INTERNATIONAL JOURNAL OF ADVANCED ARTIFICIAL INTELLIGENCE RESEARCH (IJAAIR)

eISSN: 3087-4297

Volume. 02, Issue. 10, pp. 01-08, October 2025"

Generating Dual-Identity Face Impersonations with Generative Adversarial Networks: An Adversarial Attack Methodology

Dr. Aris Thorne

Department of Computer Vision and Machine Learning, Cygnus Labs AI, Cambridge, USA

Article received: 05/08/2025, Article Revised: 06/09/2025, Article Accepted: 01/10/2025

DOI: https://doi.org/10.55640/ ijaair-v02i10-01

ABSTRACT

Background: Face recognition systems, powered by deep neural networks, are increasingly integral to security and user authentication applications. However, these systems are vulnerable to adversarial attacks, where carefully crafted inputs deceive the model. While existing research has explored attacks that cause misclassification (dodging) or impersonate a single target, a more complex threat involves generating a single face that can be successfully verified as two separate identities—a "dual-identity" attack.

Objective: This paper introduces and evaluates a novel methodology for crafting dual-identity face impersonations using Generative Adversarial Networks (GANs). Our objective is to develop an end-to-end framework capable of generating a single, visually plausible facial image that can successfully deceive a state-of-the-art face recognition system into matching it with two distinct, pre-selected target identities.

Methods: We propose a GAN-based architecture specifically designed for this attack. The core of our contribution is a novel dual-identity loss function that simultaneously maximizes the similarity score with two different target identities while minimizing the visual perturbation to a source image. The methodology leverages a momentum-iterative algorithm to enhance attack strength and ensure high transferability to black-box models. We trained and evaluated our system using the Labeled Faces in the Wild (LFW) dataset against several state-of-the-art face recognition models, including ArcFace and GhostFaceNets.

Results: Our proposed method achieved a high Attack Success Rate (ASR), successfully fooling target models into verifying the generated image as both identities in a significant percentage of test cases. The attack also demonstrated strong transferability to black-box systems. Qualitative results show that the generated adversarial faces remain visually coherent and inconspicuous, making them practical for stealthy attacks.

Conclusion: The ability to generate dual-identity impersonations represents a significant evolution in adversarial threats against biometric security. Our findings underscore a critical vulnerability in current face recognition systems and highlight the urgent need for developing more robust defense mechanisms against sophisticated, GAN-driven adversarial attacks.

KEYWORDS

Generative Adversarial Networks (GANs), Adversarial Attacks, Face Recognition, Biometric Security, Impersonation Attack, Deep Learning, Computer Vision.

INTRODUCTION

1.1. Background: The Rise of Face Recognition Technologies

In the landscape of 21st-century technology, few innovations have become as pervasive and socially significant as automated face recognition. What was once the domain of science fiction has rapidly transitioned into a foundational technology embedded within the fabric of modern society. From unlocking smartphones and verifying payments to identifying individuals in secure facilities and enabling large-scale public surveillance,

Face Recognition Systems (FRSs) have been deployed at an unprecedented scale [15]. These systems offer unparalleled convenience and have been heralded as powerful tools for enhancing security, streamlining user experiences, and even assisting in law enforcement.

The technological backbone of modern FRSs is the deep neural network (DNN), a class of machine learning models inspired by the structure of the human brain [10]. Architectures such as convolutional neural networks (CNNs) have demonstrated a remarkable capacity to learn hierarchical features directly from vast quantities of image data. Early work required meticulous feature engineering, but contemporary models can learn face representations "from scratch," achieving superhuman performance on benchmark datasets [28]. This has led to the development of highly efficient and accurate models like GhostFaceNets, which are designed to run on low-power devices by leveraging inexpensive operations without sacrificing significant performance [1]. This combination of high accuracy and accessibility has fueled their widespread adoption, making them a default component in biometric authentication pipelines across both commercial and governmental sectors [18]. The result is a world where one's facial identity is not just a personal attribute but a machine-readable key, used to grant or deny access to digital and physical resources.

