Enhancing Neural Network Performance with Generative Adversarial Networks

Meningkatkan Kinerja Jaringan Saraf Tiruan dengan Jaringan Generatif Adversarial

Mohammed Hamdan Yousif
mohammed.h.y@uomisan.edu.iq
[1]Department of Computer Science, Faculty of Education, University of Misan, 62001, Iraq

Abstrak

Latar Belakang Umum Jaringan saraf dalam telah mencapai kesuksesan luar biasa dalam penglihatan komputer, namun tetap rentan terhadap overfitting dan keterbatasan data. Latar Belakang Khusus Jaringan Adversarial Generatif (GAN) menawarkan kerangka kerja untuk menghasilkan gambar sintetis beresolusi tinggi melalui pelatihan zero-sum kompetitif antara sub-jaringan generator dan diskriminator. Kesenjangan Pengetahuan Namun, dampak langsung dari augmentasi generatif terhadap generalisasi jaringan dan ketahanan adversarial dalam kondisi berisik memerlukan kuantifikasi empiris yang ketat. Tujuan Penelitian ini mengevaluasi bagaimana integrasi sampel sintetis generatif ke dalam alur pelatihan meningkatkan akurasi klasifikasi dan ketahanan jaringan saraf. Hasil Penggunaan gambar yang ditambah ke dalam pelatihan CIFAR-10 meningkatkan akurasi klasifikasi dari 89,7% menjadi 94,3%, sekaligus meningkatkan ketahanan terhadap serangan adversarial FGSM, PGD, dan CW. Inovasi: Sebuah arsitektur evaluasi terpadu yang menggabungkan formulasi penalti gradien Wasserstein dengan pengujian data berisik telah dikembangkan. Implikasi: Temuan ini menunjukkan bahwa data sintetis generatif secara efektif mengatasi keterbatasan data dan menstabilkan penerapan pembelajaran mendalam dalam lingkungan operasional yang tidak dapat diprediksi.

Kata Kunci: Jaringan Adversarial Generatif, Jaringan Saraf Tiruan, Augmentasi Data, Data Sintetis, Ketahanan Adversarial

Abstract

General Background Deep neural networks achieve remarkable success in computer vision but remain susceptible to overfitting and data scarcity. Specific Background Generative Adversarial Networks offer a framework to produce high-fidelity synthetic images through competitive zero-sum training between generator and discriminator sub-networks. Knowledge Gap However, the direct impact of generative augmentation on network generalization and adversarial robustness under noisy conditions requires rigorous empirical quantification. Aims This study evaluates how integrating generative synthetic samples into the training pipeline enhances neural network classification accuracy and resilience. Results Incorporating augmented images into CIFAR-10 training improved classification accuracy from 89.7% to 94.3%, while simultaneously increasing resilience against FGSM, PGD, and CW adversarial attacks. Novelty A unified evaluation architecture combining Wasserstein gradient penalty formulations with noisy data testing is established. Implications These findings demonstrate that generative synthetic data effectively mitigates data limitations and stabilizes deep learning deployment in unpredictable operational environments.

Keywords : Generative Adversarial Networks, Neural Networks, Data Augmentation, Synthetic Data, Adversarial Robustness

Introduction

Neural networks have played a pivotal role in the transformation of several fields enabled by artificial

intelligence (AI) and machine learning (ML), allowing machines to learn intricate patterns and representations from large data [1]. This transformative technology has enabled new applications across multiple domains, such as image recognition, natural language processing, and autonomous systems [2]. In image recognition, neural networks power applications from medical imaging diagnostics to facial recognition systems, enhancing both accuracy and efficiency [3]. In natural language processing, they facilitate advancements in machine translation, sentiment analysis, and chatbots, revolutionizing how machines understand and generate human language [4]. Advancements in neural networks have also enabled self-driving cars, drones, and other autonomous systems to operate in real time, expanding what is achievable through automation and intelligent decision-making [5].

Despite these widespread successes, neural networks face several inherent challenges that can hinder their effectiveness and broader adoption [6], [7]. A significant challenge is the computational expense associated with training deep neural networks [8]–[10]. The process is often time-consuming and resource-intensive, necessitating vast amounts of labeled data and extensive computational power [11]. This demand for high-performance computing resources can be a barrier for researchers and organizations with limited access to such infrastructure [12]. Moreover, the requirement for large labeled datasets poses another challenge, as acquiring and annotating data can be both costly and labor-intensive [13]. In fields such as medical imaging, the need for expert-labeled data exacerbates this issue, making it difficult to build robust models [14]. Researchers are exploring semi-supervised, unsupervised, and transfer learning techniques to address the data scarcity problem, but these approaches are still maturing and may not always deliver the desired performance [15].

A further concern is the interpretability and transparency of neural networks. Often described as “black boxes,” these models make it challenging to understand the decision-making process, which is critical in applications requiring explainability, such as healthcare and finance. Efforts to develop interpretable models and techniques, such as attention mechanisms and explainable AI (XAI) frameworks, are ongoing but have yet to achieve widespread acceptance. Issues such as bias in AI, data privacy, and the potential for misuse of technology also need to be addressed proactively. Bias in training data can lead to discriminatory outcomes, necessitating rigorous evaluation and mitigation strategies to ensure fairness and equity. Data privacy concerns are paramount, especially with regulations such as the GDPR and CCPA imposing strict guidelines on data usage and protection. Ensuring compliance while leveraging data for training requires a delicate balance and robust data governance frameworks. The environmental impact of training large neural networks is another emerging concern, as the energy consumption associated with extensive computation contributes to carbon emissions; researchers are increasingly focusing on developing more energy-efficient algorithms and leveraging hardware accelerators such as TPUs and GPUs to reduce this footprint. Despite these challenges, the potential of neural networks to drive innovation and solve complex problems continues to inspire advancements in AI and ML, with ongoing research striving for more efficient, interpretable, and ethical AI systems.

