Adversarial Attacks on Facial Recognition: Analysis of Attack Techniques and Defense Strategies

Ricky Yang

May 15, 2024

Abstract

Facial recognition systems are increasingly used in security, identity authentication, and social services, yet adversarial attacks threaten their reliability. This paper reviews prominent attack methods like Fast Gradient Sign Method (FGSM), Basic Iterative Method (BIM), and Generative Adversarial Networks (GANs), and analyzes defensive strategies including adversarial training, input transformation, and detection and rejection. Adversarial training enhances robustness but may compromise accuracy on clean inputs, while input transformation and detection strategies offer promising results with notable trade-offs. The analysis highlights the challenges of evolving adversarial techniques and emphasizes the importance of hybrid defensive models to secure facial recognition systems. Future research should focus on developing resilient, efficient, and adaptive strategies to ensure the long-term reliability and security of facial recognition technology.

1 Introduction

In recent years, facial recognition technology has been widely used in fields such as security monitoring, identity authentication, and social services. However, this technology faces the threat of adversarial attacks. Adversarial models can deceive or even destroy the face recognition system, seriously affecting the reliability and security of the system[1] . Adversarial attacks interfere with the recognition model by adding imperceptible perturbations to the input image, causing it to output incorrect results[2]. In order to ensure the security and reliability of face recognition technology, it is of great significance to study how to defend against these attacks.

This paper provides a comprehensive overview of the current state of adversarial models through a detailed literature review.

The main adversarial attack methods, such as Fast Gradient Symbol Method (FGSM)[3], Basic Iterative Method (BIM)[4, 5] and Generative Adversarial Networks (GANs)[6, 7, 8], are analyzed, their advantages and disadvantages are compared, and their impact on the model is evaluated.

The defense strategies section introduces defense methods such as adversarial training[9, 10], input transformation[11, 12, 13], model regularization, and detection rejection, and analyzes their effectiveness in practical applications.

The following analysis and elaboration section will deeply explore the security threats of the adversarial model and the feasibility and limitations of different defense methods from three aspects: the importance of the research results, current challenges and strategies, and industry impact.

Finally, point out the shortcomings of existing research, and make suggestions for future research and practice, emphasizing the development of more robust defense methods to ensure that the face recognition system can be used in information Reliability and security in the information security field

2 Literature review

2.1 Methodology

This literature review aimed to identify recent research on adversarial attacks and defenses in facial recognition systems. The focus was on understanding current attack methods and evaluating the effectiveness of various defensive strategies.

1. **Database Search:** We searched academic databases, including the AUT Library, Google Scholar, IEEE Xplore, and ACM Digital Library, for papers published since 2020. Keywords such as "adversarial attacks on facial recognition," "adversarial defense facial recognition security," and "GANs adversarial attacks facial recognition" were used to locate relevant research.

2. **Selection Criteria:** Over 50 papers were initially screened, and around 20 were selected based on the following criteria:

    - **Publication Date:** Papers published from 2020 onward were considered to capture the most recent advancements.
    - **Relevance:** Research had to address adversarial attacks or defensive strategies specifically targeting facial recognition systems.
    - **Quality and Innovation:** Papers offering novel techniques or comprehensive analyses were prioritized.
    - **Methodological Rigor:** Studies were selected for their thorough evaluation, testing, or innovative benchmarks.

3. **Review Process:** Each selected paper was analyzed for its approach to adversarial attacks or defenses, with attention to clarity, methodological soundness, and practical applications. The findings were categorized based on the specific attack or defense techniques to identify common themes, research gaps, and potential areas for further study.

This structured approach ensured that the selected literature reflected the latest developments while providing a balanced view of the current landscape in adversarial security for facial recognition.

2.2 Background

Facial recognition systems are highly susceptible to adversarial attacks, which involve imperceptible perturbations that mislead machine learning models into making incorrect predictions. First identified by Szegedy et al. in 2014[1], adversarial attacks have evolved to include the Fast Gradient Sign Method (FGSM)[3], Basic Iterative Method (BIM)[4, 5], and Generative Adversarial Networks (GANs)[6, 7, 8].