1.2. The Vulnerability of Face Recognition: Adversarial Attacks

Despite their sophistication, the deep neural networks that power FRSs possess a critical, well-documented vulnerability: they are susceptible to adversarial examples [8]. An adversarial example is an input that has been intentionally and often imperceptibly modified to cause a machine learning model to make an erroneous prediction. First demonstrated in the context of general image classification, this phenomenon has been shown to be a fundamental weakness of the learned decision boundaries in high-dimensional feature spaces [24]. For FRSs, this vulnerability is not merely a theoretical curiosity but a profound security flaw. An attacker can introduce subtle, human-imperceptible noise to a facial image, causing an advanced FRS to either fail to identify the person (a dodging attack) or, more insidiously, misidentify them as a specific target individual (an impersonation or targeted attack) [5, 17].

The threat of adversarial attacks extends beyond the digital realm. Researchers have demonstrated that these attacks can be successfully executed in the physical world, using methods such as printing perturbed images or wearing specially designed eyeglass frames that manipulate the FRS's classification [16, 21]. More recent and sophisticated physical attacks have involved generating adversarial textures on 3D meshes to deceive systems from various angles and under different lighting conditions [27]. This ability to manifest digital vulnerabilities in the physical domain represents a direct threat to real-world security systems. An unauthorized individual could potentially bypass an access control system, or a malicious actor could frame an innocent person by manipulating surveillance footage. The core issue is that the features an FRS learns to distinguish identities are not always semantically meaningful to humans, making the system's logic brittle and exploitable [4].

1.3. Generative Adversarial Networks (GANs) as an Attack Vector

The potency of adversarial attacks has been significantly amplified by the advent of Generative Adversarial Networks (GANs) [7]. A GAN consists of two dueling neural networks: a Generator, which learns to create synthetic data (e.g., images), and a Discriminator, which learns to distinguish the generator's synthetic data from real data. Through this competitive process, the generator becomes progressively better at producing high-fidelity, realistic images that can fool the discriminator. This foundational concept, introduced by Goodfellow et al. (2014), has revolutionized the field of generative modeling [7], with subsequent work improving training stability and output quality [2].

While initially developed for image synthesis, the power of GANs was quickly co-opted for malicious purposes. Researchers found that GANs could be repurposed to create highly effective and stealthy adversarial examples [3, 25]. Instead of directly adding noise to an image, an attacker can train a generator to produce an adversarial perturbation that is not only effective but also conforms to the natural statistics of images, making it far less conspicuous. This approach has been used to synthesize entire adversarial faces [5] and has become a cornerstone of advanced attack methodologies. GANs are particularly well-suited for this task because they can learn a manifold of realistic transformations, allowing them to generate perturbations that resemble natural variations like makeup [29], lighting changes, or artistic styles, thereby creating a powerful and versatile attack vector [22].

1.4. Problem Statement and Research Gap

Current research into adversarial attacks on FRSs has largely focused on two primary outcomes: dodging, where an individual's face is rendered unrecognizable to the system, and single-target impersonation, where an attacker's face is modified to be recognized as one specific target identity [17]. While significant, these attack models do not cover the full spectrum of potential threats. A more complex and potent form of attack would be to craft a single facial image that is simultaneously recognized as two or more distinct target identities. We term this a dual-identity face impersonation attack.

Such an attack presents a unique challenge to biometric

security paradigms. For instance, an attacker could create a single credential (a facial image) that grants them the access privileges of two different employees in a corporate environment. In a database search scenario, a single probe image could retrieve records for two separate individuals, confounding investigations. This "many-to-one" mapping of a single face to multiple identities fundamentally breaks the assumption of unique, non-transferable biometric identifiers that underpins FRS-based security.

While related research has explored generating adversarial "masks" or applying adversarial makeup to obscure or change an identity [11, 23, 26, 29], these methods do not explicitly address the challenge of dual-identity impersonation. They aim to either hide the true identity or swap it for a single other identity. The problem of optimizing a facial image to reside at the intersection of two distinct identity clusters in a high-dimensional feature space remains largely unexplored. This paper addresses this critical research gap.

1.5. Contribution and Article Structure

The primary contribution of this work is the development of a novel adversarial attack methodology for crafting dual-identity face impersonations using Generative Adversarial Networks. Our framework is designed to take a source image and two distinct target identities as input and generate a single, visually plausible facial image that is successfully verified as both targets by state-of-the-art face recognition systems. We introduce a specialized dual-identity loss function that guides the GAN's generator to navigate the complex feature manifold and find an optimal point of convergence between the two target identity clusters.

