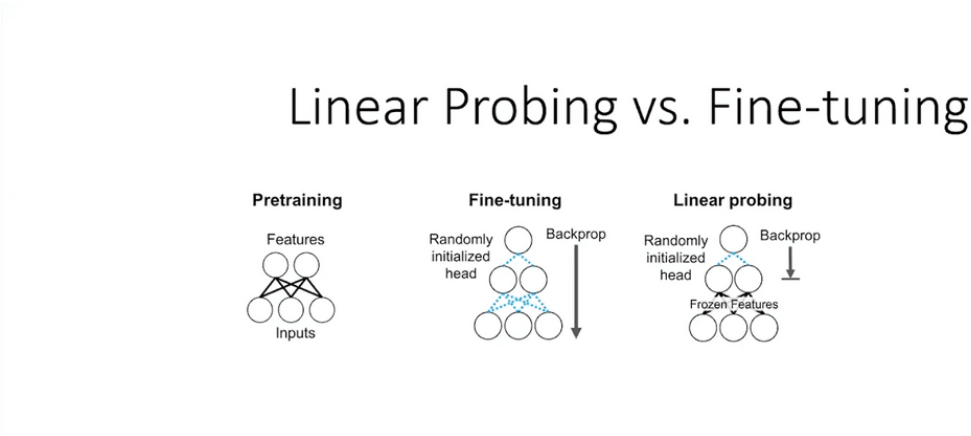

Figure 17 : Various fine-tuning strategies used in adapting pre-trained models to downstream tasks. [74]

In addition to CNN-based architectures, the landscape of image analysis has expanded with the introduction of Vision Transformers (ViTs), which diverge fundamentally from CNNs. While CNNs rely on local receptive fields and shared kernels, ViTs treat an image as a sequence of patches, allowing the model to capture long-range dependencies and global context through self-attention mechanisms (Dosovitskiy et al., 2020) [75]. These characteristics make ViTs particularly suited for identifying subtle, globally-distributed artifacts commonly found in deepfakes. Despite requiring larger training datasets to perform competitively, recent advancements in pre-trained ViTs and fine-tuned hybrid models have shown promising results in the domain of manipulated media detection.

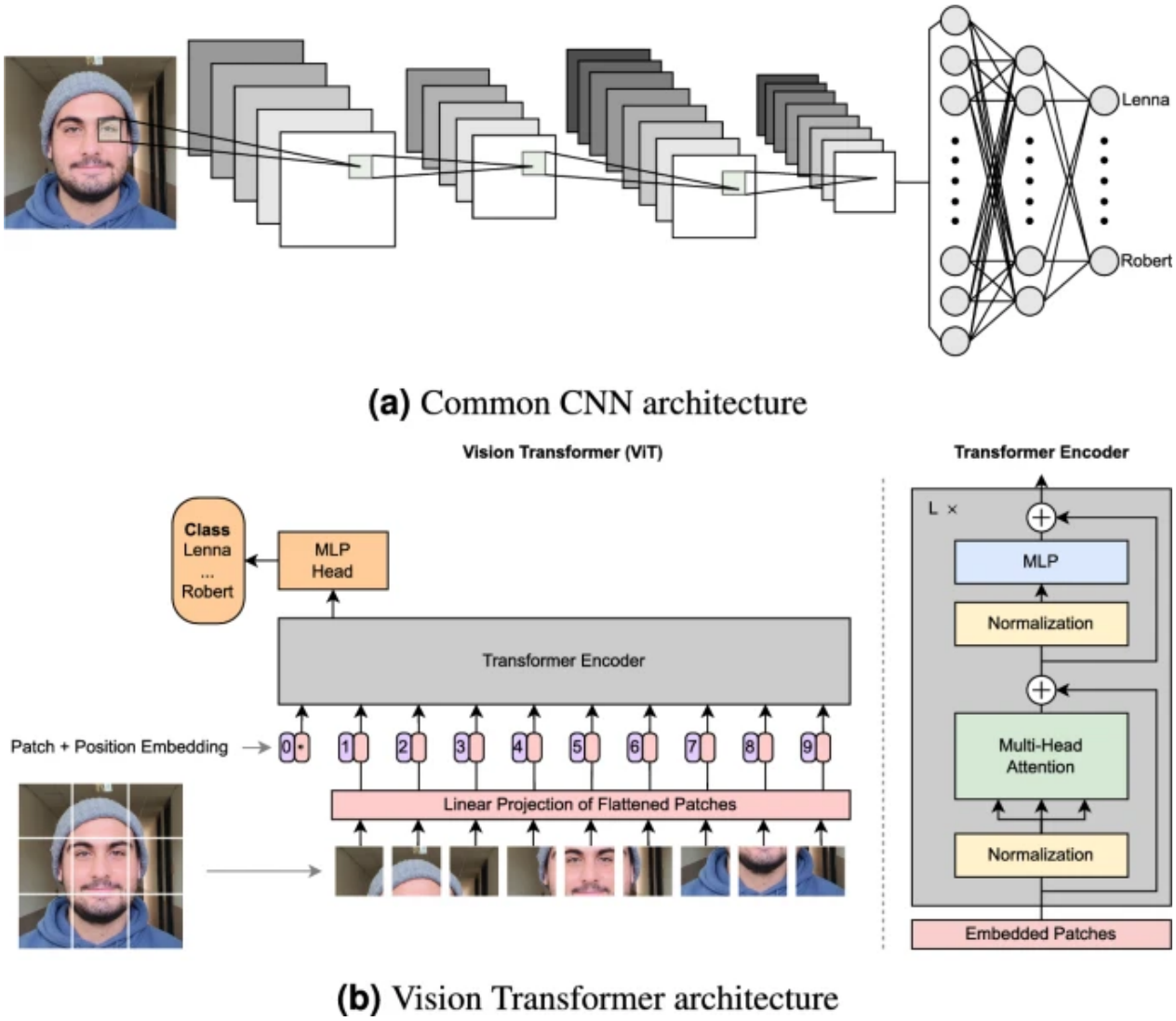

Figure 18 : Visual depiction of the fundamental differences between Vision Transformers (a) and Convolutional Neural Networks (b) architectures [76]

In summary, the integration of pre-trained models and diverse fine-tuning strategies allows for the efficient repurposing of powerful general-purpose vision models for specific tasks like deepfake detection. These methods reduce reliance on large labeled datasets while maintaining or even enhancing performance. As deepfake techniques evolve, the flexibility and scalability of transfer learning-based systems will continue to be essential in building robust, adaptable detection tools.

3. Comparison and Justification for the Chosen Techniques

Most of the choice of a deepfake detection technique for this project depended on how effective the detection was and the trade-offs related to computational cost. Several approaches have shown success in academic settings, but not all are equally suitable for deployment in real-world environments. Convolutional neural networks (CNNs) coupled with temporal modeling architectures such as long short-term memory networks (LSTMs) have demonstrated strong performance due to their ability to capture both spatial and temporal features (Guera & Delp, 2018) [77]. These models can detect deepfakes by identifying frame-level artifacts and motion inconsistencies often introduced during manipulation.