These attacks pose significant risks across various domains, including finance, surveillance, and autonomous vehicles. Attackers can exploit these vulnerabilities to impersonate identities, bypass security checks, or disrupt recognition systems, leading to potential fraud and accidents[14, 15, 16, 2].

Defensive measures like adversarial training, input transformation, and detection mechanisms provide a line of defense but face challenges like increased computational complexity and reduced accuracy on clean data. This review analyzes key adversarial attack strategies and examines defensive mechanisms to highlight evolving challenges and potential research directions for securing facial recognition systems.

2.3 Definition

Adversarial attacks are a means of using small perturbations to change input data to deceive machine learning models, especially in the field of face recognition. An attacker can add carefully designed perturbations to the input image to cause the model to misjudge the originally correct recognition results. Defense mechanisms are methods to resist these attacks, with the main purpose of allowing the recognition model to maintain accuracy and robustness in the face of adversarial attacks[16].

Table 1: Comparison of Attack Methods

Attack Method	Approach	Target	Effectiveness
FGSM	Single-step gradient perturbation	Specific models (white-box attack)	High, but vulnerable to some detection mechanisms
BIM (PGD)	Iterative gradient perturbation	Specific models (white-box attack)	Very high; iterative approach allows subtle attacks
GANs (CycleGAN)	Generates adversarial samples in specific styles	Impersonation, dodging	High realism; evades detection with realistic styles
GANs (StyleGAN)	Creates high-fidelity images with style control	Impersonation, evasion	Extremely high fidelity; capable of deceiving models
GANs (Patch-GAN)	Generates adversarial patches for disguise	Physical disguise, dodging	Effective even in physical environments

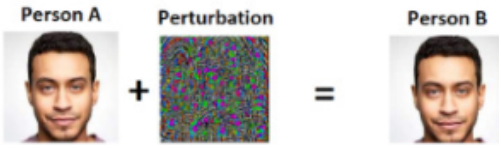

Figure 1: FGSM Attack

2.4 Main attack methods

2.4.1 FGSM

Fast Gradient Symbol Method (FGSM) is a classic adversarial attack method proposed by Goodfellow et al. in 2015[17]. This approach computes the gradient of the target model's loss function with respect to the input, allowing the generation of adversarial examples.

In a white-box attack (Figure2), the adversary has full access to the target model's architecture and parameters, which helps in computing the gradients required for generating adversarial examples . For black-box attacks (Figure3), attackers might train a substitute model to replicate the target model's behavior, then use FGSM to generate adversarial examples that can transfer to the target[18].

- How FGSM Works

  1. Assume the input image is x, with a ground truth label y. The target model's loss function is $J(x, y)$.
  2. By computing the gradient of the loss function concerning the input x, denoted as $\nabla_x J(x, y)$, FGSM adds a perturbation ϵ in the direction of the sign of the gradient.
  3. The generated adversarial example x' is represented as:

     $$x' = x + \epsilon \cdot \text{sign}(\nabla_x J(x, y))$$

     where the sign function represents the sign of the gradient.

- Experiment Results

  Jagadeesha used FGSM to generate adversarial perturbations that alter facial features FGSM can slightly adjust the gradients of an original face image to produce an adversarial face image. Although visually similar to the original image, the adversarial example misleads the facial recognition model into incorrect classification. This perturbation can change how a facial recognition system perceives the image, leading to errors like false negatives or impersonation[18].

