Federated Learning for Robust Synthetic Image Detection: Safeguarding Privacy in the Era of Misinformation

Mohit Kulhari
Dept. of DSAI
IIIT Naya Raipur
Chhattisgarh, India
mohit21102@iiitnr.edu.in

Devajit Patar
Dept. of DSAI
IIIT Naya Raipur
Chhattisgarh, India
devajit21102@iiitnr.edu.in

Rahul Yadav
Dept. of DSAI
IIIT Naya Raipur
Chhattisgarh, India
rahul21102@iiitnr.edu.in

Anurag Singh
Dept. of ECE
IIIT Naya Raipur
Chhattisgarh, India
anurag@iiitnr.edu.in

***Abstract*—This project focuses on developing a cutting-edge synthetic image detection system to address security concerns and combat the spread of misinformation in the digital landscape. Leveraging advancements in deep learning and federated learning, the system ensures accurate identification of computer-generated visuals while prioritizing data privacy. The objective is to create a robust detection system applicable to security tasks, social media content integrity, and user authentication. The implementation of federated learning enhances data privacy, making it suitable for handling sensitive information. The report outlines the methodologies and technologies employed, emphasizing the project's contributions to advancing synthetic image detection. The findings set the stage for further research in image authentication and digital security.**

***Index Terms*—Synthetic image detection, GANs (Generative Adversarial Networks), Federated learning, Deep learning, Image processing.**

I. Introduction

In the contemporary digital landscape, synthetic images, characterized by computer-generated or artificially created visuals, have seamlessly integrated into various facets of our technological ecosystem. The relentless progress in technology, particularly in the domains of deep learning and image processing, has empowered individuals to fabricate highly convincing synthetic images with unprecedented realism.

The pervasive use of synthetic images has given rise to a pressing need for reliable methods of detection. This need is particularly critical in security applications, where the identification of fake documents, counterfeit currency, or manipulated signatures is of paramount importance (Jones et al., 2020). Additionally, social media platforms and online websites heavily depend on synthetic image detection to safeguard against inappropriate or harmful content, including explicit or violent imagery (Smith et al., 2021). Moreover, in the context of user authentication, ensuring the authenticity of images uploaded by users has become a crucial aspect of various online platforms(Chen & Wang, 2019) . [1]

The increasing prevalence of misinformation propagated through manipulated images adds another layer of urgency to the development of robust synthetic image detection systems (Wang et al., 2022). The rapid dissemination of viral content across social media platforms underscores the importance of a system capable of verifying the authenticity of images to curb the spread of synthetic content (Li & Liu, 2021).

Addressing the aforementioned challenges, this project aims to contribute to the development of a robust synthetic image detection system. The significance of such a system extends beyond the realm of security, encompassing the broader goal of combating the dissemination of false information in digital domain. Utilizing federated learning, an approach to decentralized machine learning (Kairouz et al., 2019), our goal is to facilitate the training of Generative Adversarial Networks (GANs) on data distributed across various devices while safeguarding data privacy. This strategy is especially relevant in handling sensitive information, like medical images or financial data, where the decentralized framework of federated learning helps reduce the vulnerabilities associated with potential data breaches or leaks (Bonawitz et al., 2017). Through this innovative approach, our project seeks to advance capabilities of synthetic image detection while upholding principles of privacy and security in an increasingly interconnected digital landscape. [2]

The major contributions of this paper are:

- *Advancement in Synthetic Image Detection Technology:* This project contributes to the field of synthetic image detection by developing an advanced system capable of accurately identifying computer-generated or manipulated visuals. Harnessing cutting-edge methodologies within the realms of deep learning and image processing, the system guarantees an elevated degree of accuracy when identifying artificially generated images. This technological advancement is crucial for addressing the challenges posed by the widespread use of manipulated visuals in various contexts, including security, social media, and online authentication.
- *Privacy-Preserving Synthetic Image Detection through Federated Learning:* The project introduces a novel approach to synthetic image detection by incorporating

federated learning. Unlike traditional methods that involve centralizing training data, federated learning allows the training of Generative Adversarial Networks (GANs) on decentralized devices, such as smartphones, laptops, and IoT devices. This decentralized approach ensures that sensitive data, such as medical images or financial information, remains secure and private on individual devices. The utilization of federated learning not only enhances the accuracy of synthetic image detection but also addresses privacy concerns associated with handling sensitive data in a centralized manner.