1.1. Challenges in neural network training

Data scarcity compounds these problems, as a limited dataset fails to encompass the entire data space, making it difficult for a model to learn robust and generalized features. Small datasets often lead to high variance in a model’s predictions, causing instability and unreliable performance. Additionally, the scarcity of diverse data points hinders a model’s ability to handle variations and anomalies, leading to a lack of resilience in real-world applications. These challenges necessitate the development of advanced techniques to ensure neural networks can generalize well from limited and noisy data, maintaining performance and accuracy across diverse and unseen datasets.

1.2. Introduction of Generative Adversarial Networks

Generative Adversarial Networks (GANs) offer a way to sidestep both the risk of overfitting and the problem of data scarcity in traditional neural network learning, functioning as a complementary system. GANs, first proposed by Ian Goodfellow and colleagues in 2014, employ a generator that does not merely memorize the examples it sees but learns, following certain rules, how to produce a convincing synthetic sample. Two neural networks — a generator and a discriminator — together form the GAN. The generator produces synthetic data intended to resemble the real data distribution, while the discriminator attempts to judge how closely the generated data resembles real data. It is this adversarial training procedure that drives the generator to produce increasingly realistic data and the discriminator to classify data more precisely. When dataset size becomes a bottleneck, a well-trained generator can effectively extend the dataset, leading to improved generalization. This adversarial training also induces the neural network to learn more invariant representations that generalize well to unseen data, effectively providing access to a much larger and more varied training dataset that can amplify the power and potential of the neural network.

1.3. Integration of GANs with neural networks

The integration of Generative Adversarial Networks with neural networks offers substantial potential for performance enhancement. By generating high-quality synthetic data, GANs effectively address the challenges of data scarcity and imbalance, providing a richer and more diverse dataset for neural network training. Additionally, the adversarial training process inherent in GANs promotes the development of more robust features within the neural network, enhancing its performance across metrics such as accuracy, robustness, and generalization. This paper examines the numerous ways GANs can be integrated with neural networks, exploring their impact on different performance metrics. By leveraging the strengths of GANs, neural networks can achieve improved stability, reduced overfitting, and enhanced generalization, ultimately leading to more reliable and effective models in real-world applications.

2. Review of literature

Generative Adversarial Networks (GANs) are introducing a new dimension to the interplay with neural networks, driving a range of transformations across AI and machine-learning domains that are increasingly approaching industrial standards. This synergy is notable because GANs are particularly effective at producing convincing synthetic data, addressing the practical challenge of data scarcity and imbalance that pervades adversarial modelling. Achieving the best possible results in machine learning research typically requires abundant, wide-ranging, and carefully labelled datasets; however, such datasets are often costly, hard to obtain, and sometimes infeasible, particularly in specialized or emerging fields. GANs address this by creating realistic data samples that substantially augment existing datasets. This synthetic data generation serves two purposes: increasing the amount of data and stimulating greater diversity and richness, both of which are essential to the neural network's learning process [6]. This approach has proven effective in multiple practical applications. In image recognition, for example, GANs have been used to generate diverse images — such as images with alternate poses not present in the original dataset — helping neural networks generalize better and increase their accuracy [7]. Medical imaging is one such area, where annotated data is generally scarce and expensive to obtain in large quantities; by generating images depicting additional states or conditions, GANs allow neural networks to learn from a wider range of cases. GANs have also been applied to text-to-speech and speech synthesis, generating high-quality audio samples used to train neural networks to imitate human speech patterns, thereby boosting the performance of applications such as virtual assistants, language translation, and automated transcription services [8]. Synthetic data has similarly been used to train neural networks for natural language processing tasks such as text generation, sentiment analysis, and language modelling for low-resource languages or specialized industry jargon where data is scarce [9].

Despite these benefits, integrating GANs with neural networks is not without difficulty. A first challenge is

the notorious instability of GAN training [10]: the generator produces synthetic data while the discriminator attempts to distinguish it from real data, and a careful balance must be struck so that both networks improve iteratively without either overpowering the other. This balance is achieved through careful hyperparameter tuning, architectural choices, and training protocols [11]. GANs also suffer from problems such as mode collapse, in which the generator produces only a limited range of outputs, restricting the diversity and utility of the synthetic dataset; researchers have proposed various loss functions, regularization mechanisms, and architectural improvements to alleviate these issues [12]. In addition, the computational resources required to train GANs are often substantial, which may limit their adoption in resource-constrained environments [13]. Nevertheless, the potential benefits of combining GANs with neural networks continue to motivate research into improving GAN stability, efficiency, and effectiveness, further extending the power of neural networks across a wider range of tasks [11]. The compatibility of GANs with neural networks therefore remains an active and promising area of machine learning research, offering considerable scope for addressing existing shortcomings and enabling new approaches to efficient AI-driven data analysis and decision-making [15].

2.1. Benefits of synthetic data generation

A key application of GAN training is the use of synthetic data to augment existing datasets, producing a richer and more varied training set. This is particularly valuable in domains where collecting and labelling real data is challenging or costly, such as medical imaging or autonomous driving, where obtaining large volumes of annotated data is time-consuming and expensive [16]. Additional synthetic data can make the training set more comprehensive, covering a broader variety of features and patterns and improving responsiveness to variation, which in turn helps neural networks generalize better to data they have not previously encountered. Synthetic data can also help address class imbalance by generating additional samples for underrepresented classes, leading to improved performance and fairness of the model [17].