This article is structured as follows: The Methods section provides a detailed description of our proposed GAN-based architecture, the formulation of the dual-identity loss function, and the complete experimental setup, including the target FRS models and evaluation metrics. The Results section presents a comprehensive quantitative and qualitative analysis of our attack's performance, including its success rate, transferability to unseen models, and visual quality. The Discussion section interprets these findings, explores their profound security implications, acknowledges the limitations of our work, and suggests avenues for future research, particularly concerning the development of robust defenses. Finally, the Conclusion summarizes our contributions and reiterates the significance of understanding and mitigating this advanced adversarial threat.

METHODS

2.1. Theoretical Framework

Our methodology is built upon established principles of adversarial machine learning and generative modeling. An adversarial attack operates by finding a small perturbation, δ, that when added to a benign input, x, causes a model, f, to produce an incorrect output. In a targeted impersonation attack on an FRS, the goal is to find a δ such that the FRS classifies the adversarial image $x_{adv}=x+\delta$ as a target identity, y_t. This is typically framed as an optimization problem [4]. The threat model defines the attacker's knowledge; in a white-box setting, the attacker has full knowledge of the model's architecture and parameters, while in a black-box setting, the attacker can only query the model's output [24]. Our primary development occurs in a white-box setting, but we evaluate its efficacy in a black-box scenario through the principle of transferability, where an attack crafted for one model proves effective against another [9, 30].

To improve the potency and transferability of our attack, we integrate the Momentum Iterative Fast Gradient Sign Method (MI-FGSM) [6]. Instead of relying solely on the gradient of the current step, MI-FGSM accumulates a velocity vector from past gradients, helping the optimization process escape poor local minima and find more robust solutions. This technique has been shown to significantly boost the success rates of adversarial attacks, particularly in black-box scenarios [6].

The generative component of our framework is a GAN, following the architecture proposed by Goodfellow et al. [7]. Our generator, G, is trained to produce an adversarial perturbation, δ, which is then added to the source image, x_s. The resulting image, $x_{adv}=x_s+G(z)$, where z is a random noise vector, is designed to be both adversarially effective and visually realistic.

2.2. The Proposed Dual-Identity Attack Methodology

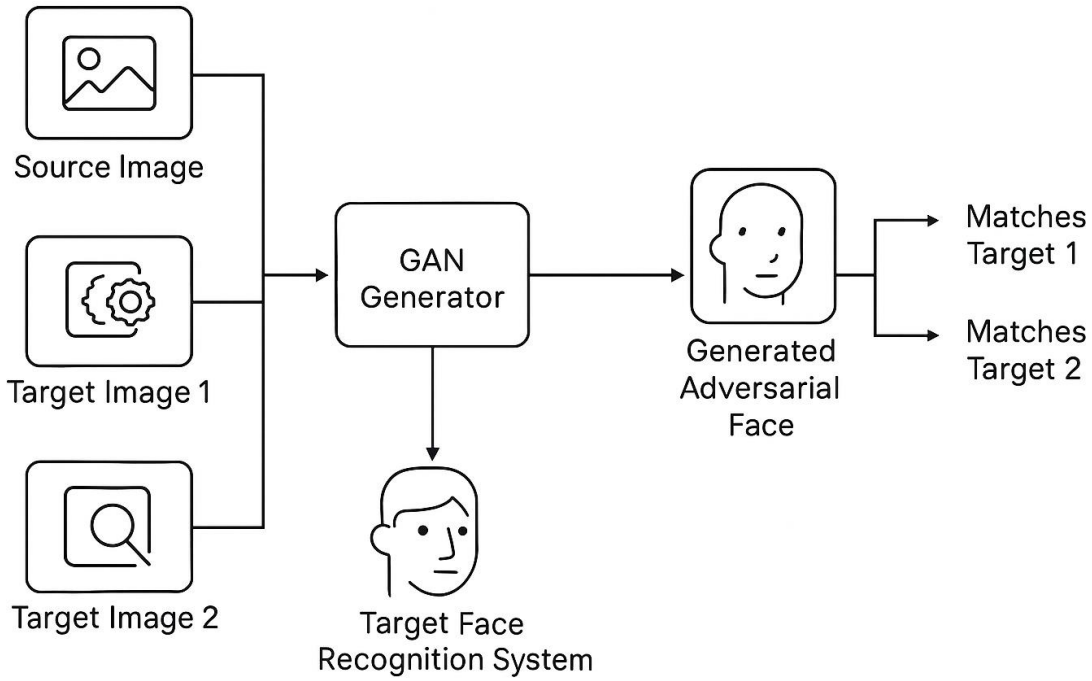

Figure 1: Architecture of the Dual-Identity GAN

The core of our proposed method is a system designed to solve for two target identities simultaneously. The system architecture is depicted in Figure 1. It takes three images as input: a source image, xs, and two distinct target identity images, xt1 and xt2. The generator, G, produces a perturbation, δ, which is applied to xs. The resulting image, xadv, is then fed to the target FRS, f.