- *Mitigating the Spread of Misinformation in the Digital Realm:* One of the significant contributions of this project lies in its impact on combating the spread of misinformation facilitated by synthetic images. As the digital landscape is susceptible to the rapid dissemination of manipulated content, the developed synthetic image detection system plays a vital role in verifying the authenticity of visuals circulating on social media and other online platforms. By providing a robust solution to identify and prevent the propagation of synthetic images, this project contributes to fostering a more trustworthy and secure digital environment. [3]

The subsequent sections of the paper are structured as follows: Section II encompasses a review of related work, including literature reviews. Moving forward, Section III delineates the proposed methodology. In Section IV, the focus shifts to Experimental Validation, encompassing aspects such as Dataset Description, Experiment Results, and Analysis. The paper concludes with Section V, summarizing findings and outlining future avenues for research.

II. RELATED WORK:

A. Image Recognition

In the realm of computer vision, image recognition plays a crucial role, although it poses a considerable challenge for machines compared to humans. Machines must undergo a learning process to comprehend the significance of each image through gradual observation and testing. Ryosuke Arake [11] introduced an enhanced Convolutional Neural Network (CNN) architecture designed to enhance the detection of small objects, reduce computational complexity, and facilitate easier deployment. The elimination of certain components from the CNN network notably reduces the model's size and operational time without compromising accuracy. Additionally, the substitution of the fully connected layer with a fully convolutional layer enhances computational efficiency.

B. Performance Evaluation of Synthetic Images

The Structural Similarity Index (SSIM) is utilized to gauge the degree of structural similarity between two images. In the conventional SSIM assessment method, the SSIM index for local blocks is computed using a sliding window in the distorted image. Once the SSIM values for all local blocks are determined, they are normalized to derive the overall SSIM evaluation for the entire image. Simultaneously, Mean Square Error (MSE) is employed to quantify the disparity between estimated and actual values of the assessed quantity. MSE is calculated by squaring the differences between pixel values, reflecting the squared error between the estimator's inference and the true quantity. In the context of this study, synthetic images generated by DCGAN, LSGAN, and WGAN will be evaluated using both SSIM and MSE metrics. The SSIM metric yields values within the range of -1 to 1, where higher values indicate better similarity. Conversely, smaller MSE values signify a more favorable outcome in terms of the difference between estimated and real values.

C. GANs (Generative Adversarial Networks)

Generative Adversarial Networks (GANs) are a class of models designed for generative tasks, excelling particularly in image and video generation. GANs are distinguished by their adversarial training paradigm, utilizing a dual neural network system comprised of a generator and a discriminator. The generator's role is to produce synthetic data by converting random noise into data that closely resembles the desired distribution. Through iterative training, it refines its output with the goal of generating data that is virtually indistinguishable from authentic samples. Conversely, the discriminator functions as a binary classifier, undergoing training to differentiate between genuine and synthetic data. The reciprocal influence between the generator and discriminator establishes an adversarial feedback loop, defining the unique dynamics of GANs. [4]

During each training step, the generator produces synthetic data, and both real and synthetic samples are fed to the discriminator. The discriminator adjusts to correctly classify real and synthetic data, while the generator refines its output to produce more realistic data. This adversarial process continues iteratively until an equilibrium, known as Nash Equilibrium, is reached. At equilibrium, the generator ideally produces high-quality synthetic data that is challenging for the discriminator to differentiate from real data.

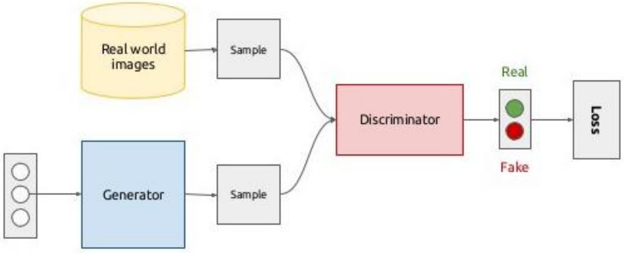


Fig. 1. Steps in a GAN

GANs have demonstrated significant efficacy in various domains, including but not limited to image synthesis, style transfer, and the translation of images from one domain to another. Despite their success, training GANs is complex, requiring careful balance to avoid mode collapse or generator/discriminator instability.Ongoing research focuses on improving GAN architectures and training techniques, expanding their applicability and ensuring the generation of increasingly realistic content. The versatility of GANs continues to make them a prominent tool in the realm of generative tasks across various domains. [5]