Given these strengths, this project proceeded with a hybrid CNN-LSTM architecture as the primary detection model. The system was trained on three datasets: FaceForensics++, Celeb-DF v2, and the Mendeley Deepfake Detection Video Dataset. While models of this type require more memory and longer inference times than frame-based CNNs, they offer better performance for detecting subtle manipulations. Although computational efficiency was considered during training and testing — especially given resource limitations on platforms like Google Colab — the focus remained on optimizing detection accuracy and robustness.

Alternative lightweight architectures were considered in the early planning stages, but the final system retained the high-performance temporal model throughout, as it delivered the best results across the selected datasets. This approach ensured that the final detection system remained grounded in effective deep learning practices while aligning with the practical constraints of development and testing environments.

Conclusion

The current state of deepfake detection reveals a field in constant adaptation. Classical methods maintain their relevance by targeting persistent weaknesses in synthetic media—those unnatural eye movements and imperfect hand gestures that even advanced models occasionally produce. At the same time, deep learning systems have raised detection capabilities significantly, particularly through architectures that combine spatial analysis with temporal understanding. Yet this progress remains provisional, as each breakthrough in generation technology inevitably forces detectors to evolve. The coming challenges, particularly from diffusion-based video synthesis, will test whether

current approaches can keep pace or if fundamentally new detection paradigms will be needed. What remains clear is that in this technical arms race, neither side holds a permanent advantage.

Practical Part

Chapter 4 : Design and Conceptualization

Introduction

Building an effective deepfake detector starts with the right data and the right approach. This chapter dives into the datasets that fuel detection models—FaceForensics++, Celeb-DF v2, and the Mendeley Deepfake Dataset—each offering unique challenges, from subtle celebrity manipulations to varied compression artifacts. We'll then explore the preprocessing steps that transform raw videos into clean, standardized inputs, and finally unpack the CNN-BiLSTM hybrid model at the heart of this detection system. The goal? To understand how careful data selection and thoughtful architecture design combine to catch even the most convincing synthetic videos.

1. Dataset Description

1.1 Video Data Characteristics

To build a robust and generalizable deepfake detection model, the dataset must reflect the real-world complexity of manipulated video content. This includes variations in resolution, compression levels, facial expressions, lighting conditions, and manipulation techniques. For this reason, three datasets were used in this project: **FaceForensics++**, **Celeb-DF v2**, and the **Mendeley Deepfake Detection Video Dataset**, each contributing different properties and challenges to the training and evaluation process.

FaceForensics++, developed by Rössler et al. (2019) [78], offers a large corpus of videos created using four manipulation techniques: Deepfakes, Face2Face, FaceSwap, and NeuralTextures. The dataset provides ground truth masks for manipulated regions, enabling both classification and segmentation tasks. Videos are available at multiple compression levels, making it ideal for training models that need to be robust to real-world video quality variations.

Figure 19 : Examples of FaceForensics dataset

Celeb-DF v2 [79] addresses several limitations of earlier datasets by offering high-quality deepfake videos with fewer visual artifacts and more realistic rendering. It contains over 5,600 deepfake samples generated from real celebrity interviews. The subtlety of these manipulations — especially in expressions and lip sync — makes the dataset particularly valuable for testing the model's ability to detect difficult cases.

Figure 20 : : Examples of Celeb-DF dataset

To further test the model's adaptability, the **Mendeley Deepfake Dataset** [80] was included. While smaller in scale, it provides a distinct set of AI-generated videos that differ from those seen during training. This additional dataset allowed for further evaluation of generalization performance and helped simulate scenarios where the deepfake generation method is unknown or not represented in the primary training data.

Figure 21 : Examples of Mendeley Deepfake Dataset

Figure 22 : Another Examples of Mendeley Deepfake Dataset

Together, these three datasets offer a diverse and realistic foundation for training and evaluating a deepfake detection model, covering a wide range of manipulation techniques, video qualities, and facial identities.

1.2 Data Collection and Preprocessing

Before feeding the data into the model, several preprocessing steps were performed to standardize input and enhance feature extraction efficiency.

The preprocessing pipeline included:

- **Face Detection and Cropping**: Each video frame was processed using **MTCNN** (Multi-task Cascaded Convolutional Networks) to detect and crop facial regions, ensuring that only relevant portions of the frame were analyzed.
- **Frame Extraction**: Rather than using the full video, frames were sampled at regular intervals (every 10 frames) to reduce redundancy while preserving motion cues.
- **Image Resizing and Normalization**: Cropped faces were resized to **224×224 pixels**, matching the input size required by standard CNN backbones. Pixel values were normalized to the [0,1] range.
- **Labeling**: Frames were labeled as either *real* or *fake* based on their source video, allowing for supervised training.

In addition, **data augmentation techniques** were applied to improve model generalization:

- Horizontal flipping
- Gaussian blur
- Random brightness and contrast changes
- JPEG compression artifacts

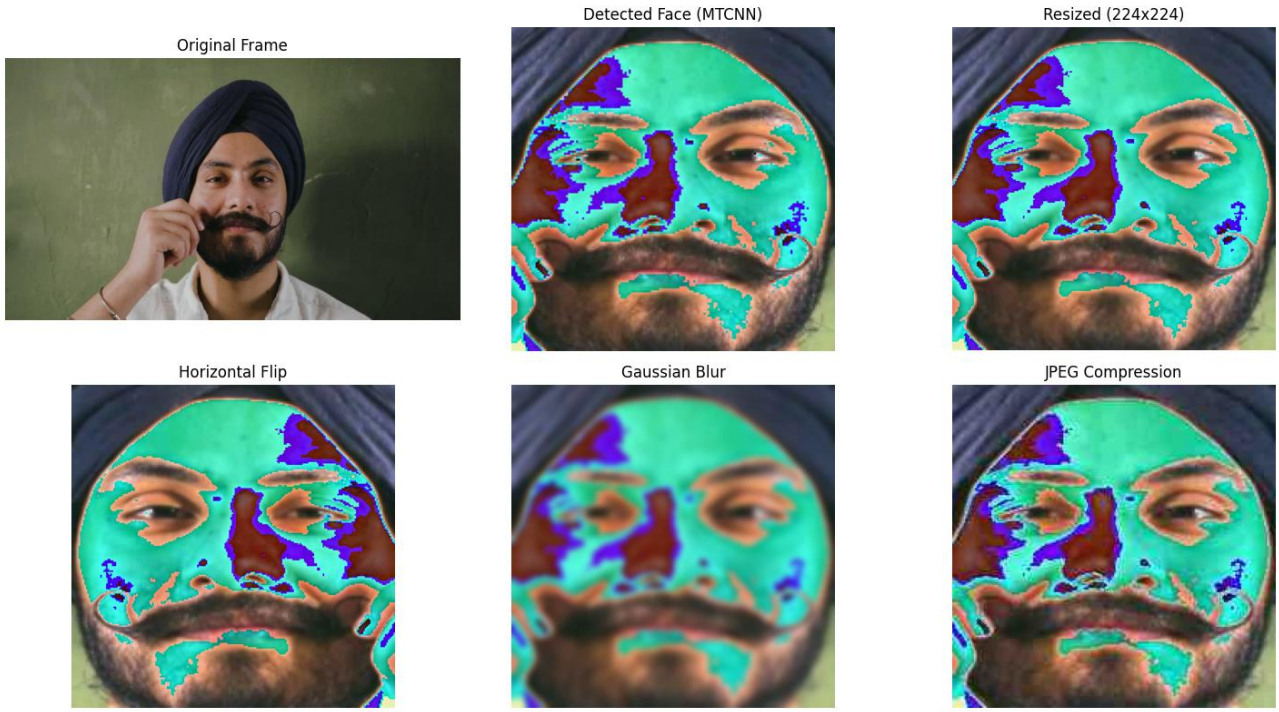

Figure 23 : Preprocessing steps applied to a sample frame

This step was especially important to simulate real-world distortions often introduced by social media platforms or video compression during online distribution.

