Proceedings of CONF-MLA 2025 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2025.LD28611

Performance Bottlenecks and Solutions in Deep Learning for Image Recognition

Zikai Chen

Department of Physics, King's College London, London, United Kingdom
weudawn@163.com

Abstract. Deep learning has advanced image recognition, achieving strong results in medical imaging, autonomous driving, and security. Yet significant bottlenecks still limit deployment. This paper reviews three main challenges: weak robustness, high computational demands, and reliance on large labeled datasets. Recent studies identify the causes of these issues, including growing model complexity, distribution shifts between training and real data, and lack of security-aware design. To address these problems, various strategies have been developed in the past five years. For robustness, adversarial training, data augmentation, and domain adaptation have been widely applied. To enhance the efficiency of deep learning models, techniques including network pruning, parameter quantization, and lightweight architectures (e.g., MobileNet and EfficientNet) are widely adopted—often augmented by knowledge distillation and hardware-aware neural architecture search (NAS). To mitigate reliance on large-scale labeled datasets, approaches such as transfer learning, self-supervised learning frameworks (e.g., SimCLR and BYOL), and multimodal models (e.g., CLIP) have demonstrated promising performance. While progress is evident, trade-offs remain. Future work should focus on combining these strategies to achieve models that are simultaneously accurate, efficient, and robust for real-world applications.

Keywords: deep learning, image recognition, robustness, efficiency, transfer learning

1. Introduction

Image recognition is a fundamental task in computer vision, with applications in healthcare, autonomous driving, and security surveillance [1,2]. Recent advancements in deep learning, particularly convolutional neural networks (CNNs) and vision transformers (ViTs), have substantially improved recognition accuracy on large-scale benchmarks [3-5]. These models now achieve state-of-the-art performance and are increasingly deployed in practice.

Despite this progress, several challenges remain for real-world deployment. A primary concern is robustness: models that perform well in controlled environments often fail when exposed to noise, occlusion, or adversarial perturbations [6]. In addition, modern architectures typically contain tens of millions of parameters, leading to high computational and storage requirements that hinder deployment on mobile or embedded devices [7,8]. Finally, these models heavily rely on large annotated datasets, which are costly and time-consuming to obtain, particularly in specialized domains such as medical imaging and satellite monitoring [9].

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.LD28611

This paper reviews these three critical bottlenecks in deep learning–based image recognition: robustness, efficiency, and data dependence. While previous surveys have broadly examined robustness in computer vision [5] and efficiency-oriented techniques [7,8], few have considered the combined challenges addressed here. By synthesizing recent studies, this review identifies the causes of these limitations and evaluates strategies proposed in the past five years to mitigate them. The findings provide insights for guiding future research and advancing more robust, efficient, and scalable image recognition systems.

2. Current situation and challenges

2.1. Lack of robustness

Deep learning has dramatically advanced image recognition, yet robustness remains one of its weakest points. Models that achieve over 90% accuracy on clean benchmark datasets can fail catastrophically when exposed to subtle changes. Wang et al. [6] classify robustness issues into adversarial robustness, corruption robustness, and distributional robustness, showing that many state-of-the-art models score far lower in stress tests than on benchmarks. For instance, a ResNet-50 trained on ImageNet can lose more than 40% Top-1 accuracy under Gaussian noise or JPEG compression, despite performing well on the clean test set.

Adversarial examples are the most widely discussed form of fragility. Small, carefully designed pixel-level perturbations-often invisible to humans-can cause misclassifications with high confidence. Javed et al. [1] highlight the implications in medical imaging: chest X-ray classifiers that detect pneumonia correctly in controlled environments may fail once adversarial noise is introduced, risking serious clinical consequences [1]. Similar vulnerabilities have been demonstrated in autonomous driving, where simply adding inconspicuous stickers to a stop sign can cause misrecognition, potentially leading to accidents.

Robustness is also challenged by natural variations. In practice, lighting conditions, partial occlusions, or sensor-specific artifacts degrade model performance significantly. Balendran et al. [2] observe that medical imaging systems trained on high-quality hospital data often underperform on community-acquired datasets with lower resolution or inconsistent preprocessing. Attempts to address these gaps, such as extensive data augmentation or domain adaptation, help but do not fully resolve the issue.

These observations demonstrate that although deep learning models achieve high accuracy under ideal conditions, they remain structurally brittle in real-world operational scenarios—thereby constraining their deployment in high-stakes application domains.

2.2. Computational and storage costs

Another critical bottleneck in deep learning–based image recognition lies in the computational and storage demands of modern architectures. As models become deeper and wider, their parameter counts and memory requirements rise dramatically. For example, ResNet-152 contains roughly 60 million parameters, while recent Vision Transformers (ViTs) and Swin Transformers can exceed 80–100 million [3]. This scale translates into billions of floating-point operations (FLOPs), making training and inference both time-consuming and energy-intensive. While such models perform impressively on benchmarks like ImageNet, their deployment on mobile devices, surveillance cameras, or medical imaging equipment remains highly impractical.