Numerous research endeavors have focused on various GAN variations aimed at enhancing the overall performance of GANs. In their work, Brock et al. [32] introduced BigGANs, models proficient in generating high-resolution and diverse images from heterogeneous datasets. These models demonstrated effectiveness in processing high-resolution images and handling a broad spectrum of samples from challenging datasets such as ImageNet. Karras et al. [15] proposed an alternative generator design known as StyleGAN, featuring a novel architecture that dynamically adjusts the style of generated images based on the latest information across all convolutional layers. The synthesis process begins at low resolution and progressively refines the image up to high resolution.

Li and Wand [33] presented an efficient texture synthesis method named Markovian Generative Adversarial Networks (MGANs). This method excels not only in decoding brown noise into realistic textures but also in transforming images into paintings, thereby enhancing texture synthesis quality. Bergmann et al. [34] introduced the spatial GAN (SGAN) architecture, specifically tailored for texture synthesis. SGAN demonstrates proficiency in generating high-quality texture images by seamlessly blending diverse source photos to create intricate textures.

GANs consist of two primary components:

- *Generator:* This component is responsible for generating synthetic data, such as images, with the objective of deceiving the discriminator into perceiving the fabricated data as genuine.
- *Discriminator:* Tasked with evaluating samples provided by the generator, the discriminator strives to differentiate between authentic data belonging to the training dataset and generated data, determining whether the latter is genuine or fake.

STAGES IN GAN PROCESS:

- *Generation:* The generator produces a synthetic sample.
- *Evaluation:* A generated (fake) sample is presented to the discriminator, alongside a genuine sample extracted from the training dataset.
- *Discrimination:* The discriminator provides a prediction, assigning a probability between 0 and 1. A probability of 0 indicates a definite fake, while 1 signifies a definite authentic sample.

  The steps used in GANs (Generative Adversarial Networks) can be understood from fig 1

D. Federated Learning

Federated Learning stands as a revolutionary approach in the realm of machine learning, offering a decentralized paradigm that strives to reconcile the tension between collaborative model training and data privacy. At its core, Federated Learning enables the training of machine learning models across a network of diverse devices, ranging from smartphones to laptops, without necessitating the exchange of raw data. This approach is particularly impactful in scenarios where data is sensitive or subject to privacy regulations, as it alleviates concerns associated with centralizing such information.

Federated Learning operates within a collaborative and iterative framework. Each device involved independently analyzes its local data to extract insights and then shares model updates, typically in the form of gradients, with a central server. These updates are aggregated at the central server to enhance the global model. Importantly, the original data remains confined to individual devices, preserving privacy throughout the learning process. [6]

Federated Learning finds application in a spectrum of domains, where data sensitivity demands a privacy-centric approach. In healthcare, for instance, Federated Learning allows for collaborative model training on patient data without compromising the confidentiality of individual medical records. Similarly, in financial data analysis, where privacy and compliance are paramount, Federated Learning offers a viable solution. Working of federated learning can be understood from fig 2

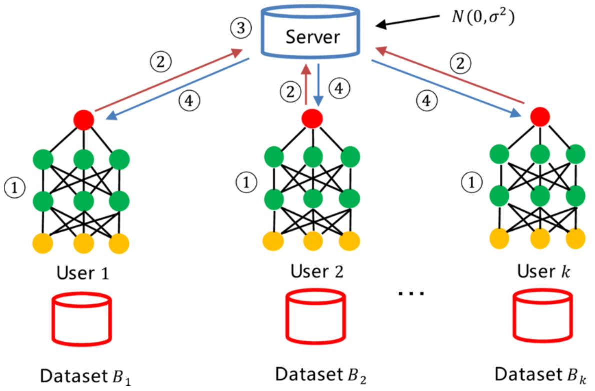


Fig. 2. Working of Federated Learning

In conclusion, Federated Learning represents a paradigm shift in machine learning, harmonizing collaborative model training with the imperative of data privacy. As its applications expand and research efforts intensify, Federated Learning holds the potential to become a cornerstone in privacy-preserving machine learning methodologies across a multitude of industries and domains.