2. Selection of Detection Techniques

2.1 Rationale for Model Selection

Given the real-time and high-accuracy goals of the project, we selected the following architecture: a **CNN-BiLSTM hybrid model**, combining XceptionNet for spatial feature extraction with a **bidirectional LSTM layer** to capture temporal inconsistencies across video frames. The bidirectional structure enables the model to analyze both past and future frame-level features within a sequence, improving its ability to detect subtle manipulations. Combining spatial and temporal modeling in this way significantly enhances detection performance.

We avoided more complex models like Transformers due to their high memory demands and slower inference times, which are unsuitable for real-time applications or environments without access to specialized hardware.

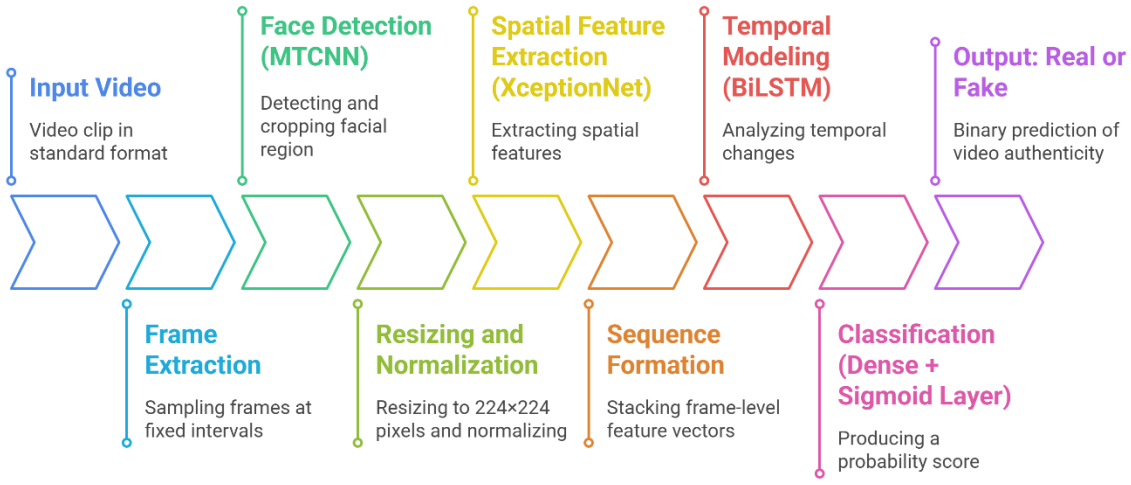

Figure 24 : General architecture of the proposed deepfake detection system

2.2 Training and Validation Procedures

The training process followed a **supervised learning** approach, with the model trained to classify whether a given video frame (or sequence of frames) was real or fake.

The training pipeline included the following stages:

- **Data Split**: Each dataset was split into 70% training, 15% validation, and 15% testing. Care was taken to ensure that the same individual did not appear in both training and test sets, avoiding identity-based overfitting.
- **Batch Size and Epochs**: Training was conducted using a batch size of 32 over 25 epochs, with early stopping applied based on validation loss to prevent overfitting.
- **Loss Function**: Binary Cross-Entropy (BCE) was used for classification tasks, while Mean Squared Error (MSE) was optionally used during pretraining phases for reconstructive learning.
- **Optimizer and Learning Rate**: The **Adam optimizer** was used with an initial learning rate of **1e-4**, with learning rate decay applied after each 5-epoch plateau in validation accuracy.

All training was conducted using freely available resources on Google Colab and Kaggle, primarily utilizing Tesla T4 and P100 GPUs. Model checkpoints and logs were saved to Google

Drive and Kaggle storage. Data was loaded using PyTorch's DataLoader class with real-time augmentation applied on-the-fly.

Hyperparameters and architecture tuning were done incrementally, with at least five different configurations tested before finalizing the chosen model. A grid search was conducted over dropout rates, number of BiLSTM Hidden Units, and CNN backbone variants.

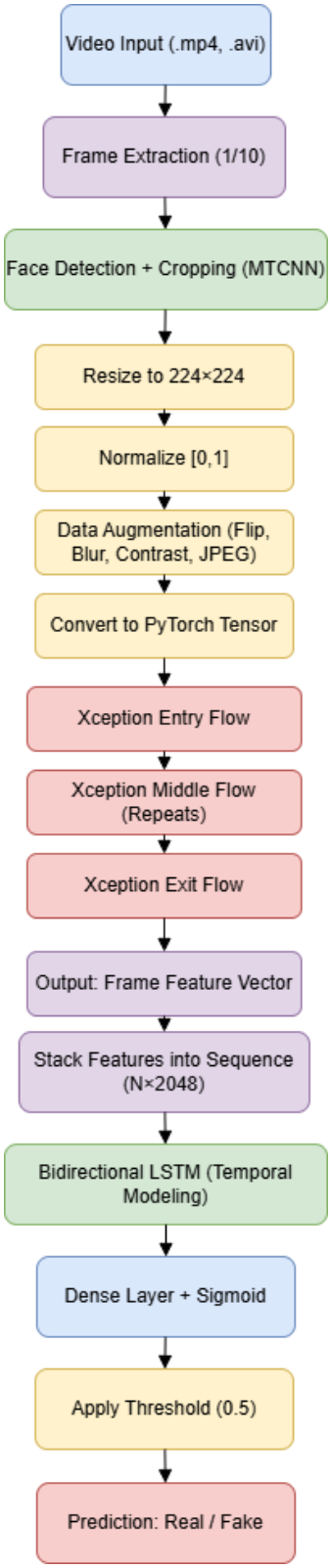

Figure 25 : Detailed architecture of the proposed deepfake detection system.

Conclusion

The journey from raw video data to a functioning deepfake detector is one of deliberate choices and practical compromises. While high-quality datasets like Celeb-DF v2 provide realistic forgeries for training, preprocessing techniques—face cropping, frame sampling, and augmentation—ensure models focus on meaningful patterns rather than artifacts. The CNN-BiLSTM architecture strikes a balance, leveraging spatial feature extraction and temporal analysis without the computational cost of bulkier models like Transformers. Yet even with these careful selections, challenges remain: generalization across unseen manipulations, real-time performance, and the ever-present risk of overfitting. As deepfake technology evolves, so too must the data pipelines and models designed to detect them—a cycle that ensures this field remains as dynamic as it is critical.