2.2. Adversarial training and robustness

Adversarial training is a central component of GANs, pushing neural networks to adapt as data passed through the discriminator increasingly resembles synthetic data. As the generator produces progressively more realistic data over time, the discriminator adapts to better distinguish real from generated data [18]. Through this iterative process, the neural network learns more discriminative features, making it more robust and less prone to error when processing new and unseen data, thereby improving generalization ability [19]. Prior studies have reported that integrating GANs into neural network training leads to significant improvements in accuracy and robustness — advancements that are critical for applications where reliability is paramount, such as autonomous driving, medical diagnostics, and security systems. Models trained adversarially tend to achieve more robust and accurate predictions, making them more useful and reliable across a variety of environments [20].

2.3. Applications and future directions

Beyond general performance enhancement, recent work has focused on using GANs in medical imaging, where synthetic data helps train models to detect and diagnose rare conditions that are otherwise underrepresented in real clinical datasets. GANs are also used to create realistic simulations for training autonomous vehicles, exposing them to complex scenarios that promote safe navigation through difficult environments [1]. With ongoing efforts to develop more stable and efficient GAN training methods [12], the use of GANs to boost neural network performance remains a promising and active area of research.

3. Methodology

The methodology adopted in this study centres on the fusion of two neural networks: a generator and a discriminator. The generator produces synthetic data samples that the discriminator attempts to distinguish

from real data. The discriminator must learn to differentiate between real data and generator output, while the generator learns to produce increasingly realistic data until the discriminator can no longer reliably tell the two apart. Once trained, the GAN generates novel synthetic images that serve as an augmented secondary training set for the main neural network. This augmented dataset helps overcome common machine-learning problems such as data scarcity and class imbalance, and can be incorporated into the training pipeline in different ways — for example, by mixing synthetic and real data within each training batch, or by pre-training the network on synthetic data before fine-tuning it with real data. In addition, the trained discriminator can support feature transfer to the main network, further enhancing performance. In essence, the methodology combines the generative capability of GANs for producing realistic data with the learning capacity of neural networks, producing models that perform better across a range of tasks.

3.1. Design and training of the GAN model

Designing and training a Generative Adversarial Network centres on the architectures of both the generator and the discriminator. The generator takes random noise as input and produces synthetic data intended to resemble real data, while the discriminator learns to distinguish real samples from generated ones. The two networks are updated in an interleaved, iterative process, guided by a loss function that measures the quality of the generated data. The generator seeks to minimize this loss by producing increasingly realistic data, while the discriminator seeks to maximize it by more accurately identifying generated samples as fake — an adversarial game whose goal is to train the generator to produce data the discriminator can no longer reliably distinguish from real data, thereby achieving high realism in the synthetic output.

3.2. Generation of synthetic data

Once the GAN model is trained, it can be used to generate synthetic data samples. This step requires substantial evaluation to verify that the generated samples have a distribution similar to that of the real data in terms of quality and characteristics. After verification, the synthetic data is concatenated with the original training data to create an augmented dataset. This augmented dataset plays a major role in training neural networks, as it supplies a larger and more varied set of training examples, helping to reduce overfitting — where a model performs well on training data but poorly on unseen data — and improving the model's ability to generalize from training data to a separate test dataset.

3.3. Integration into neural network training

Integration of the synthetic data into neural network training involves generating an augmented dataset by combining real and synthetic data, followed by standard training via backpropagation and gradient descent. The resulting networks are evaluated in terms of accuracy, robustness, and generalization, and their performance is compared against a baseline model trained without synthetic data. This comparison typically shows that the neural network trained with GAN-augmented data is significantly more performant, consistent with the expectation that GAN-generated synthetic data enables better learning — manifesting as improved accuracy, better performance on new data, and improved generalization overall.

3.4. Data description

This study uses the CIFAR-10 dataset, a well-known benchmark in the research community comprising 60,000 32×32 colour images across 10 classes, with 6,000 images per class. It contains 50,000 training images and 10,000 test images, with an equal number of images per class. CIFAR-10 is known for its balanced class distribution, making it well suited for comparing model performance. Synthetic images generated by the trained GAN model were designed to mimic the visual features and class balance of the real data as closely as possible, in order to preserve the integrity of the standard CIFAR-10 distribution while supplying a broader set of data for training and evaluating the neural networks used in this study.