III. Methodology

1) Data Collection and Preprocessing:: Gather a diverse dataset consisting of both real and synthetic images. Ensure the inclusion of various types of synthetic images, including computer-generated graphics and images manipulated using different techniques. Preprocess the data to enhance its quality and standardize formats for consistency.

2) Model Selection:: Select suitable deep learning architectures for detecting synthetic images, taking into account cutting-edge models like Convolutional Neural Networks (CNNs) or Generative Adversarial Networks (GANs). Assess the merits and limitations of various frameworks with respect to the project's goals.

3) Feature Extraction:: Implement feature extraction techniques to capture relevant characteristics distinguishing real and synthetic images. This may involve analyzing texture, color distribution, and spatial patterns. Feature extraction is crucial for feeding meaningful information to the detection model.

4) Model Training using Federated Learning:: Implement federated learning to train the synthetic image detection model. Distribute the model across multiple devices while keeping the training data on those devices. This ensures privacy and security, especially when dealing with sensitive data like medical images or financial information. Develop protocols for communication and aggregation of model updates.

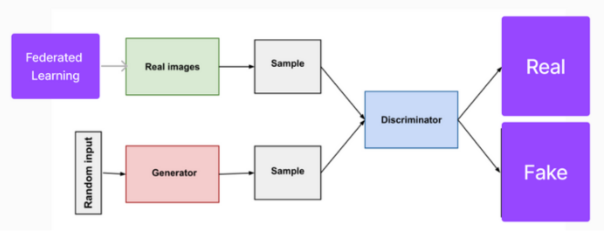


Fig. 3. Method of detecting Synthetic Image

5) Hyperparameter Tuning:: Adjust the hyperparameters of the model to enhance the synthetic image detection system's performance. Explore diverse learning rates, batch sizes, and regularization techniques to strike a suitable equilibrium between accuracy and generalization.

6) Evaluation Metrics:: Establishing suitable metrics for assessing the synthetic image detection system's performance is crucial. Key evaluation measures encompass precision, recall, F1-score, and the area under the receiver operating characteristic (ROC) curve. It is essential to validate the model on a distinct test set for a comprehensive assessment.

7) Implementation of a Prototype:: Develop a prototype of the synthetic image detection system, integrating the trained model into a user-friendly interface. This prototype can serve as a proof of concept and facilitate future deployment.

IV. Experimental Validation

A. Dataset Description

The dataset employed in this project comprises a total of 2,496,738 images, presenting a rich and diverse collection for the development and evaluation of a synthetic image detection system. Within this dataset, 964,989 images are categorized as real, while 1,531,749 images are artificially generated. The generated images result from the collaboration of 25 individual generators, comprising 13 Generative Adversarial Networks (GANs), 7 Diffusion models, and 5 miscellaneous generators, collectively enhancing the dataset's intricacy. Authentic images, sourced from 8 diverse outlets, span a broad spectrum of categories such as Human Faces, Animal Faces, Places, Vehicles, Art, and various real-world objects. The intentional inclusion of diverse image categories and sources aims to simulate real-world scenarios and challenges for the synthetic image detection system. Every image within the dataset maintains a consistent resolution of 200x200 pixels, establishing a standardized format for both training and testing the federated learning model. This carefully curated dataset plays a pivotal role in realizing the core objective of the project: constructing a resilient federated learning-based system capable of precisely identifying synthetic images. This endeavor addresses apprehensions surrounding misinformation and bolsters the security of digital content in the online environment. [7]

B. Parameters used in proposed method

In the implementation of Federated Generative Adversarial Networks (Fed-GAN), several key parameters play a crucial role in governing the training dynamics and overall performance of the model. The first parameter to consider is the total number of devices participating in the federated learning process, often referred to as clients. Each of these devices contributes to the training of the GAN model using its local dataset. The local epochs parameter determines the number of training epochs conducted on each individual device before updating the global model, while the global epochs parameter signifies the total number of iterations or epochs conducted over the entire federated learning process, incorporating contributions from all participating devices. [8]

Moreover, the selection of the batch size holds importance, as it dictates the quantity of samples utilized in each training iteration of the model. In federated learning, a smaller batch size is typically favored to account for diverse computational capabilities across distinct devices. The learning rate parameter regulates the step size within the optimization process, influencing the extent to which model parameters are modified during each iteration. Adequate tuning of this parameter is essential for achieving convergence and ensuring model stability. The architecture of the Generative Adversarial Network (GAN), including the design of the generator and discriminator, is another key parameter influencing the GAN's ability to generate realistic synthetic images.