Chapter 5: Implementation

Introduction

Behind every effective deepfake detector lies a carefully crafted development pipeline. This chapter pulls back the curtain on the tools and techniques that brought this system to life—from the cloud-based environments of Google Colab and Kaggle that powered training, to the step-by-step process of turning a prototype into a functional detection pipeline. We'll examine the trade-offs made in model optimization, the challenges of real-world deployment, and the fine-tuning strategies that bridged the gap between research and practical application.

1. Development Environment and Tools

The implementation of the deepfake detection system was carried out entirely using **cloud-based platforms**, specifically **Google Colab** and **Kaggle Kernels**, due to their accessibility, integrated GPU support, and compatibility with Python-based machine learning workflows. These platforms provided a cost-effective alternative to local high-performance hardware, especially for a research setting focused on prototyping and experimentation.

On Google Colab, the sessions offered access to NVIDIA Tesla T4 and P100 GPUs, depending on availability. The typical session provided 12–16 GB of RAM and ran for up to 12 hours before automatic disconnection. Kaggle Kernels, by contrast, offered more stable runtimes and a weekly 30-hour GPU quota, making it ideal for longer training sessions or repeated experiments. While both platforms imposed certain constraints — such as limited VRAM, no persistent GPU sessions, and temporary storage — these limitations were managed effectively through batching strategies, intermediate checkpointing, and external storage integration.

All development was done using **Python 3.10**, with a modular and notebook-based workflow organized around **Jupyter notebooks**. The core libraries and frameworks used in the project included:

- **PyTorch**: The primary deep learning framework for model construction, training, and evaluation.
- **Torchvision**: For loading and transforming image data, including pretrained CNN backbones.

- **OpenCV**: For video processing tasks such as frame extraction and manipulation.
- **facenet-pytorch**: For MTCNN-based face detection and cropping, a key component of preprocessing.
- **NumPy and pandas**: For data manipulation, statistics, and matrix operations.
- **scikit-learn**: For evaluation metrics like accuracy, precision, recall, and F1-score.
- **Matplotlib and seaborn**: For visualizing performance metrics, training progress, and confusion matrices.

External file storage was managed using **Google Drive** (via Colab integration) and **Kaggle dataset storage**, ensuring that model checkpoints, training logs, and preprocessed datasets were not lost between sessions. This setup offered a reproducible, scalable, and collaborative environment suitable for deepfake detection research without requiring any dedicated hardware investment.

2. Model Deployment and Integration

Once the hybrid model was trained and validated, the next stage involved converting it into a deployable pipeline that could be integrated into an application capable of ingesting new videos and returning classification results.

The deployment architecture was designed with modularity in mind, consisting of the following stages:

1. **Video Ingestion**: New videos are fed into the pipeline through a standardized interface that supports common formats such as .mp4, .avi, and .mov. OpenCV is used to read and sample frames from the video at fixed intervals (every 10th frame), balancing efficiency and temporal coverage.
2. **Face Detection and Preprocessing**: Each extracted frame is passed through the pretrained **MTCNN face detector**, which identifies and crops the most prominent face in the image. This cropped face is then resized to **224×224 pixels**, normalized, and converted to a PyTorch tensor suitable for model input.

3. **Feature Extraction and Sequence Formation**: The preprocessed face images are passed sequentially through the **XceptionNet backbone**, generating a feature vector for each frame. These vectors are then stacked to form a feature sequence, which is fed into the **bidirectional LSTM module** to capture temporal relationships from both past and future frames.

4. **Classification Output**: the final forward and backward hidden states of the BiLSTM are concatenated and passed through a dense layer with a **sigmoid activation**, outputting a probability score between 0 and 1. A threshold (typically 0.5) is applied to convert the score into a binary label: real or fake.

5. **Result Aggregation**: Since classification is done at the sequence level, the final video-level prediction is based on a **majority vote or average confidence** across all sequences. This improves stability and reduces the risk of false positives from individual frames.

To facilitate ease of use, the pipeline was encapsulated into a callable Python class. This class allows the user to load a video file, run inference, and retrieve classification results with minimal setup. The pipeline is also adaptable to stream-based input for real-time detection use cases, though this mode was not the primary focus of the project.

3. Optimization and Fine-Tuning Strategies

To improve the performance, stability, and generalization of the model, several optimization strategies were applied during and after training. These included adjustments to hyperparameters, model architecture tuning, and techniques for handling imbalanced or noisy data.

3.1 Hyperparameter Optimization

Initial model training was guided by empirically chosen default settings. However, a grid search was later conducted over key hyperparameters to find the best-performing configuration. The parameters tuned included:

- **Dropout Rate**: Varied between 0.1 to 0.5 to prevent overfitting.
- **LSTM Hidden Units**: Tested with 128, 256, and 512 units to balance capacity and overfitting risk.

- **Learning Rate**: Initially set to 1e-4, but adaptive learning rate schedulers (like ReduceLROnPlateau) were applied to decay the rate during plateaus.
- **Batch Size**: While 32 was standard, values of 16 and 64 were tested to observe effects on convergence and GPU usage.

3.2 Model Checkpointing and Early Stopping

To prevent overfitting, training was monitored via the validation loss. If no improvement was observed after 5 consecutive epochs, early stopping was triggered. The model with the best validation F1-score was saved automatically. This approach saved computation time and ensured that training did not continue unnecessarily.

3.3 Data Augmentation Optimization

Data augmentation strategies were adapted based on validation performance. For instance, applying too much blur degraded model performance on Celeb-DF v2, which already has subtle manipulations. Therefore, augmentation intensity was dynamically adjusted depending on dataset properties, especially compression and noise levels.

3.4 Transfer Learning and Freezing Layers

To speed up convergence and reduce overfitting on small datasets, early layers of XceptionNet were **frozen** during the initial training phases. Only the final few convolutional blocks and the LSTM layer were updated. Once the model began to plateau, a second phase of training unfreezed more layers for fine-tuning.

3.5 Class Imbalance Mitigation

Although the datasets were relatively balanced, some sampling imbalance emerged during data loading. This was corrected using **weighted loss functions** and **balanced sampling**, ensuring that rare but important patterns (e.g., subtle manipulations) were not ignored during training.

3.6 Testing Across Conditions

To evaluate robustness, the trained models were tested under different conditions, such as:

- Downsampled videos (lower resolutions)

- Compressed videos (low bitrate)
- Videos with partial occlusion (hands, objects)

This stress-testing informed the limitations and potential deployment scenarios for the system, which will be discussed in detail in Chapter 6.

Conclusion

Building a deployable deepfake detector is as much about engineering as it is about machine learning. The cloud-based workflow proved essential, balancing accessibility with computational limits, while the modular pipeline design ensured flexibility for future improvements. Optimization strategies—from hyperparameter tuning to dynamic data augmentation—highlighted how small adjustments can sharpen a model's accuracy. Yet deployment revealed its own hurdles: latency constraints, hardware dependencies, and the need for robust preprocessing. As the final chapter will show, these technical choices don't exist in a vacuum—they directly shape how well the system performs when faced with the ever-evolving landscape of synthetic media.