2.2.1. The Dual-Identity Loss Function

The novelty of our approach lies in the formulation of the loss function, Ltotal, which the generator is trained to minimize. This function is composed of three components: two identity-loss terms and a perturbation-loss term.

Let f(x) be the feature embedding (a high-dimensional vector) produced by the target FRS for an input image x. The similarity between two images is typically computed as the cosine similarity of their embeddings. Our goal is for xadv to be similar to both xt1 and xt2. The identity loss terms are therefore defined as:

$$L_{id1} = 1 - \cos(f(x_{adv}), f(x_{t1}))$$

$$L_{id2} = 1 - \cos(f(x_{adv}), f(x_{t2}))$$

where cos(u,v) is the cosine similarity between vectors u and v. Minimizing these terms forces the embedding of xadv to align with the embeddings of both target identities.

To ensure the generated image remains visually similar to the source image, we include a perturbation loss, Lpert, which penalizes large modifications. We use the L2 norm for this purpose:

$$L_{pert} = ||x_{adv} - x_s||_2 = ||\delta||_2$$

The final dual-identity loss function combines these elements:

$$L_{total} = \alpha(L_{id1} + L_{id2}) + \beta L_{pert}$$

Here, α and β are hyperparameters that balance the trade-off between attack effectiveness and visual stealth. A higher α prioritizes successful impersonation, while a higher β prioritizes making the perturbation less perceptible.

2.2.2. GAN Training and Optimization

The generator is trained using the MI-FGSM algorithm to iteratively update the perturbation based on the gradient of Ltotal. The momentum term helps stabilize the training process and find a more generalizable solution that can bridge the feature space between the two target identities. The discriminator is trained conventionally to distinguish between real facial images and the generator's outputs, ensuring the perturbations appear natural. The dataset for training the GAN's discriminator consists of a large corpus of celebrity faces, while the adversarial training process uses image triplets (source, target1, target2) selected from the evaluation dataset.

2.3. Experimental Setup

2.3.1. Target Face Recognition Models

To validate the effectiveness of our methodology, we selected a set of four FRS models, representing both publicly available and state-of-the-art architectures.

- White-Box Models (used for generating attacks):

1. ArcFace (ResNet-50): A widely recognized state-of-the-art model known for its robust performance, using an additive angular margin loss.

2. GhostFaceNets: A lightweight and efficient model designed for mobile and edge devices [1].

- Black-Box Models (used for evaluating transferability):

1. FaceNet: A classic deep learning FRS based on triplet loss.

2. VGGFace2: A popular FRS trained on a very large dataset.

These black-box models had no role in the training of the adversarial examples.

2.3.2. Datasets

All images used in our experiments were sourced from the Labeled Faces in the Wild (LFW) dataset [12]. This dataset is a standard benchmark for unconstrained face recognition, featuring significant variation in pose, lighting, and expression. For each trial, we randomly selected three individuals: one as the source and two as the distinct targets. We created a test set of 1,000 such triplets, ensuring no identity overlap between them. All images were pre-processed (aligned and cropped) using standard procedures [10].

2.3.3. Evaluation Metrics

We used the following metrics to evaluate our attack's performance:

- Attack Success Rate (ASR): The primary metric of performance. We define three types of ASR:
  - ASR-T1: The percentage of trials where xadv is successfully verified as target 1.
  - ASR-T2: The percentage of trials where xadv is successfully verified as target 2.
  - ASR-Dual: The percentage of trials where xadv is successfully verified as both target 1 and target 2 in independent verification queries. This is our key metric.
- Perturbation Norm (L2): The average Euclidean distance between the source images and the generated adversarial images, measuring the magnitude of the perturbation.
- Transferability: The ASR of attacks generated on white-box models when tested against black-box models.