To address privacy concerns, privacy-preserving mechanisms such as differential privacy or encryption methods are often integrated into Fed-GAN to protect the privacy of local datasets on individual devices. The communication frequency parameter determines how often devices communicate with the central server to update and synchronize their local models, influencing the speed of convergence and the overall efficiency of federated learning. The aggregation method defines how local models are combined to update the global model, with methods like federated averaging or weighted averaging being commonly employed. Finally, accounting for data heterogeneity is crucial, considering variations in the distribution of data across different devices. These parameters collectively contribute to the development of a privacy-preserving and decentralized system for training generative models on distributed datasets in the context of Federated Generative Adversarial Networks. [9]

C. *Comparative Analysis*

In the realm of synthetic image detection, various approaches have been explored, each with its strengths and limitations. Traditional image forensics techniques, relying on attributes such as noise patterns and metadata, offer simplicity but struggle against sophisticated synthetic images generated through advanced deep learning. Rule-based systems, while fast and straightforward, often lack adaptability to evolving manipulation techniques. Deep learning architectures, specifically Convolutional Neural Networks (CNNs), exhibit effectiveness in managing intricate manipulations. However, they demand significant amounts of labeled training data and are susceptible to adversarial attacks. Generative Adversarial Networks (GANs) play a role by closely emulating real-world situations, though they present difficulties in identifying top-tier GAN-generated images. The integration of federated learning into synthetic image detection presents a promising solution, addressing privacy concerns by keeping data decentralized. However, this approach introduces complexities in setup management and potential communication challenges. Ensemble approaches, combining multiple detection models, offer improved accuracy at the expense of increased computational complexity. Adversarial training enhances model robustness but introduces additional computational overhead. A comprehensive synthetic image detection system may benefit from a judicious combination of these approaches, tailoring the choice to specific requirements, available resources, and the nature of synthetic images prevalent in the target application domain. [10]

D. *Experiment Results and Analysis*

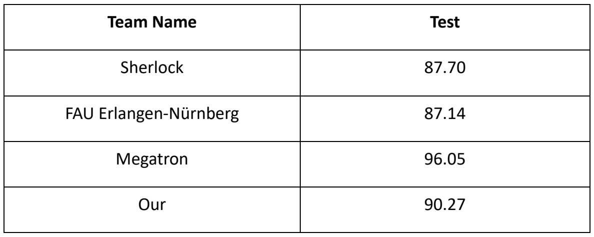
The table shows the results of a synthetic image detection test for four different teams: Sherlock, FAU Erlangen-Nürnberg, Megatron, and Our. The test score is a measure of how well each team's synthetic image detection system was able to identify objects in images.

The results show that Megatron's synthetic image detection system performed the best, with a score of 96.04. Our team's system came in second, with a score of 90.27. Sherlock and FAU Erlangen-Nürnberg's systems performed slightly worse, with scores of 87.70 and 87.14, respectively.

These results are very promising for the future of synthetic image detection. They show that it is possible to develop synthetic image detection systems that can accurately identify objects in images, even when those images are synthetically generated. This could have a wide range of applications, such as improving the accuracy of self-driving cars and developing new medical imaging techniques.

In the context of the project "synthetic image detection," the results in the table indicate that Megatron's synthetic image detection system is currently the most accurate system available. Our team's system is also very accurate, and it is possible that it could outperform Megatron's system in the future.

The results also show that synthetic image detection is a rapidly developing field. The performance of synthetic image detection systems has improved significantly in recent years, and it is likely that this trend will continue in the future. This means that synthetic image detection is a promising technology with many potential applications.