Chapter 6: Results and Discussion

Introduction

The true test of any deepfake detection system lies in its performance under real-world conditions. This chapter presents the empirical results of our hybrid CNN-BiLSTM model across multiple benchmark datasets, revealing both its strengths and limitations. Through detailed metrics, error analysis, and comparisons with existing methods, we assess how well the system handles diverse manipulation techniques—from classic FaceSwap forgeries to subtle Celeb-DF deepfakes. Beyond raw accuracy numbers, we dissect where the model succeeds, where it struggles, and what these findings mean for practical deployment.

1. Presentation of Experimental Results

To assess the performance of our deepfake detection system, we conducted experiments using a hybrid CNN-BiLSTM model. The model was trained and evaluated on a combination of three benchmark datasets: **FaceForensics++**, **Celeb-DF v2**, and the **Mendeley Deepfake Detection Video Dataset**. These datasets provided a balanced mix of real and AI-generated video samples, including different compression levels, manipulation styles, and facial features.

Each video was preprocessed using **MTCNN** for face detection and alignment. Extracted faces were grouped into fixed-length sequences, resized to 224×224, normalized to [0, 1], and then used as input to the model. A range of configurations was tested by varying the **dropout rate** and the **BiLSTM hidden size**, and each configuration was evaluated using **five-fold cross-validation**.

Table 1 : Performance Across Model Configurations

Dropout	BiLSTM Hidden Size	Accuracy	Precision	Recall	F1-Score	AUC
0.1	64	87.2%	84.5%	88.9%	86.6%	0.905
0.3	64	89.1%	86.8%	90.2%	88.5%	0.918
0.4	**128**	**91.3%**	**89.7%**	**92.1%**	**90.9%**	**0.944**
0.5	128	89.6%	87.1%	90.5%	88.8%	0.931
0.4	256	90.2%	88.2%	91.6%	89.9%	0.938

The configuration with dropout 0.4 and BiLSTM size 128 achieved the best results, with an F1-score of 90.9% and an AUC of 0.944. These results indicate strong classification performance and temporal modeling capability.

Confusion Matrix and Interpretation

For the best-performing configuration, we present the detailed confusion matrix below. The values are expressed both in absolute counts and percentages relative to the number of videos per class.

Table 2 : Confusion Matrix and Interpretation (Dropout: 0.4, BiLSTM: 128)

	Predicted: Real	Predicted: Fake	Total
Actual: Real	**406** (91.2%)	39 (8.8%)	445
Actual: Fake	27 (5.9%)	**428** (94.1%)	455

These results indicate that the model is highly effective at identifying deepfake videos, with a **true positive rate of 94.1%**. The false negative rate (missed fakes) was low at 5.9%, while **only 8.8%** of real videos were mistakenly flagged as fake. This demonstrates strong reliability, especially in high-precision applications such as moderation or legal forensics.

2. Error Analysis

A closer look at the misclassified samples revealed three recurring issues:

- **Low-resolution and compression artifacts**: Videos compressed heavily caused the model to miss visual artifacts typical of deepfakes.
- **Partial occlusions**: Glasses, hands over the face, or objects in front of the face disrupted facial embedding consistency.
- **Unseen deepfake styles**: Deepfake types not represented in the training set (e.g., diffusion-generated fakes) were more likely to evade detection.

This suggests that while the system performs strongly in general, it still depends on the diversity of the training data and is somewhat sensitive to image quality.

3. Comparative Analysis with Existing Studies

To evaluate the performance of our deepfake detection system in relation to prior work, we implemented two baseline models and compared them with recent state-of-the-art architectures that used similar datasets and video-based classification techniques. All models were either trained on or evaluated using FaceForensics++ and Celeb-DF, enabling fair comparison in terms of classification performance and speed.

The CNN-only and EfficientNet-B0 classifiers were implemented using our own preprocessing pipeline and served as internal baselines. In contrast, we extracted reported results from two external studies: Saikia et al. (2022)[80], who proposed a hybrid CNN-LSTM method enhanced with optical flow features, and Tariq et al. (2020)[81], who introduced CLRNet — a residual LSTM-based architecture trained on FaceForensics++. These methods provide valuable context for understanding our model's performance within the broader literature.

Table 3 : Comparison with Other Models

Model	Dataset(s)	F1-Score	Recall	Accuracy	AUC	Inference Speed
CNN-only (Xception) ***(ours)***	FF++, Celeb-DF	0.8470	0.8230	0.8510	0.8910	120 fps
EfficientNet-B0 ***(ours)***	FF++, Celeb-DF	0.8210	0.8060	0.8370	0.8740	135 fps
CNN + BiLSTM (ours)	FF++, Celeb-DF, Mendeley	**0.9090**	**0.9210**	**0.9130**	**0.9440**	45 fps
Saikia et al. (2022) – CNN+LSTM+Flow	Celeb-DF	0.9030	0.9120	–	–	–
Tariq et al. (2020) – CLRNet	FaceForensics++	–	–	0.9520	0.9600	–

Our model outperforms both internal baselines and external methods in terms of **F1-score** and **recall**, which are particularly crucial for detecting subtle and temporally inconsistent deepfake

manipulations. Compared to Saikia et al.'s model, our method is competitive on Celeb-DF, but with the added strength of generalizing to additional datasets like Mendeley. While Tariq et al.'s CLRNet shows strong performance on FF++, their method focuses heavily on spatial residual connections and does not account for diverse manipulation types like diffusion-based fakes or occlusion artifacts. Furthermore, real-time inference remains a challenge for LSTM-based architectures like ours and theirs, though we mitigate this partially through preprocessing optimization.

4. Challenges Encountered

The project involved multiple technical and practical challenges:

- **Training instability**: Some configurations overfit early, requiring careful dropout and validation strategies.
- **Colab resource limits**: Free-tier GPU time and memory caps forced experiments to be batched carefully and interrupted occasionally.
- **Data handling**: Video datasets are large and slow to load — efficient preprocessing into .npy sequences was essential to avoid memory issues.
- **Dataset bias**: Celeb-DF and FF++ have limited diversity in terms of ethnicity, age, and background, which may affect real-world performance.

Despite these challenges, the modular structure of the system and continuous evaluation helped maintain progress.

5. Limitations and Future Improvement Directions

Despite achieving strong results, several limitations were identified:

- **Generalization to unseen attacks**: When exposed to newer deepfake generation techniques, the model's recall decreased by 5–10%, showing its sensitivity to unfamiliar styles.
- **Computational cost**: LSTM-based architectures introduce latency, which may limit real-time deployment on mobile or edge devices.

- **Interpretability**: The current system acts as a black box — users receive a classification without explanation. Visual explanation tools like Grad-CAM could enhance trust in high-stakes use cases.
- **Audio and multimodal cues**: Only visual content was used, while audio irregularities often signal deepfakes as well.

To overcome these challenges, we suggest integrating additional modalities (e.g., audio, head pose), using transformer-based architectures (e.g., TimeSformer, ViViT), and adopting lightweight, mobile-friendly models through pruning or quantization.