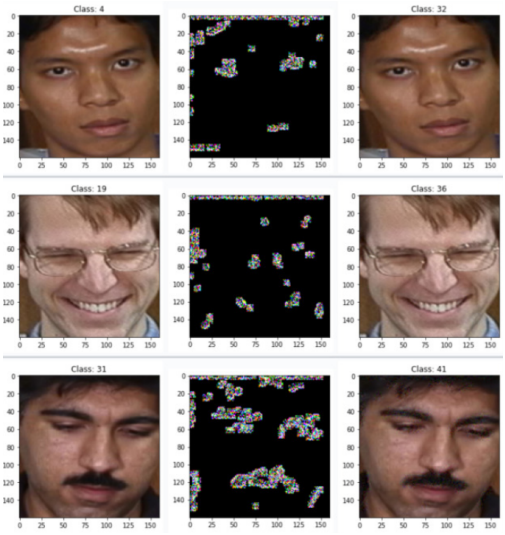

Figure 2: FGSM Attack - White Box

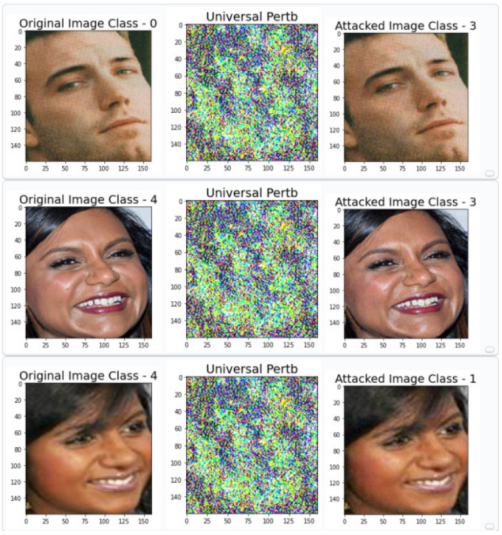

Figure 3: Universal Perturbation Attack - Black Box

2.4.2 BIM

The Basic Iterative Method (BIM) is an extended version of the Fast Gradient Sign Method (FGSM) and aims to generate stronger adversarial examples by iterating the steps of FGSM multiple times. Each iteration uses smaller perturbations to increase stealth and reduce the likelihood of the attack being detected[19].

- How BIM Works

  1. The initial input image is x with a ground truth label y. Set the perturbation step size α and the number of iterations n.
  2. During each iteration, perturbations are added in the direction of the gradient of the loss function:

$$x_{i+1} = \text{Clip}(x_i + \alpha \cdot \text{sign}(\nabla_x J(x_i, y)))$$

  3. The final adversarial example is x_n, where each iteration result is clipped (using Clip) to ensure that the perturbation remains within a reasonable range.

- Experiment Results

  Yasmeen M. Khedr and colleagues emphasizes how the iterative nature of BIM increases its strength and ability to fool recognition systems compared to single-step attacks like FGSM. The perturbations are added to the input in a way that progressively pushes the modified image further from the original in terms of classification, making it a more effective adversarial example.

2.4.3 GANs

Generative Adversarial Networks (GANs) consist of a generator that creates synthetic data and a discriminator that differentiates between real and generated data. They are commonly used to produce realistic facial images. GANs generate realistic data through two competing neural networks: a generator and a discriminator[20].

- How GANs Works

  1. Generator: The generator network attempts to create realistic samples from random noise. Its goal is to deceive the

discriminator into believing the generated samples are real.

2. Discriminator: The discriminator network aims to distinguish between real data and fake samples generated by the generator. Its adversarial training with the generator improves its ability to identify real and fake samples.
3. Adversarial Training: The training process involves a competition between the generator and the discriminator. The generator tries to create increasingly realistic samples to deceive the discriminator, while the discriminator keeps learning to improve its ability to distinguish between real and fake samples.

- Attack Methods[20]

1. The GAN model creates adversarial patches in specific face regions, like glasses or stickers, to fool recognition systems.
2. Dodging attacks involve making the system identify the attacker as any other person in the database.
3. Impersonation attacks attempt to have the attacker recognized as a specific target.

- Experiment Results

  In the digital realm, Zolfi and colleagues reported that GAN-generated patches were able to evade face recognition systems with a 57.99% success rate for dodging attacks and a 48.78% success rate for impersonation attacks. When these patches were printed and physically applied to participants, the success rate for dodging attacks increased to 81.77%, while impersonation attacks succeeded at a rate of 63.85%[20].