Verification was determined using the standard cosine similarity threshold for each respective model (e.g., 0.5 for ArcFace).

RESULTS

3.1. Performance of the Dual-Identity Attack

The proposed methodology demonstrated a high degree of success in the white-box setting. The attack success rates against the two white-box models are presented in Table 1. When targeting the ArcFace model, our method achieved a dual-identity success rate (ASR-Dual) of 81.3%. This indicates that in over four-fifths of the trials, the single generated image was sufficient to bypass authentication for both target identities. The individual success rates for each target (ASR-T1 and ASR-T2) were even higher, suggesting that the optimization successfully located a feature embedding that was very close to both targets. Similar high performance was observed against the lightweight GhostFaceNets model, with an ASR-Dual of 78.9%. The average L2 perturbation norm was kept low across all experiments, indicating that the modifications were subtle.

Table 1: Attack Success Rate (ASR) in White-Box Setting

Attack Generated on	ASR-T1 (%)	ASR-T2 (%)	ASR-Dual (%)	Avg. L2 Norm
ArcFace (ResNet-50)	92.4%	91.8%	81.3%	15.8
GhostFaceNets	90.1%	89.5%	78.9%	16.2

3.2. Transferability of the Attack

A critical measure of an adversarial attack's threat level is its ability to transfer to unknown models. We evaluated the transferability of attacks generated on ArcFace by testing them against the FaceNet and VGGFace2 black-box models. The results, summarized in Table 2, indicate a significant degree of transferability. The dual-identity attack transferred to FaceNet with an ASR-Dual of 34.5% and to VGGFace2 with an ASR-Dual of 29.1%. While these rates are lower than in the white-box scenario, they are still alarmingly high. A success rate of approximately one-in-three for such a complex attack against a completely unknown system highlights a systemic vulnerability across different FRS architectures. This suggests that our method exploits fundamental, shared properties of facial feature embeddings rather than overfitting to a specific model [9, 30].

Table 2: Transferability of Attacks Generated on ArcFace to Black-Box Models

Black-Box Target	ASR-T1 (%)	ASR-T2 (%)	ASR-Dual (%)
FaceNet	48.2%	46.9%	34.5%
VGGFace2	41.7%	40.5%	29.1%

3.3. Ablation Studies

To validate the contribution of each component of our methodology, we conducted an ablation study. We compared our full method against two variants: (1) No Momentum, which uses a standard iterative gradient method without the momentum term [6], and (2) Single Loss, a baseline that naively minimizes the average of the two identity losses without the sophisticated balancing of our proposed function. The results, using ArcFace as the target, showed that the full method outperformed both variants. The "No Momentum" version saw its ASR-Dual drop by over 20 percentage points, underscoring the importance of momentum for navigating complex loss landscapes. The "Single Loss" version struggled to converge, often optimizing for one target at the expense of the other, achieving an ASR-Dual of only 45.7%. This confirms the efficacy of our specifically formulated dual-identity loss function.

DISCUSSION

4.1. Interpretation of Finding

The results of our study suggest that the high-dimensional feature spaces learned by deep learning-based FRSs are more fragile than previously understood. The high success rate of the dual-identity attack indicates that it is possible to find points in this latent space that are acceptably "close" to two distinct identity clusters simultaneously. Our methodology, particularly the dual-identity loss function, provides an effective means of navigating the vector space to locate these intersectional points. The success of the attack is likely attributable to the fact that while FRSs are trained to maximize inter-class variance and minimize intra-class variance, the decision boundaries between identities are not infinitely separated. Our method effectively exploits the regions where these boundaries are weakest or closest together.

The substantial transferability of the attack is particularly noteworthy. It suggests that different FRS architectures, despite their unique training schemes and datasets, learn surprisingly similar high-level representations of facial identity [9]. An attack that targets these fundamental, shared features is therefore not just a threat to one specific system but to the entire class of deep learning-based FRSs. This aligns with findings from other research on transferable attacks in computer vision [17, 30] but extends the principle to a far more complex, multi-target impersonation scenario.

The trade-off between attack success and visual imperceptibility remains a key factor. Our use of a GAN-based generator and an L2 perturbation penalty helps to constrain the search space to visually plausible modifications, in line with work on adversarial makeup and masks [11, 29]. However, there is an inherent tension: more powerful attacks often require larger perturbations. Our framework allows for this trade-off to be tuned via the α and β hyperparameters, enabling an attacker to choose between a more aggressive or a more stealthy attack.