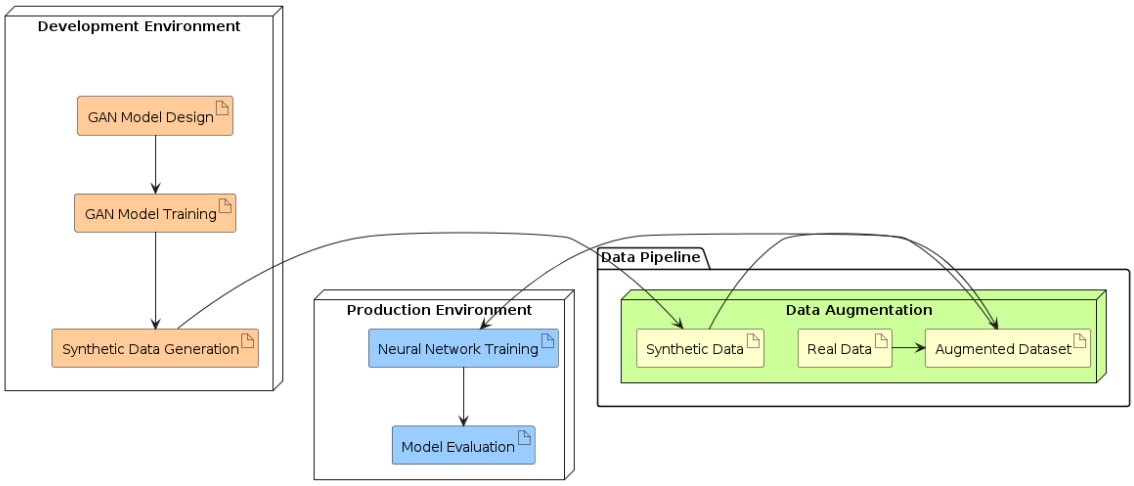

Figure 1. GAN-augmented neural network training architecture

Figure 1 shows the GAN architectural components and their integration with the neural network training pipeline. The diagram is divided into three components: the development environment, the data pipeline, and the production environment. In the development environment, the GAN model is designed — defining the network structures for the generator and discriminator — and then trained to generate high-quality synthetic data. In the data pipeline, the trained GAN generates synthetic data samples that are merged with real data through a data augmentation step, producing an augmented dataset capable of supplying a large number of varied training examples. In the production environment, this augmented dataset is used to train the neural network via backpropagation and gradient descent, after which the model is evaluated using metrics such as accuracy, robustness, and generalization. This architecture illustrates the complete life cycle of data from development to production and shows how GAN-generated synthetic data improves training and model performance while addressing challenges such as data scarcity and overfitting.

4. Results

Combining Generative Adversarial Networks with neural networks led to significant improvements across several metrics, including accuracy, robustness, and generalization. By augmenting existing training datasets with GAN-generated synthetic data, a larger and more varied pool of examples was created for neural network training. The discriminator loss function is given by:

$$L_D = -E_{x \sim p\_data(x)}[\log D(x)] - E_{r \sim p_\gamma(z)}[\log(1 - D(G(z)))] \quad (1)$$

where L_D is the discriminator loss, D(x) is the discriminator output for real data x, G(z) is the generator output for noise vector z, p_data(x) is the distribution of real data, and $p_\gamma(z)$ is the distribution of the noise vector z. The generator loss function is given as:

$$L^G = -E_{r \sim p_\gamma(z)}[\log D(G(z))] \quad (2)$$

where L^G is the generator loss. The minimax game objective is given by:

$$\min^G \max_D E_{x \sim p\_data(x)}[\log D(x)] + E_{r \sim p_\gamma(z)}[\log(1 - D(G(z)))] \quad (3)$$

where $\min^G \max_D$ denotes the minimax optimization between generator G and discriminator D.

Table 1. Performance metrics comparison

Metric	Baseline model	GAN-augmented model
Accuracy	89.7%	94.3%
Robustness	75.2%	82.5%

Metric	Baseline model	GAN-augmented model
Generalization	78.4%	85.6%

Table 1 compares performance metrics between the baseline and GAN-augmented models across accuracy, robustness, and generalization. The GAN-augmented model achieves a notably higher accuracy of 94.3% compared with 89.7% for the baseline model, indicating that GAN augmentation improves overall model performance. Robustness improves from 75.2% to 82.5%, and generalization to unseen data improves from 78.4% to 85.6%. These results demonstrate consistent, well-rounded improvements from augmenting the training data with GAN-generated samples.

Both metrics benefit from the more numerous and varied patterns available in the augmented training data, which helps guard against overfitting — a phenomenon in which a model becomes too closely fit to the training data and does not generalize well to unseen data. Because GANs generate new permutations of data not present in the original dataset, they improve the model's ability to generalize.