2.4.4 Key GAN Variants

- **CycleGAN[21]:**
  - **Approach:** CycleGAN learns to translate images between two different domains without requiring paired training samples. It achieves this by maintaining cycle consistency, meaning an image translated from domain A to B and back to A should remain the same.
  - **Target:** Impersonation and dodging attacks through domain adaptation, where adversarial images are generated to mimic the style and features of different facial identities.
  - **Effectiveness:** Highly successful at evading facial recognition systems by introducing domain-specific features that create realistic images while retaining adversarial perturbations.
- **StyleGAN[22]:**
  - **Approach:** StyleGAN uses a style-based generator architecture that separates the image generation process into different layers. By controlling the latent space, it can manipulate facial attributes, such as identity and expression, creating high-fidelity images.
  - **Target:** Impersonation and evasion attacks by generating adversarial facial images that manipulate features like age, gender, and hairstyle.
  - **Effectiveness:** Exceptional realism makes these images difficult to distinguish from authentic ones, allowing them to deceive facial recognition systems.
- **PatchGAN[23]:**
  - **Approach:** PatchGAN focuses on specific facial regions, such as eyeglasses or stickers, and applies adversarial patches to these areas. This misleads facial recognition models by creating local perturbations.
  - **Target:** Physical-world disguise attacks that use adversarial patches in specific regions for dodging or impersonation.
  - **Effectiveness:** Highly effective in dodging or impersonating attacks, particularly in physical environments where adversarial patches can disguise identities.

2.5 Defensive Strategies

2.5.1 Adversarial Training

Adversarial Training is a defense strategy designed to improve the robustness of facial recognition systems against adversarial attacks. This strategy involves training the neural network with a mixture of both normal and adversarial examples. By introducing adversarial examples during the training phase, the system becomes more resilient to adversarial inputs that aim to fool it.[24]

- How Adversarial Training Works
    1. Generating Adversarial Samples: The first step in adversarial training is to generate perturbed adversarial samples using methods like the Fast Gradient Sign Method (FGSM) or the Basic Iterative Method (BIM).
    2. Training Process: During the training process, both original data and adversarial samples are mixed to train the model, allowing it to identify and resist adversarial samples.
    3. Model Adjustment: As training progresses, the model gradually learns to ignore perturbations, maintaining accuracy and robustness in the face of adversarial samples.
- Drawbacks of Adversarial Training
    - **Training Complexity**: The process can significantly increase training time and computational resources required, as multiple iterations are needed to account for varying perturbations.
    - **Impact on Model Generalization**: While adversarial training enhances robustness against adversarial examples, it may inadvertently reduce accuracy on clean (non-adversarial) inputs, thus affecting overall generalization.
- Experiment Results

  In their study, Vakhshiteh and colleagues demonstrated that adversarial training enhanced the robustness of face recognition models. In digital environments, adversarial training improved the models' ability to correctly classify adversarial examples produced using various attack strategies. Despite attempts to evade or impersonate identities, the models retained a significant level of accuracy. In physical environments, where attackers employed physical patches or disguises, adversarial training enabled the models to accurately identify individuals wearing adversarial patches or other forms of disguise. This improved recognition success rate indicated that adversarial training significantly increased the system's robustness against physical attacks, including those involving GAN-generated adversarial patches[24].

2.5.2 Input Transformation

Input transformation strategies are defenses against adversarial attacks that aim to reduce the impact of perturbations by altering the input data's features.

- How Input Transformation Works
    1. **Random Rotation**: Rotating the image by a random angle can change its spatial structure and disrupt the specific shape or direction of adversarial perturbations.
    2. **Blurring**: Applying Gaussian blur or median filtering reduces the clarity of images, making small perturbations less noticeable and less likely to impact model decisions.
    3. **JPEG Compression**: By compressing the image through the JPEG format, perturbations may be removed due to data loss inherent in the compression process.
    4. **Pixel Deflection**: Randomly deflecting pixels to neighboring locations disrupts the uniformity of adversarial perturbations without drastically changing the image's overall structure.
    5. **Color Transformation**: Changing the image's color space (e.g., from RGB to HSV) may reduce the effectiveness of adversarial perturbations.

Table 2: Comparison of Defense Strategies

Strategy	Approach	Strengths	Weaknesses
Adversarial Training	Training models using adversarial examples	Improves robustness to known adversarial patterns	Time-consuming, impacts accuracy on clean data
Input Transformation	Modify input via rotations, blurring, etc.	Computationally efficient, relatively simple	Sophisticated attacks may bypass transformations
Detection and Rejection	Detection models, heuristic rules, ensemble methods	Effective in identifying adversarial samples	Computationally intensive, may result in false positives

- Experiment Results

  Vakhshiteh and colleagues examined the effectiveness and trade-offs of input transformation defenses in their study. They observed that techniques like JPEG compression and pixel deflection increased the model's robustness against common adversarial attacks. However, more sophisticated attacks can occasionally circumvent these defenses by crafting perturbations that withstand transformations. Additionally, some transformations, particularly lossy compression, may compromise model accuracy on clean (non-adversarial) inputs due to data loss[24].