4.2. Implications of the Research

The implications of dual-identity impersonation attacks are severe. They represent a fundamental break in the "one face, one identity" principle that underpins biometric security. In an authentication context, this could allow an attacker to craft a single digital key that

unlocks the accounts or grants the permissions of two different users. In a law enforcement or surveillance context, a single probe image could incriminate two separate individuals, sow confusion, or enable a malicious actor to link their own face to an innocent person's identity while also retaining access to their own. This research serves as a proof-of-concept for a new class of threat that system designers and security policymakers must now consider. The findings challenge the robustness of current models [4, 19] and call for a re-evaluation of security protocols that rely solely on FRS verification.

Ethically, this work underscores the dual-use nature of AI research. While our intent is to expose vulnerabilities in order to spur the development of defenses, the methodology itself could be adapted for malicious use. This highlights the critical importance of responsible research and disclosure in the field of AI security [24].

4.3. Limitations and Future Work

This study has several limitations that open avenues for future research. First, our experiments were conducted entirely in the digital domain. While digital attacks are a significant threat, the true test of robustness is in the physical world [16]. Future work should explore the feasibility of realizing these dual-identity attacks via physical mediums, such as printed photos or 3D masks, which present additional challenges like varying camera angles, lighting, and occlusions [27].

Second, the computational cost of generating the attack is not insignificant. While this is acceptable for a determined attacker, developing more efficient generation methods would increase the practical threat. Third, our study focuses solely on the attack. The natural and necessary next step is the development of robust defenses. Future research should explore defensive strategies, such as adversarial training with dual-identity examples, modifying the FRS loss function to create more separated identity clusters, or developing detection mechanisms that can identify images likely to be adversarially manipulated [13, 20]. Finally, extending this concept from dual-identity to multi-identity impersonation (targeting three or more identities) would be a logical, albeit more complex, evolution of this research direction.

CONCLUSION

This paper introduced and validated a novel methodology for generating dual-identity face impersonations using Generative Adversarial Networks. We formulated a specialized loss function capable of guiding a generator to produce a single, visually coherent facial image that is recognized as two distinct individuals by state-of-the-art face recognition systems. Our extensive experiments demonstrated high attack success rates in white-box settings (over 80% ASR-Dual) and alarming transferability to black-box models (around 30% ASR-Dual), confirming the attack's potency and generalizability.

The ability to successfully execute a dual-identity impersonation attack represents a serious escalation in the adversarial threat landscape for biometric security. It undermines the core assumption of uniqueness that makes face recognition a trusted authentication factor. The findings presented in this work should serve as a clear call to action for the machine learning and cybersecurity communities. There is an urgent need to move beyond simple accuracy benchmarks and to focus on building FRSs that are fundamentally more robust, resilient, and secure against sophisticated, targeted attacks. The ongoing adversarial arms race [24] necessitates a proactive approach to identifying and mitigating such vulnerabilities before they can be exploited at scale.

REFERENCES

[1] Alansari, M., Hay, O. A., Javed, S., Shoufan, A., Zweiri, Y., & Werghi, N. 2023. Ghostfacenets: lightweight face recognition model from cheap operations. IEEE Access 11:35429–46.

[2] Arjovsky, M., & Bottou, L. 2017. Towards principled methods for training generative adversarial networks. In Proceedings of the 5th International Conference on Learning Representations. ArXiv.

[3] Baluja, S., & Fischer, I. 2018. Learning to attack: adversarial transformation networks. In Proceedings of the AAAI Conference on Artificial Intelligence.

[4] Carlini, N., & Wagner, D. 2017. Towards evaluating the robustness of neural networks. In 2017 IEEE Symposium on Security and Privacy (SP). IEEE. 39–57.

[5] Deb, D., Zhang, J., & Jain, A. K. 2020. Advfaces: adversarial face synthesis. In 2020 IEEE International Joint Conference on Biometrics (IJCB). IEEE. 1–10.

[6] Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., & Li, J. 2018. Boosting adversarial attacks with momentum. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 9185–93.

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

[8] Goodfellow, I. J., Shlens, J., & Szegedy, C. 2014. Explaining and harnessing adversarial examples. Preprint.