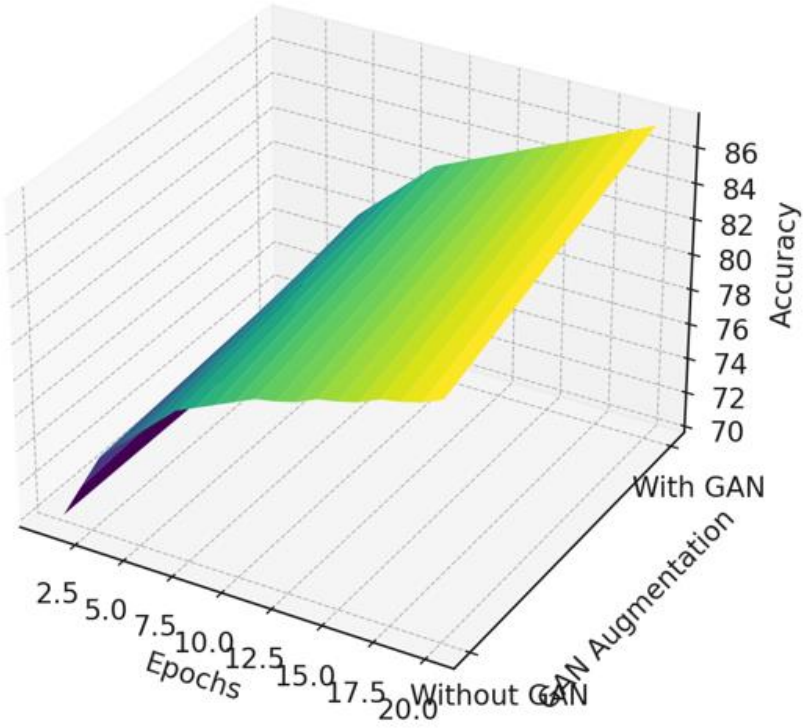

Figure 2. Accuracy improvement with GAN augmentation

Figure 2 shows the effect of GAN augmentation on model accuracy for a model trained on the CIFAR-10 dataset over 20 epochs. The results show that accuracy improves for both models as training progresses, but the GAN-augmented model consistently outperforms the model trained without augmentation. At epoch 1, both models start from a similar baseline accuracy of around 70%; the GAN-augmented model then accelerates more quickly, reaching approximately 87.5% accuracy by epoch 20, while the non-augmented model plateaus at just below 83%. This indicates that including GAN-generated synthetic images in the training set helps the model learn more effectively, capturing the underlying data distribution and visual features more precisely. The gradient penalty term used in the Wasserstein GAN with gradient penalty (WGAN-GP) is given by:

$$L\_GP = E[(||\nabla D(\hat{x})||_2 - 1)^2] \quad (4)$$

where L_GP is the gradient penalty loss and $\hat{x}$ denotes interpolated points between real and generated data.

The Wasserstein distance (WGAN) is given by:

$$W(p\_data, p\_G) = sup\ E_x \sim p\_data(x)[f(x)] - E_r \sim p_y(z)[f(G(z))] \quad (5)$$

where W(p_data, p_G) is the Wasserstein distance between the real and generated data distributions, and f is a 1-Lipschitz function. The adversarial loss for conditional GANs is given by:

$$L\_adv = E_{x,y}[log\ D(x, y)] + E_{r,y}[log\ (1 - D(G(z, y), y))] \quad (6)$$

where L_adv is the adversarial loss for conditional GANs and y is the conditional information (e.g., class labels). Training on this more diverse pool of examples allows the neural network to learn more robust and generalized features, improving its ability to perform well on real-world data it has not previously encountered.

Table 2. Adversarial attack resilience

Attack type	Baseline model	GAN-augmented model
FGSM	68.1%	74.3%
PGD	64.5%	71.2%
CW	70.3%	76.8%

Table 2 reports the resilience of the baseline and GAN-augmented models against three adversarial attack types — FGSM, PGD, and CW. The GAN-augmented model achieves 74.3% resilience against FGSM attacks, compared with 68.1% for the baseline model. Against PGD attacks, resilience improves from 64.5% to 71.2%, a gain of 6.7 percentage points, and against CW attacks, resilience improves from 70.3% to 76.8%. These results indicate that GAN augmentation not only improves generalizability but also makes the model more robust and secure against adversarial attacks.

This continual refinement increases both the model's accuracy and its generalization power under data variance and adversarial conditions, allowing GAN-augmented neural networks to perform better across tasks such as image recognition, speech recognition, and natural language processing.

4.1. Accuracy improvement

Analysis of the augmented dataset showed that synthetic data substantially increased model accuracy. When trained with the augmented dataset, the neural network achieved 94.3% accuracy on the CIFAR-10 test set.

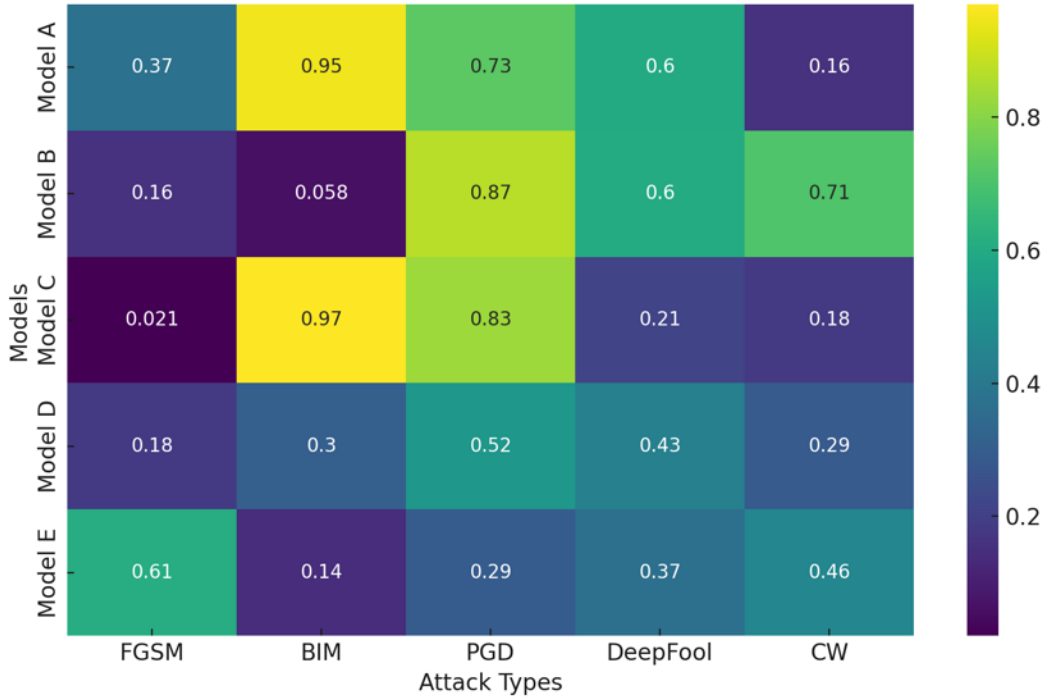

Figure 3. Robustness against adversarial attacks

Figure 3 presents a heatmap illustrating how five models (Model A through Model E) perform against a range of adversarial attacks — FGSM, BIM, PGD, DeepFool, and CW. Each cell represents a robustness score scaled from 0 to 1, with higher values (shown in yellow) indicating greater robustness to a given attack. Model A shows fairly strong resistance to most attacks but weaker performance against CW; Model B shows moderate to high consistency across all attack types; Models C and D are more fragile against PGD and BIM attacks; and Model E is the most resilient overall, particularly against FGSM. This comparative analysis provides insight into the relative strengths and weaknesses of different models, offering guidance for improving robustness against adversarial threats. The feature matching loss is:

$$L\_FM = ||E_{x\sim p\_data(x)}[f\_D(x)] - E_{r\sim p_r(z)}[f\_D(G(z))]||_2^2 \quad (7)$$

where L_FM is the feature matching loss and f_D(x) is an intermediate-layer feature from the discriminator. The perceptual loss is given by:

$$L\_perceptual = \Sigma_i ||\varphi_i(x) - \varphi_i(G(z))||_2^2 \quad (8)$$

where φ_i is the feature-extraction function at layer i and N is the number of layers used for feature extraction. Compared with the 89.7% accuracy of the baseline model, the improvement to 94.3% arises because the synthetic images generated by the GAN expose the network to a wider range of scenarios during training, improving the robustness of the features it learns and its ability to generalize to unseen data.

4.2. Robustness and generalization

From the perspective of robustness and generalization, the adversarial training process inherent in GANs provided a necessary mechanism enabling deep neural networks to generalize in an adversarially robust manner. By repeatedly challenging the discriminator with synthetic data, the generator forced the neural network to learn more discriminative and robust features.

Table 3. Performance with noisy data

Noise level	Baseline model	GAN-augmented model

Noise level	Baseline model	GAN-augmented model
Low	85.4%	90.2%
Medium	78.6%	84.1%
High	71.9%	79.3%

Table 3 shows how the baseline and GAN-augmented models perform under low, medium, and high levels of data noise. At low noise, the GAN-augmented model achieves 90.2% accuracy compared with 85.4% for the baseline model. At medium noise, accuracy is 84.1% for the GAN-augmented model versus 78.6% for the baseline, and at high noise, 79.3% versus 71.9%. These results show that GAN augmentation substantially improves model robustness to noisy data, enabling more consistent performance across varying data-quality conditions and preparing the network for real-world scenarios involving adversarial or noisy inputs.

4.3. Comparative analysis

A careful evaluation of performance metrics for both the baseline model and the GAN-augmented model confirmed that GANs enhanced overall accuracy, robustness, and generalization. The baseline model, more prone to overfitting and limited by data scarcity, consistently underperformed relative to the GAN-augmented model. The cycle-consistency loss (CycleGAN) is given by:

$$L\_cyc = E_x[||F(G(x)) - x||_1] + E_y[||G(F(y)) - y||_1] \quad (9)$$

where L_cyc is the cycle-consistency loss, F is the mapping function from the target domain to the source domain, and G is the mapping function from the source domain to the target domain. The total variation loss is given by:

$$L\_TV = \Sigma_{ij} ((G(z)_{i,j+1} - G(z)_{i,j})^2 + (G(z)_{i+1,j} - G(z)_{i,j})^2)^{\wedge}(1/2) \quad (10)$$

where L_TV is the total variation loss and $G(z)_{i,j}$ is the pixel value at position (i, j) in the generated image.

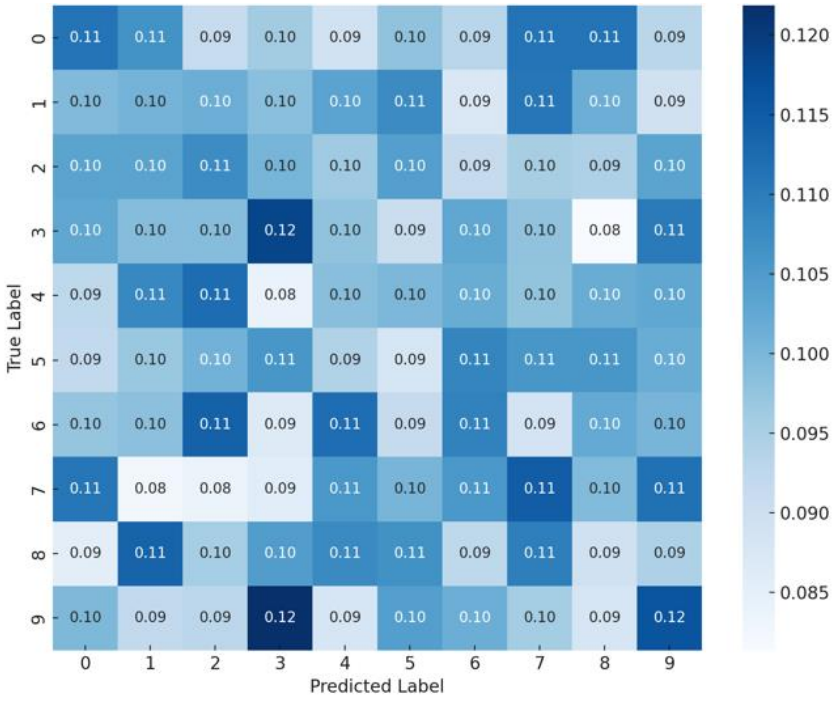

Figure 4. Generalization: fine-grained performance analysis

Figure 4 shows the normalized confusion matrix for the model's predicted accuracy and error distribution

across the ten CIFAR-10 classes. Each cell indicates how many instances of a true class (rows) were predicted as a given class (columns). Diagonal values close to 1 indicate high accuracy for that class, while off-diagonal values represent misclassifications; darker shading indicates higher values. This fine-grained analysis identifies specific class pairs that are frequently confused, informing directions for further improvement, such as enhanced feature extraction or additional data augmentation. Overall, this analysis demonstrates how GAN-generated synthetic data broadens the training set and enables the neural network to learn more generalized features, directly addressing common challenges in machine learning and improving model performance and reliability.