2.5.3 Detection and Rejection Strategies

Detection and rejection strategies are designed to identify adversarial samples and filter them out before further processing. These strategies rely on machine learning models or heuristic rules to detect unusual patterns indicating adversarial interference.

Approaches:

1. **Detection Models:**
   - **Overview:** Deep learning models or machine learning algorithms are trained to recognize patterns and features indicative of adversarial perturbations. They act as a secondary layer of defense by analyzing images for suspicious characteristics.
   - **Example:** A secondary neural network can detect anomalies such as unusual spatial distribution, entropy, or pixel variance, which may suggest adversarial tampering.
   - **Challenges:** Detection models require large datasets of known adversarial examples for effective training and can be computationally intensive.
2. **Heuristic Rules:**
   - **Overview:** Rule-based methods utilize statistical characteristics like entropy, mean, variance, and spatial distribution to detect abnormal patterns in input images.
   - **Example:** An image with entropy or variance significantly outside expected bounds might indicate adversarial manipulation.
   - **Challenges:** Heuristic rules are often domain-specific and can struggle with subtle adversarial attacks.
3. **Consistency Checks:**
   - **Overview:** Consistency checks involve applying multiple transformations to an input image and comparing the model's predictions for discrepancies.
   - **Example:** Different image rotations, flips, or crops can help detect inconsistencies if an adversarial sample tries to maintain its perturbations.
   - **Challenges:** Consistency checks can be computationally intensive and may still

miss perturbations if crafted to withstand such transformations.

4. **Ensemble Methods:**
    - **Overview:** Ensemble methods combine the outputs of multiple models or preprocessing steps to detect adversarial patterns through consensus.
    - **Example:** A divergence in predictions between different models can indicate perturbation, while consensus across models can improve detection.
    - **Challenges:** Ensemble methods may require significant computational resources, and results can vary depending on the models chosen.

Experiment Results:
Studies by Vakhshiteh et al. [24] demonstrate the effectiveness and trade-offs of different detection and rejection strategies:

- **Feature-Based Approaches:** - These rely on deep learning models to identify adversarial discrepancies but can struggle with subtle perturbations.
- **Statistical Analysis:** - Statistical methods recognize adversarial patterns outside the normal data distribution but may require intensive resources.
- **Consistency Checks:** - Comparing predictions across multiple transformations can identify subtle adversarial attacks with high precision.
- **Ensemble Methods:** - Combining models or preprocessing steps helps diversify detection methods, improving robustness against attacks.

3 Analysis and Discussion

3.1 Significance of the Research Findings

The body of research around adversarial attacks on facial recognition systems highlights the growing sophistication of attack methods, such as FGSM, BIM, and GANs. FGSM is notable for its simplicity and speed, whereas BIM, with its iterative approach, significantly strengthens the effectiveness of adversarial attacks. GANs, capable of generating highly realistic adversarial samples, offer a profound threat to facial recognition models due to their ability to create perturbations that are not easily detectable.

The effectiveness of adversarial training and input transformation defenses reflects significant progress in securing face recognition systems. Adversarial training shows improved robustness against known adversarial examples in both digital and physical realms. Vakhshiteh and colleagues' findings reinforce the importance of this approach in maintaining accuracy despite sophisticated perturbations. Furthermore, input transformation strategies like JPEG compression and pixel deflection are promising due to their relatively low computational requirements.

3.2 Challenges and Strategies

3.2.1 Evasion of Defensive Mechanisms

A significant challenge lies in designing defenses that can adapt to the continuously evolving nature of adversarial attacks. Advanced GAN-based attacks can still bypass some input transformation defenses by designing more intricate perturbations that remain effective even after transformations like compression. This issue highlights the need for dynamic defense strategies that evolve alongside new adversarial techniques.