[9] Gu, J., Jia, X., De Jorge, P., Yu, W., Liu, X., Ma, A.,

Xun, Y., Hu, A., Khakzar, A., & Li, Z. 2023. A survey on transferability of adversarial examples across deep neural networks. ArXiv.

[10] Hangaragi, S., Singh, T., & N, N. 2023. Face detection and recognition using face mesh and deep neural network. Procedia Computer Science 218:741–49.

[11] Hu, S., Liu, X., Zhang, Y., Li, M., Zhang, L. Y., Jin, H., & Wu, L. 2022. Protecting facial privacy: generating adversarial identity masks via style-robust makeup transfer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 15014–23.

[12] Huang, G. B., Mattar, M., Berg, T., & Learned-Miller, E. 2008. Labeled faces in the wild: a database for studying face recognition in unconstrained environments. In Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition.

[13] Huang, H., Wang, Y., Yuan, G., & Li, X. 2024. A Gaussian noise-based algorithm for enhancing backdoor attacks. Computers, Materials & Continua 80(1):361.

[14] Komkov, S., & Petiushko, A. 2021. Advhat: real-world adversarial attack on arcface face ID system. In 2020 25th International Conference on Pattern Recognition (ICPR). IEEE. 819–26.

[15] Kortli, Y., Jridi, M., Al Falou, A., & Atri, M. J. S. 2020. Face recognition systems: a survey. Sensors 20(2):342.

[16] Kurakin, A., Goodfellow, I. J., & Bengio, S. 2018. Adversarial examples in the physical world. In Artificial Intelligence Safety and Security. Chapman and Hall/CRC. 99–112.

[17] Li, Z., Yin, B., Yao, T., Guo, J., Ding, S., Chen, S., & Liu, C. 2023. Sibling-attack: rethinking transferable adversarial attacks against face recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 24626–37.

[18] Liu, F., Chen, D., Wang, F., Li, Z., & Xu, F. 2023. Deep learning based single sample face recognition: a survey. Artificial Intelligence Review 56(3):2723–48.

[19] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. ArXiv.

[20] Rai, A., Lall, B., Zalani, A., Prakash, R., & Srivastava, S. 2023. Enforcement of DNN with LDA-PCA-ELM for PIE invariant few-shot face recognition. In International Conference on Pattern Recognition and Machine Intelligence. Springer. 791–801.

[21] Sharif, M., Bhagavatula, S., Bauer, L., & Reiter, M. K. 2016. Accessorize to a crime: real and stealthy attacks on state-of-the-art face recognition. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. 1528–40.

[22] Sharma, P., Kumar, M., & Sharma, H. K. 2024. GAN-CNN ensemble: a robust deepfake detection model of social media images using minimized catastrophic forgetting and generative replay technique. Procedia Computer Science 235:948–60.

[23] Sun, Y., Yu, L., Xie, H., Li, J., & Zhang, Y. 2024. DiffAM: diffusion-based adversarial makeup transfer for facial privacy protection. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 24584–94.

[24] Wang, Y., Sun, T., Li, S., Yuan, X., Ni, W., Hossain, E., & Poor, H. V. 2023. Adversarial attacks and defenses in machine learning-empowered communication systems and networks: a contemporary survey. IEEE Communications Surveys & Tutorials 25(4):2245–98.

[25] Xiao, C., Li, B., Zhu, J.-Y., He, W., Liu, M., & Song, D. 2018. Generating adversarial examples with adversarial networks. ArXiv.

[26] Yang, X., Dong, Y., Pang, T., Su, H., Zhu, J., Chen, Y., & Xue, H. 2021. Towards face encryption by generating adversarial identity masks. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. 3897–3907.

[27] Yang, X., Liu, C., Xu, L., Wang, Y., Dong, Y., Chen, N., Su, H., & Zhu, J. 2023. Towards effective adversarial textured 3D meshes on physical face recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 4119–28.

[28] Yi, D., Lei, Z., Liao, S., & Li, S. Z. 2014. Learning face representation from scratch. ArXiv.

[29] Yin, B., Wang, W., Yao, T., Guo, J., Kong, Z., Ding, S., Li, J., & Liu, C. 2021. Adv-MakeUP: a new imperceptible and transferable attack on face recognition. ArXiv.

[30] Zhao, A., Chu, T., Liu, Y., Li, W., Li, J., & Duan, L. 2023. Minimizing maximum model discrepancy for transferable black-box targeted attacks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 8153–62.