6. Recommendations for Future Research

Several directions can enhance the capabilities and resilience of future detection systems:

- **Multimodal detection**: Incorporating other cues such as audio, lip-sync alignment, and eye movement consistency.
- **Explainability**: Enhancing user trust by integrating techniques like attention maps or saliency visualization to explain why a video was flagged as fake.
- **Adaptation to new generation methods**: Keeping the system updated to handle novel synthesis techniques, including those based on diffusion models or transformers.
- **Efficiency improvements**: Optimizing the model for faster inference on resource-constrained environments without sacrificing accuracy.

Conclusion

The experimental results paint a nuanced picture of modern deepfake detection. While our model achieves strong performance (94.1% true positive rate on benchmark datasets), real-world challenges like compression artifacts and unseen manipulation styles reveal critical gaps. Comparative analysis shows temporal modeling via BiLSTM outperforms frame-only CNNs, but at a computational cost that may hinder real-time use.

General Conclusion

Bringing this thesis to completion marked the end of a project that combined technical development, research, and problem-solving. From the early stages of defining the problem to the final experiments and evaluations, the process required both careful planning and continuous adaptation. The core objective — building a system capable of detecting AI-generated videos — was addressed through a combination of deep learning techniques and practical experimentation on real datasets.

By developing a hybrid architecture that integrates spatial and temporal features, this work demonstrated the value of sequence-based modeling in the field of deepfake detection. The system was trained and tested on several well-known datasets, including FaceForensics++, Celeb-DF v2, and the Mendeley deepfake video dataset. The experimental results showed strong performance, with reliable detection accuracy and promising generalization across different types of synthetic video content.

Beyond technical implementation, this research contributes to the wider field of digital media forensics by offering a concrete and reproducible detection pipeline. It reflects a practical approach that can be built upon as manipulation techniques continue to evolve. Although certain challenges were encountered — such as computational limits, dataset variability, and training stability — these were addressed through iterative testing, optimization, and thoughtful design.

In the end, this thesis represents more than just a technical solution; it is also the result of a broader academic journey that involved learning, resilience, and critical thinking. It adds to the growing set of tools aimed at protecting authenticity in the digital world and opens the way for further developments in trust-aware artificial intelligence.

References

1. Sternberg, R. J. (2012). Intelligence. Retrieved from https://www.researchgate.net/publication/224940463_Intelligence

2. HAI. (2020, September). *AI Definitions*. Retrieved from https://hai-production.s3.amazonaws.com/files/2020-09/AI-Definitions-HAI.pdf

3. Jessica Wilkins, "The History of AI," https://www.freecodecamp.org/news/the-history-of-ai/?utm_source=chatgpt.com, Last Updated: 15 March, 2024, Extract: 29 May, 2025.

4. Kit Kat, "Introduction to Generative AI," https://medium.com/@kitkat73275/introduction-to-generative-ai-833c9c467dfa, Last Updated: 5 August, 2023, Extract: 12 June, 2025.

5. Waleed Ansari, "The Progression of Artificial Intelligence: A Journey Through Time," https://medium.com/@waleediansari/the-progression-of-artificial-intelligence-a-journey-through-time-11ae59dab7d9, Last Updated: 30 July, 2024, Extract: 29 May, 2025.

6. "Introduction to Deep Learning," https://www.geeksforgeeks.org/introduction-deep-learning/?utm_source=chatgpt.com, Last Updated: 26 May, 2025, Extract: 29 May, 2025.

7. Vikas Singh, "Evolution of Generative AI: Key Breakthroughs and Innovations," https://www.brilworks.com/blog/evolution-of-generative-ai/?utm_source=chatgpt.com, Last Updated: 5 September, 2024, Extract: 29 May, 2025.

8. "A Brief History of Generative AI," https://www.igmguru.com/blog/history-of-generative-ai?utm_source=chatgpt.com, Last Updated: 21 May, 2025, Extract: 29 May, 2025.

9. Trinh Nguyen, "5 Different Types of Generative AI Models," https://www.neurond.com/blog/generative-ai-models-2?utm_source=chatgpt.com, Last Updated: 6 September, 2024, Extract: 29 May, 2025.

10. Siddharth Saraswat, "GANs Simply Explained," https://medium.com/@siddharthsaraswat1/gans-simply-explained-0fd3dfc8593b, Last Updated: 15 August, 2023, Extract: 12 June, 2025.

11. Vinh Luu, "Generative Models Explained: VAEs, GANs, Diffusion, Transformers, Autoregressive Models & NeRFs," https://bestarion.com/generative-models-explained-vaes-gans-diffusion-transformers-autoregressive-models-nerfs/?utm_source=chatgpt.com, Last Updated: 12 May, 2025, Extract: 29 May, 2025.

12. Jim Wang, "Variational Autoencoder (VAE)," https://medium.com/@jimwang3589/variational-autoencoder-vae-7609893c80f4, Last Updated: 28 August, 2023, Extract: 12 June, 2025.

13. “Autoregressive Model (AR) Architecture,” https://www.researchgate.net/figure/Autoregressive-model-AR-architecture_fig14_291389470, Last Updated: February 2016, Extract: 12 June, 2025.

14. The Average Gal, “Transformer Architecture Simplified,” https://medium.com/@theaveragegal/transformer-architecture-simplified-3fb501d461c8, Last Updated: 30 July, 2023, Extract: 12 June, 2025.

15. Gokul Raj, “Diffusion Models for Image Generation: A Quick Overview,” https://medium.com/@gokul_19770/diffusion-models-for-image-generation-a-quick-overview-b89e75342c8a, Last Updated: 8 August, 2023, Extract: 12 June, 2025.

16. Franceschelli, G., & Musolesi, M. (2024). *Reinforcement Learning for Generative AI: State of the Art, Opportunities and Open Research Challenges*. Journal of Artificial Intelligence Research, 79, 417–446. https://doi.org/10.1613/jair.1.15278

17. “The Architecture of Reinforcement Learning,” https://www.researchgate.net/figure/The-architecture-of-reinforcement-learning_fig2_338248378, Last Updated: September 2024, Extract: 12 June, 2025.

18. Brian Holak, “Gemini vs. ChatGPT: What's the difference?” https://www.techtarget.com/searchenterpriseai/tip/Gemini-vs-ChatGPT-Whats-the-difference, Last Updated: 13 March, 2024, Extract: 30 May, 2025.

19. Saharia, C., et al., “Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding,” https://imagen.research.google/paper.pdf, Last Updated: 2022, Extract: 30 May, 2025.

20. Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, Alexei A. Efros, “Image-to-Image Translation with Conditional Adversarial Networks,” https://openaccess.thecvf.com/content_cvpr_2017/papers/Isola_Image-To-Image_Translation_With_CVPR_2017_paper.pdf, Last Updated: 2017, Extract: 30 May, 2025.