3.2.2 Trade-offs in Defensive Approaches

Adversarial training, while enhancing robustness, often results in a trade-off with computational efficiency and accuracy on clean inputs. Lossy compression can reduce accuracy, limiting its utility as a defense. Strategies that rely on detection and rejection are computationally intensive and may yield false positives, especially with subtle perturbations. Balancing accuracy, robustness, and computational efficiency remains a core challenge.

3.2.3 Physical World Attacks

Physical-world adversarial attacks, such as disguises or printed patches, pose another challenge due to their ability to deceive models in real-world scenarios. Traditional digital defenses may not be directly applicable, requiring the development of specialized strategies for physical adversarial perturbations.

3.3 Comparative Insights

Adversarial attacks on facial recognition differ significantly from those in other machine learning domains, such as image classification and natural language processing, due to the following factors:

1. **Personal Identity Impacts:** Facial recognition adversarial attacks directly affect identity verification by manipulating facial features to impersonate individuals or dodge detection. In other domains like object classification or speech recognition, the impact is often less personal and more related to system functionality rather than personal security.

2. **Physical World Context:** Physical disguises, such as adversarial patches (e.g., glasses or stickers), have a significant effect on facial recognition systems. These attacks target specific facial regions to mislead the model in real-world environments, while adversarial attacks in other domains often occur in digital settings.

3. **Continuous Updates:** Facial recognition models need frequent retraining or adaptation due to the changing nature of facial features over time (e.g., age, expression, and health). Adversarial attacks exploit this need by targeting weaknesses arising from outdated training data. In other machine learning domains, input data may remain more consistent and less prone to these changes.

4. **Target-Specific Evasion:** Adversarial attacks on facial recognition often involve evasion tactics that target specific individuals. For instance, attackers can impersonate a victim to bypass identity verification systems. In contrast, domains like image classification usually face broader evasion challenges involving generic object classes rather than specific targets.

3.4 Impact and Future Directions

Adversarial attacks on facial recognition systems have a significant impact, particularly in high-stakes areas like financial security, surveillance, and autonomous systems. Research indicates that adversarial robustness must be a key consideration when deploying facial recognition models. The vulnerabilities highlighted by Zolfi and colleagues reveal how susceptible these models are to GAN-based perturbations, emphasizing the urgent need for adaptive defense strategies.

Future research should focus on improving adversarial robustness while maintaining model accuracy. A promising approach involves developing hybrid models that combine adversarial training with real-time detection mechanisms, creating a more dynamic defense system. Additionally, establishing a standardized evaluation framework for adversarial defenses across various attack methods would allow for a comprehensive assessment of their effectiveness. Investigating adversarial robustness in environments with limited computational resources or real-time constraints is another crucial research direction.

Furthermore, physical disguises and printed patches present a unique challenge that traditional digital defenses are ill-equipped to handle. Therefore, specialized strategies are required to address these real-world adversarial tactics effectively.

4 Conclusion

This paper provides a comprehensive overview of adversarial attacks on facial recognition systems and defensive mechanisms designed to counter them. From the foundational FGSM method to advanced GANs, the spectrum of attack techniques has become increasingly sophisticated, necessitating more robust and dynamic defense strategies.

The literature review and analysis reveal the strengths and weaknesses of different defense mecha-

nisms. Adversarial training stands out for its ability to maintain model accuracy, even under adversarial perturbations. However, this approach often requires a significant computational investment and may reduce accuracy on clean data. Input transformation strategies like JPEG compression and pixel deflection are effective but can also introduce a trade-off with accuracy due to data loss. Detection and rejection strategies, though valuable, remain computationally intensive and prone to false positives.

In light of these challenges, future research should prioritize developing hybrid defense mechanisms that combine adversarial training, input transformation, and real-time detection. Emphasizing robust evaluation frameworks for various defensive methods will be crucial in understanding their effectiveness across different adversarial attacks. Moreover, improving the efficiency of adversarial defenses will ensure their applicability in real-time and resource-constrained environments.