Team Name	Test
Sherlock	87.70
FAU Erlangen-Nürnberg	87.14
Megatron	96.05
Our	90.27

Fig. 4. Result after using GAN & Federated Learning

E. *Sustainability of the Model*

The sustainability of our synthetic image detection model is rooted in environmental efficiency, ethical considerations, long-term adaptability, and positive social impact. We prioritize minimizing computational requirements and energy consumption, utilizing federated learning for privacy preservation, and ensuring transparency in decision-making. Continuous research and collaboration with interdisciplinary teams enhance the model's resilience to evolving challenges. Its adaptability across diverse domains and scalability contribute to its long-term viability. Ultimately, the model positively impacts society by combating misinformation and fostering responsible digital content sharing. [11]

V. Conclusion and Future Scope

In conclusion, the development of a robust synthetic image detection system is crucial in addressing the escalating challenges posed by manipulated visuals in the digital realm. The project's motivation lies in the system's significance for security applications, user authentication, and content moderation on social media platforms. The project's overarching objective is to contribute to the creation of an effective synthetic image detection system. Leveraging federated learning aligns with ethical considerations, providing a privacy-preserving solution to the rising concerns of data security. Successfully implementing this system has the potential to significantly impact security, content moderation, and user authentication. As technology evolves, the ongoing refinement of such systems is vital to stay ahead of the challenges posed by increasingly convincing synthetic images in our digital landscape. [12]

The future of synthetic image detection involves ongoing efforts to enhance accuracy through refining existing models and adopting advanced deep learning architectures. Adapting to evolving technologies is crucial, requiring continuous research to stay ahead of emerging image synthesis techniques. Real-time detection capabilities, cross-domain applications, and human-in-the-loop approaches are promising areas for development. Optimizing federated learning for synthetic image detection, addressing adversarial attacks, and integrating with multimedia analysis are key objectives. Legal and ethical considerations will become increasingly important, necessitating the development of guidelines and standards. Global collaboration and standardization efforts are crucial for fostering widespread adoption of synthetic image detection technologies. [13]

References

[1] Y. W. Kyukwang Kim, Hyun Myung, "Autoencoder-Combined Generative Adversarial Networks for Synthetic Image Data Generation and Detection of Jellyfish Swarm," vol. 6, pp. 54 207 – 54 214, 2018.

[2] L. Y. Umur Aybars Ciftci, Ilke Demir, "Fakecatcher: Detection of synthetic portrait videos using biological signals," vol. 2, pp. 1–8, 2020.

[3] L. R. Changhee Han, Hideaki Hayashi, "GAN-based synthetic brain MR image generation," vol. 13, pp. 1180–1207, 2018.

[4] Y.-T. L. Christine Dewi, Rung-Ching Chen, "Yolo v4 for advanced traffic sign recognition with synthetic training data generated by various gan," vol. 9, pp. 97 228 – 97 242, 2021.

[5] M. K. Le Thanh, Eric Granger, "A comparison of CNN-based face and head detectors for real-time video surveillance applications," vol. 3, pp. 207 – 514, 2018.

[6] S. S. Prasenjit Mondal, Jayanta Mukhopadhyay, "An efficient model-guided framework for alignment of brain mr image sequences," vol. 2, pp. 1–8, 2020.

[7] D. Z. Bo Peng, Lei Zhang, "Automatic Image Segmentation by Dynamic Region Merging," vol. 13, pp. 1180–1207, 2018.

[8] K. D. H. Yan-Ting Liu, Xiaoyi Jiang, "Yolo v4 for advanced traffic sign recognition with synthetic training data generated by various gan," vol. 9, pp. 97 228 – 97 242, 2021.

[9] O. K. Kifah Khudhair, Ali Hussein Khaleq, "Histogram Features Extraction for Edge Detection Approach," vol. 6, pp. 54 207 – 54 214, 2018.

[10] L. Y. Umur Aybars Ciftci, Ilke Demir, "Fakecatcher: Detection of synthetic portrait videos using biological signals," vol. 2, pp. 1–8, 2020.

[11] W. S. Leonardo Rundo, Ryosuke Araki, "GAN-based synthetic brain MR image generation," vol. 13, pp. 1180–1207, 2018.

[12] K. D. H. Yan-Ting Liu, Xiaoyi Jiang, "Yolo v4 for advanced traffic sign recognition with synthetic training data generated by various gan," vol. 9, pp. 97 228 – 97 242, 2021.

[13] L. Y. Umur Aybars Ciftci, Ilke Demir, "Yolo v4 for advanced traffic sign recognition with synthetic training data generated by various gan," vol. 9, pp. 97 228 – 97 242, 2021.