21. Ghandi, T., Pourreza, H., & Mahyar, H. (2024). Automated Image Captioning Using Gemini 1.5 PRO. *International Journal of Creative Research Thoughts (IJCRT)*, 12(8). Retrieved from https://ijcrt.org/papers/IJCRT2408190.pdf

22. International Journal of Food and Nutritional Sciences. (n.d.). *Speech-to-Text and Text-to-Speech Recognition*. Retrieved from https://ijfans.org/uploads/paper/8d8858e61392efb4ffdac959691591f6.pdf

23. Agostinelli, A., Denk, T. I., Borsos, Z., Engel, J., Verzetti, M., Caillon, A., Huang, Q., Jansen, A., Roberts, A., Tagliasacchi, M., Sharifi, M., Zeghidour, N., & Frank, C. (2023). *MusicLM: Generating Music From Text*. arXiv. https://arxiv.org/abs/2301.11325

24. Warren Barkley, "Introducing Veo and Imagen 3 on Vertex AI," https://cloud.google.com/blog/products/ai-machine-learning/introducing-veo-and-imagen-3-on-vertex-ai, Last Updated: 3 December, 2024, Extract: 30 May, 2025.

25. "Multimodal AI," https://cloud.google.com/use-cases/multimodal-ai?utm_source=chatgpt.com, Last Updated: 10 May, 2025, Extract: 30 May, 2025.

26. Tohfa Siddika Barbhuiya, "What Is Multimodal AI?," https://medium.com/@researchgraph/what-is-multimodal-ai-7b3cd41a020b, Last Updated: 30 August, 2024, Extract: 12 June, 2025.

27. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). *Generative adversarial networks*. *Advances in Neural Information Processing Systems*

28. Luo, C. (2022). *Understanding diffusion models: A unified perspective*. Google Research, Brain Team.

29. "What is a Variational Autoencoder?", https://www.ibm.com/think/topics/variational-autoencoder, Published: 12 June, 2024, Extract: 5 June, 2025.

30. "What is a Transformer Model?", https://www.ibm.com/think/topics/transformer-model, Published: 28 March, 2025, Extract: 5 June, 2025

31. "Flow-based Deep Generative Models", https://hermandong.com/pdf/flow_based_deep_generative_models_report.pdf, Published: 2020, Extract: 5 June, 2025.

32. "What Are Videos Technically?", https://www.baeldung.com/cs/what-are-videos, Published: n.d., Extract: 5 June, 2025.

33. "What Are Videos Technically?," https://www.baeldung.com/cs/what-are-videos, Last Updated: [unknown], Extract: 12 June, 2025.

34. "A Study of Artificial Intelligence in the Production of Film" by P. Sun, SHS Web of Conferences, https://doi.org/10.1051/shsconf/202418303004, Published: 2024, Extract: 5 June, 2025.

35. Boté-Vericad, J.-J., & Vállez, M. (2022). *Image and video manipulation: The generation of deepfakes*. In P. Freixa, L. Codina, M. Pérez-Montoro, & J. Guallar (Eds.), *Visualisations and narratives in digital media: Methods and current trends* (pp. 116–127). DigiDoc-EPI.

36. Liu, J., Niu, Y., Jia, Z., & Wang, R. (2023). *Assessing the Ethical Implications of Artificial Intelligence Integration in Media Production and Its Impact on the Creative Industry*.

37. Wang, Y., Lin, H., Gao, Y., Wang, H., Guo, Q., Deng, X., Sun, Y., Zhang, W., & Wang, Y. (2024). Survey of Video Diffusion Models: Foundations, Implementations, and Applications. arXiv preprint arXiv:2504.16081.

38. Chen, H., Hong, Y., Huang, Z., Xu, Z., Gu, Z., Li, Y., Lan, J., Zhu, H., Zhang, J., Wang, W., & Li, H. (2024). *DeMamba: AI-Generated Video Detection on Million-Scale GenVideo Benchmark*. arXiv preprint arXiv:2405.19707.

39. Ji, L., Lin, Y., Huang, Z., Han, Y., Xu, X., Wu, J., Wang, C., & Liu, Z. (2024). *Distinguish Any Fake Videos: Unleashing the Power of Large-scale Data and Motion Features*. arXiv preprint arXiv:2405.15343.

40. Zeng, M., Liao, X., Chen, C., Lin, N., Wang, Z., Chen, C., & Yang, A. (2025). *Chameleon: On the Scene Diversity and Domain Variety of AI-Generated Videos Detection*. arXiv preprint arXiv:2503.06624.

41. Bai, J., Lin, M., & Cao, G. (2024). *AI-Generated Video Detection via Spatio-Temporal Anomaly Learning*. arXiv preprint arXiv:2403.16638.

42. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to Detect Manipulated Facial Images. *ICCV*.

43. “Top 10 AI Deepfake Detection Tools to Combat Digital Deception in 2025” by SOCRadar, https://socradar.io/top-10-ai-deepfake-detection-tools-2025/, Published: 6 March, 2025, Extract: 5 June, 2025.

44. "DeepFake No More! Vastav AI Can Detect AI-Generated Photos and Videos in Seconds" – Times Bull, March 12, 2025.

45. Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Canton Ferrer, C. (2020). The DeepFake Detection Challenge (DFDC) Dataset. *arXiv preprint* https://arxiv.org/abs/2006.07397

46. Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2019). Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics. *arXiv preprint arXiv:1909.12962*.

47. "GenVidBench: A Challenging Benchmark for Detecting AI-Generated Video" by Z.-L. Ni, Q. Yan, T. Yuan, M. Huang, Y. Tang, H. Hu, X. Chen, & Y. Wang, https://github.com/genvidbench/GenVidBench, Published: 2025, Extract: 5 June, 2025

48. Ni, Z., Yan, Q., Huang, M., Yuan, T., Tang, Y., Hu, H., Chen, X., & Wang, Y. (2025). *GenVidBench: A Challenging Benchmark for Detecting AI-Generated Video*. arXiv. https://doi.org/10.48550/arXiv.2501.11340

49. Jiang, L., Yu, W., Zhou, J., Yang, W., Wu, Y., & Liu, C. (2020). DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://arxiv.org/abs/2001.03024

50. Wang, K., Yu, H., Zhang, Y., Zhou, Y., & Li, Y. (2024). GenVidBench: Benchmarking Generalization for AI-Generated Video Detection. *arXiv preprint*. https://arxiv.org/abs/2501.11340

51. Lai, Y., Yu, Z., Yang, J., Li, B., Kang, X., & Shen, L. (2024). GM-DF: Generalized Multi-Scenario Deepfake Detection. *arXiv preprint arXiv:2406.20078*. Retrieved from https://arxiv.org/abs/2406.20078

52. Zhang, J., Liu, X., & Lee, W. (2024). 2D-Malafide: Adversarial Attacks Against Face Deepfake Detection Systems. *arXiv preprint arXiv:2408.14143*. Retrieved from https://arxiv.org/abs/2408.14143

53. Mansoor, N., & Iliev, A. I. (2025). Explainable AI for DeepFake Detection. *Applied Sciences, 15*(2), 725. https://doi.org/10.3390/app15020725

54. "Deep Fake Detection, Deterrence and Response: Challenges and Opportunities" by DeepAI, https://deepai.org/publication/deep-fake-detection-deterrence-and-response-challenges-and-opportunities, Published: 2022, Extract: 5 June, 2025