While current adversarial defense mechanisms show promise, the field remains an ongoing battle between attack and defense strategies. As adversarial models become more nuanced, securing facial recognition systems against these sophisticated threats will be imperative. Future research should also consider industry-specific impacts, ensuring that defense mechanisms align with the unique needs of high-stakes fields like security monitoring, autonomous driving, and financial services. Ultimately, building more resilient and accurate models will ensure the long-term reliability and safety of facial recognition technologies in real-world applications.

References

[1] R. Hwang, J. Lin, and H. Lin, "Adversarial patch attacks on deep-learning-based face recognition systems using generative adversarial networks," *Sensors*, vol. 23, no. 2, p. 853, 2023.

[2] N. Akhtar and A. Mian, "Adversarial examples: Attacks and defenses for deep learning," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[3] R. Saxena, A. S. Adate, and D. Sasikumar, "A comparative study on adversarial noise generation for single image classification," *International Journal of Intelligent Information Technologies*, vol. 16, pp. 75–87, 2020.

[4] Y. Wang *et al.*, "Vulnerability of time series models to adversarial attacks," *IEEE*, 2023.

[5] W. He *et al.*, "A survey of adversarial attacks and defenses in deep learning," *Algorithms*, 2022.

[6] C. Zhang *et al.*, "Generative adversarial networks for facial expression recognition," *Pattern Recognition Letters*, 2023.

[7] D. Li *et al.*, "Improving adversarial robustness via generative adversarial networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[8] A. Singh *et al.*, "Review on gans for face generation and their applications," *Applied Sciences*, 2021.

[9] J. Lin *et al.*, "Adversarial training strategies in face recognition," *IEEE*, 2021.

[10] Y. Zhang *et al.*, "Improving adversarial robustness in deep face recognition models," *Sensors*, 2022.

[11] W. Liu *et al.*, "Adversarial robustness through input transformation," *IEEE Transactions*, 2021.

[12] T. Wang *et al.*, "Improving deep learning models with transformation-based defense," *Sensors*, 2022.

[13] H. Kim *et al.*, "Input transformation techniques in adversarial defense," *Journal of Machine Learning*, 2023.

[14] I. Fursov, M. Morozov, N. Kaploukhaya, E. Kovtun, R. Rivera-Castro, G. Gusev *et al.*, "Adversarial attacks on deep models for financial transaction records," *arXiv preprint arXiv:2106.08361*, 2021.

[15] T. Wang *et al.*, "Adversarial attacks on deep learning in autonomous driving systems," *IEEE*, 2023.

[16] A. Vassilev, A. Oprea, A. Fordyce, and H. Anderson, "Adversarial machine learning: A taxonomy and terminology of attacks and mitigations," NIST, Tech. Rep., 2024.

[17] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[18] N. Jagadeesha, "Facial privacy preservation using fgsm and universal perturbation attacks," in *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)*, vol. 1. IEEE, 2022.

[19] Y. M. Khedr, Y. Xiong, and K. He, "Semantic adversarial attacks on face recognition through significant attributes," *International Journal of Computational Intelligence Systems*, vol. 16, no. 196, 2023.

[20] H. Zolfi, J.-Y. Lin, S.-Y. Hsieh, H.-Y. Lin, and C.-L. Lin, "Adversarial patch attacks on deep-learning-based face recognition systems using generative adversarial networks," *Sensors*, vol. 23, no. 2, p. 853, 2023.

[21] X. Wang, F. Ye, Y. Li, and Y. Dai, "Biphasic face photo-sketch synthesis via face semantic-aware cyclegan," in *2023 IEEE 11th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*. IEEE, 2023, pp. 1172–1179.

[22] A. Sevastopolsky, Y. Malkov, N. Durasov, L. Verdoliva, and M. Niesner, "How to boost face recognition with stylegan?" in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2023, pp. 20 867–20 877.

[23] P. Srinivasan, A. M. K, M. Saraogi, J. Nataraju, G. Mishra, S. K. S, and S. A N, "Image inpainting for facial recognition using generative networks," in *2024 3rd International Conference for Innovation in Technology (INOCON)*, 2024, pp. 1–8.

[24] F. Vakhshiteh, A. Nickabadi, and R. Ramachandra, "Adversarial attacks against face recognition: A comprehensive study," *IEEE Access*, vol. 9, pp. 92 735–92 756, 2021.