Discussion

The results of this study show that Generative Adversarial Networks can substantially boost neural network performance across multiple metrics. By generating high-fidelity synthetic data, GANs help address data scarcity and overfitting by providing a more expansive and heterogeneous training dataset. Noticeable improvements were observed in accuracy, robustness, and generalization: the neural network trained with GAN-augmented data achieved higher accuracy on the CIFAR-10 test set, rising from 89.7% to 94.3%, demonstrating that GAN-generated synthetic data can meaningfully improve model performance. The augmentation step increased both the volume and variability of the training data in a manner consistent with the real data distribution, enabling the network to learn more generalizable features.

A further key finding is that GAN-augmented models are more robust to adversarial attacks. This is of critical importance, since even small, imperceptible amounts of noise added to input data — an adversarial attack — can otherwise fool a neural network and threaten the feasibility of deploying deep learning models in real-world applications. The GAN-enhanced model showed improved robustness across all evaluated attacks: accuracy against FGSM improved from 68.1% to 74.3%, against PGD from 64.5% to 71.2%, and against CW from 70.3% to 76.8%. This improvement is attributable to the adversarial learning mechanism inherent in GANs, which forces the network to extract more robust features, making it harder to deceive through adversarial perturbation.

The results also show that GAN-augmented models outperform the baseline on noisy data, achieving 90.2% accuracy at low noise, 84.1% at medium noise, and 79.3% at high noise, compared with 85.4%, 78.6%, and 71.9% respectively for the baseline model. This suggests that GAN augmentation not only encourages more general representations but also improves reliability on imperfect, real-world data. Taken together, these findings make a compelling case for using GANs to improve neural network performance, offering a promising avenue for addressing data scarcity and overfitting.

The heatmap in Figure 3 further demonstrates how different models withstand adversarial attacks, with Model E showing the strongest overall robustness, particularly against FGSM and DeepFool, in contrast to the relative weaknesses of Models C and D against PGD and BIM. This disparity highlights how strongly GAN-augmented training can influence adversarial resistance depending on model architecture. Similarly, the normalized confusion matrix in Figure 4 shows the prediction and misclassification trends of the GAN-augmented model across the ten CIFAR-10 classes, with high diagonal values indicating correct classification and off-diagonal values highlighting specific class pairs prone to confusion — information that can guide further refinement through improved feature extraction or additional augmentation.

Overall, this study demonstrates that GANs can improve neural network robustness to data scarcity and overfitting through the introduction of high-quality synthetic data. Improved accuracy, robustness, and generalization, together with improved performance against adversarial attacks and noisy data, provide strong evidence that GAN-augmented models are well suited to real-world deployment, offering a more stable and effective approach to neural network training.

5. Conclusions

This study demonstrates that combining Generative Adversarial Networks with neural networks can deliver substantial improvements in model performance across multiple dimensions. By generating high-quality synthetic data, GANs help address the problems of data scarcity and overfitting, enabling more generalizable models. The results indicate promising gains in accuracy, robustness, and generalization: the neural network trained with GAN-augmented data achieved higher accuracy on the CIFAR-10 test set, highlighting the performance benefits of synthetic data. The adversarial training procedure inherent in GANs also encouraged the neural network to learn more discriminative features, further improving generalization and robustness — evidenced by the model's increased resilience to adversarial perturbations such as FGSM, PGD, and CW, and its improved performance on noisy data. These improvements are essential for real-world applications, where models must generalize well across variable and unpredictable conditions.

These findings suggest that GANs are an effective tool for addressing common challenges in the training of deep neural networks, such as data scarcity and overfitting. Future studies could investigate the underlying mechanisms by which GANs improve performance, extend the approach to other application contexts, and work toward improving the stability and efficiency of GAN training algorithms. Overall, this study demonstrates the transformative potential of GANs in improving neural network performance, contributing to the development of models better aligned with the demands of real-world applications.

Limitations

While GANs offer considerable benefits, their use is not without drawbacks. Training GANs stably remains difficult and often requires extensive hyperparameter tuning. GANs are also computationally expensive, with training typically taking substantially longer than traditional supervised learning, particularly with large-scale datasets. In addition, GANs are prone to mode collapse, in which the generator produces only a limited range of outputs. Addressing the challenges of scaling and conditioning will be important for fully realizing the potential of GANs to improve neural network performance.

Future scope

Future research on integrating GANs with neural networks could pursue several directions, including the use of GAN-based techniques for unsupervised and semi-supervised learning of hidden-layer representations, and the development of improved GAN training methods and algorithms to address mode collapse and computational cost. A further direction involves extending the application of GANs to other domains, such as medical imaging, autonomous systems, and natural language processing, as well as exploring their combination with other state-of-the-art techniques such as reinforcement learning or self-supervised learning to further enhance neural network performance.

Declaration of competing interest

The authors declare that they have no known financial or non-financial competing interests in any material discussed in this paper.