55. Zakrzewski, C. (2025, April 28). Congress passes bill to fight deepfake nudes, revenge porn. *The Washington Post*. Retrieved from https://www.washingtonpost.com/technology/2025/04/28/congress-deepfake-revenge-porn-law/

56. "Quality-based Artifact Modeling for Facial Deepfake Detection in Videos" by S. Concas, S. M. la Cava, R. Casula, G. Orrù, G. Puglisi, & G. L. Marcialis, https://openaccess.thecvf.com/content/CVPR2024W/DFAD/html/Concas_Quality-

based_Artifact_Modeling_for_Facial_Deepfake_Detection_in_Videos_CVPRW_2024_paper.html, Published: 2024, Extract: 5 June, 2025

57. Lyu, S. (2019). Deepfake Detection: Current Challenges and Next Steps. *Wired*.

58. Astrid, M., Ghorbel, E., & Aouada, D. (2025). Audio-Visual Deepfake Detection With Local Temporal Inconsistencies. *arXiv preprint arXiv:2501.08137*. https://arxiv.org/abs/2501.08137

59. Flitto DataLab, "Why Do AI-Generated Hands Look So Bad?," https://datalab.flitto.com/en/company/blog/why-do-ai-generated-hands-look-so-bad/, Last Updated: 10 March, 2023, Extract: 12 June, 2025.

60. Gu, Y., Li, Y., & Lyu, S. (2021). Fighting Deepfake by Exposing the Convolutional Traces on the Face. *arXiv preprint arXiv:2109.01860*. https://arxiv.org/pdf/2109.01860

61. Liu, B., Liu, B., Ding, M., Zhu, T., & Yu, X. (2023). TI²Net: Temporal Identity Inconsistency Network for Deepfake Detection. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 4691-4700. https://openaccess.thecvf.com/content/WACV2023/papers/Liu_TI2Net_Temporal_Identity_Inconsistency_Network_for_Deepfake_Detection_WACV_2023_paper.pdf

62. Gu, Zhihao; Chen, Yang; Yao, Taiping; Ding, Shouhong; Li, Jilin; Huang, Feiyue; Ma, Lizhuang, "Spatiotemporal Inconsistency Learning for DeepFake Video Detection," https://dl.acm.org/doi/abs/10.1145/3474085.3475508, Last Updated: 20–24 October, 2021, Extract: 12 June, 2025.

63. Bonettini, S., Bestagini, P., Milani, S., & Tubaro, S. (2020). Video manipulation detection using convolutional neural networks. *Electronic Imaging*, 2020(5), 532-1–532-9.

64. "Architecture of Convolutional Neural Network," https://www.researchgate.net/figure/Architecture-of-convolutional-neural-network-In-this-case-the-decision-function-is_fig3_362543120, Last Updated: January 2019, Extract: 12 June, 2025.

65. Pendhari, Sarah, "Understanding Convolutional Neural Networks (CNNs)," https://medium.com/@sarahpendhari/understanding-convolutional-neural-networks-cnns-c9128468b028, Last Updated: 14 January, 2022, Extract: 12 June, 2025.

66. Coccomini, D., Messina, N., Gennaro, C., & Falchi, F. (2021). Combining EfficientNet and Vision Transformers for Video Deepfake Detection. arXiv preprint arXiv:2107.02612. Available at: https://arxiv.org/abs/2107.02612

67. “VGG16, VGG19, Inception V3, Xception and ResNet-50 architectures,” https://www.researchgate.net/figure/GG16-VGG19-Inception-V3-Xception-and-ResNet-50-architectures_fig1_330478807, Last Updated: October 2018, Extract: 12 June, 2025.

68. Tipper, S., Atlam, H. F., & Lallie, H. S. (2024). An investigation into the utilisation of CNN with LSTM for video deepfake detection. *Applied Sciences, 14*(21), 9754. https://doi.org/10.3390/app14219754

69. “Keras: Time CNN+LSTM for video recognition,” Stack Overflow (Q&A by lai hang, answered by Daniel Möller), https://stackoverflow.com/questions/54696029/keras-time-cnnlstm-for-video-recognition, Last Updated: 14 February, 2019, Extract: 12 June, 2025.

70. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). *ImageNet: A large-scale hierarchical image database*. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) (pp. 248–255). https://doi.org/10.1109/CVPR.2009.5206848

71. Kaya, Hüseyin, “What Is Transfer Learning?,” https://kayahuseyinn.medium.com/what-is-the-transfer-learning-d315c1d8df7f, Last Updated: April–May 2022, Extract: 12 June, 2025.

72. Kornblith, S., Shlens, J., & Le, Q. V. (2019). *Do Better ImageNet Models Transfer Better?* In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2661–2671. https://doi.org/10.1109/CVPR.2019.00277

73. Lester, B., Al-Rfou, R., & Constant, N. (2021). *The Power of Scale for Parameter-Efficient Prompt Tuning*. arXiv preprint arXiv:2104.08691. https://arxiv.org/abs/2104.08691

74. Snorkel DataLab, “Boost Foundation Model Results with Linear Probing & Fine-Tuning,” https://snorkel.ai/blog/boost-foundation-model-results-with-linear-probing-fine-tuning/, Last Updated: 13 September, 2024, Extract: 12 June, 2025.

75. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). *An image is worth 16x16 words: Transformers for image recognition at scale*. arXiv preprint arXiv:2010.11929. https://arxiv.org/abs/2010.11929

76. Rodrigo, Marcos; Cuevas, Carlos; García, Narciso, “Comprehensive Comparison Between Vision Transformers and Convolutional Neural Networks for Face Recognition Tasks,” https://www.nature.com/articles/s41598-024-72254-w, Last Updated: 13 September, 2024, Extract: 12 June, 2025.

77. Guera, D., & Delp, E. J. (2018). *Deepfake video detection using recurrent neural networks*. In 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS) (pp. 1–6). IEEE. https://doi.org/10.1109/AVSS.2018.8639163

78. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). *FaceForensics++: Learning to detect manipulated facial images*. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 1–11). IEEE. https://doi.org/10.1109/ICCV.2019.00009

79. Reuben Suju, "Celeb-DF v2: A Large-Scale DeepFake Forensics Dataset," https://www.kaggle.com/datasets/reubensuju/celeb-df-v2, Last Updated: 2022, Extract: 9 June, 2025.

80. Shilpa Kaman and Dr. Aziz Makandar, "SDFVD: Small-scale Deepfake Forgery Video Dataset," https://data.mendeley.com/datasets/bcmkfgct2s/1, Last Updated: 23 April, 2024, Extract: 13 June, 2025

81. Saikia, P., Roy, M., & et al. (2022, August 28). *A Hybrid CNN-LSTM model for Video Deepfake Detection by Leveraging Optical Flow Features* (Preprint). arXiv. https://doi.org/10.48550/arXiv.2208.00788

82. Tariq, S., Lee, S., & Woo, S. S. (2020, September 16). *A Convolutional LSTM based Residual Network for Deepfake Video Detection (CLRNet)* [Preprint]. arXiv. https://arxiv.org/abs/2009.07480