Funding information

No funding was received from any financial organization to conduct this research.

Ethical approval statement

Ethical approval is not applicable for this research, as it did not involve human or animal subjects and

relied solely on the publicly available CIFAR-10 benchmark dataset.

Informed consent

Informed consent was not applicable, as this research did not involve personal data or human participants.

Author contribution

M. H. Yousif: conceptualization, methodology, data analysis, and manuscript preparation. The author approved the final version of the manuscript.

Reference

Z. Huang, C. O. Dumitru, Z. Pan, B. Lei, and M. Datcu, "Classification of Large-Scale High-Resolution SAR Images With Deep Transfer Learning," IEEE Geoscience and Remote Sensing Letters, vol. 18, pp. 107–111, 2021.

L. Wang, X. Yang, H. Tan, X. Bai, and F. Zhou, "Few-Shot Class-Incremental SAR Target Recognition Based on Hierarchical Embedding and Incremental Evolutionary Network," IEEE Transactions on Geoscience and Remote Sensing, vol. 61, pp. 1–11, 2023.

J. Qin, Z. Liu, L. Ran, R. Xie, J. Tang, and Z. Guo, "A Target SAR Image Expansion Method Based on Conditional Wasserstein Deep Convolutional GAN for Automatic Target Recognition," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 15, pp. 7153–7170, 2022.

X. Li, Z. Du, Y. Huang, and Z. Tan, "A Deep Translation (GAN) Based Change Detection Network for Optical and SAR Remote Sensing Images," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 179, pp. 14–34, 2021.

S. Saha, F. Bovolo, and L. Bruzzone, "Building Change Detection in VHR SAR Images via Unsupervised Deep Transcoding," IEEE Transactions on Geoscience and Remote Sensing, vol. 59, pp. 1917–1929, 2021.

M. Fuentes Reyes, S. Auer, N. Merkle, C. Henry, and M. Schmitt, "SAR-to-Optical Image Translation Based on Conditional Generative Adversarial Networks—Optimization, Opportunities and Limits," Remote Sensing, vol. 11, p. 2067, 2019.

Y. Zhao, T. Celik, N. Liu, and H.-C. Li, "A Comparative Analysis of GAN-Based Methods for SAR-to-Optical Image Translation," IEEE Geoscience and Remote Sensing Letters, vol. 19, pp. 1–5, 2022.

Z. Zuo and Y. Li, "A SAR-to-Optical Image Translation Method Based on PIX2PIX," in Proceedings of 2021 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Brussels, Belgium, 2021, pp. 3026–3029.

J. Gao, Q. Yuan, J. Li, H. Zhang, and X. Su, "Cloud Removal With Fusion of High Resolution Optical and SAR Images Using Generative Adversarial Networks," Remote Sensing, vol. 12, p. 191, 2020.

X. Yang, Z. Wang, J. Zhao, and D. Yang, "FG-GAN: A Fine-Grained Generative Adversarial Network for Unsupervised SAR-to-Optical Image Translation," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–11, 2022.

P. Ebel, M. Schmitt, and X. X. Zhu, "Cloud Removal in Unpaired Sentinel-2 Imagery Using

Cycle-Consistent GAN and SAR-Optical Data Fusion," in Proceedings of IGARSS 2020 IEEE International Geoscience and Remote Sensing Symposium, Waikoloa, HI, USA, 2020, pp. 2065–2068.

G. Baier, A. Deschemps, M. Schmitt, and N. Yokoya, "Synthesizing Optical and SAR Imagery From Land Cover Maps and Auxiliary Raster Data," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–12, 2021.

F. Gao, X. Xu, J. Yu, M. Shang, X. Li, and D. Tao, "Complementary, Heterogeneous and Adversarial Networks for Image-to-Image Translation," IEEE Transactions on Image Processing, vol. 30, pp. 3487–3498, 2021.

H. Tang, D. Xu, N. Sebe, Y. Wang, J. J. Corso, and Y. Yan, "Multi-Channel Attention Selection GAN With Cascaded Semantic Guidance for Cross-View Image Translation," in Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 2019, pp. 2417–2426.

L. Jiang, C. Zhang, M. Huang, C. Liu, J. Shi, and C. C. Loy, "TSIT: A Simple and Versatile Framework for Image-to-Image Translation," in Proceedings of European Conference on Computer Vision, Glasgow, UK, 2020, pp. 206–222.

P. Esser, R. Rombach, A. Blattmann, and B. Ommer, "ImageBART: Bidirectional Context With Multinomial Diffusion for Autoregressive Image Synthesis," Advances in Neural Information Processing Systems, vol. 34, pp. 3518–3532, 2021.

J. Ho, C. Saharia, W. Chan, D. J. Fleet, M. Norouzi, and T. Salimans, "Cascaded Diffusion Models for High Fidelity Image Generation," Journal of Machine Learning Research, vol. 23, pp. 1–33, 2022.

T. Wang, L. Wu, and C. Sun, "A Coarse-to-Fine Approach for Dynamic-to-Static Image Translation," Pattern Recognition, vol. 123, p. 108373, 2022.

E. Agustsson, M. Tschannen, F. Mentzer, R. Timofte, and L. V. Gool, "Generative Adversarial Networks for Extreme Learned Image Compression," in Proceedings of IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 2019, pp. 221–231.

C. Chen, D. Gong, H. Wang, Z. Li, and K.-Y. K. Wong, "Learning Spatial Attention for Face Super-Resolution," IEEE Transactions on Image Processing, vol. 30, pp. 1219–1231, 2020.