DOI: 10.54254/2755-2721/2025.LD28611

The issue becomes more acute in real-time applications. In autonomous driving, for instance, vehicle perception systems must process high-resolution video streams at 30 frames per second or more. However, large networks introduce inference delays that jeopardize safety. Similarly, in telemedicine and portable health monitoring, devices with limited hardware struggle to support such computationally heavy models. Liu et al. [4] emphasize that the high training costs—often requiring hundreds of GPU hours—are an additional barrier, particularly for research groups or organizations with limited resources.

Beyond cost and speed, there is also the environmental impact. Training a single large model can emit as much carbon as several cars over their lifetime. Paula et al. [7] compare model compression strategies and demonstrate that pruning and quantization can reduce energy usage by up to 50%, with minimal accuracy loss. Khan et al. [8] further highlight hardware-aware optimization as an emerging direction for sustainable AI. While cloud computing can offload part of the workload, transmitting large amounts of image data raises latency and privacy concerns, particularly in healthcare. Thus, efficiency is not simply an optimization problem—it is central to whether deep learning systems can be deployed safely and responsibly at scale.

2.3. Data dependence

A further obstacle to the widespread deployment of deep learning in image recognition is its heavy reliance on large, labeled datasets. Landmark datasets such as ImageNet and MS-COCO have enabled enormous progress in computer vision, but they are not representative of many specialized domains. In medical imaging, for example, creating labeled datasets requires expert radiologists to annotate CT or MRI scans, which is both costly and time-consuming. Trigka [9] notes that this scarcity of annotated data has slowed the adoption of deep learning methods in healthcare, despite their potential benefits. Similar challenges exist in satellite and remote-sensing applications, where images often need expert geoscientists to provide accurate ground-truth labels.

Even when annotated data is available, domain shift remains a significant issue. Models trained on one dataset may fail to generalize to new conditions. For instance, face recognition systems trained on high-quality, balanced datasets often misclassify in surveillance environments with poor lighting and lower resolution. Likewise, a diagnostic model trained on chest X-rays from one hospital may lose accuracy when applied to data from another due to differences in imaging devices, preprocessing protocols, or patient demographics. These mismatches demonstrate that the success of deep learning in controlled environments often does not translate smoothly to real-world scenarios.

Techniques such as transfer learning and data augmentation have reduced the demand for extensive labeling, but they do not fully eliminate the problem. More recently, self-supervised learning and multimodal representation learning have emerged as promising approaches to leverage vast amounts of unlabeled data. However, as Trigka [9] emphasizes, ensuring these methods achieve consistent reliability across domains remains a critical challenge.

3. Causes of bottlenecks

3.1. Model complexity

The pursuit of higher accuracy has encouraged researchers to design increasingly deep and wide architectures. This growth in complexity leads to a dramatic rise in the number of parameters and floating-point operations, which significantly inflates computational demands and memory usage [3]. As a result, large parameter spaces make optimization more difficult: gradient updates become

DOI: 10.54254/2755-2721/2025.LD28611

less stable, and training requires careful regularization to avoid divergence. In addition, highly complex models tend to overfit when data is limited, as they memorize training patterns rather than learning robust representations. Liu et al. [4] emphasize that the relationship between model capacity and generalization is nonlinear—beyond a certain threshold, additional parameters offer marginal accuracy gains while sharply increasing computational and storage costs. Thus, model complexity itself becomes a structural source of inefficiency and fragility.

3.2. Data distribution shift

A second cause of bottlenecks is the violation of the fundamental assumption in supervised learning: that training and test data are drawn from the same distribution. In reality, this assumption rarely holds. Image data collected in different settings can vary in resolution, color balance, noise level, or background patterns. When the statistical properties of input data shift, models trained on one dataset fail to generalize because they have learned correlations specific to the training distribution rather than universal features. Trigka [9] points out that this is not simply a lack of data volume, but a mismatch in data distributions, which undermines the stability of learned representations. Unlike the challenges outlined in Section 2, which highlight empirical performance drops, the deeper cause here is that deep learning models lack inherent mechanisms for adaptation to out-of-distribution inputs.

3.3. Lack of security-aware design

Finally, many deep learning systems are developed with a singular optimization goal: maximizing predictive accuracy on clean datasets. This design philosophy neglects adversarial robustness and security considerations. Wang et al. [6] note that the loss functions commonly used in training do not incorporate robustness terms, leaving models vulnerable to imperceptible perturbations. Furthermore, standardized evaluation protocols for robustness are still underdeveloped, meaning that weaknesses remain hidden until deployment. Javed et al. [1] argue that this issue reflects a cultural bias in AI research toward benchmark accuracy, rather than a holistic consideration of reliability. The absence of robustness-aware design means that even models with excellent accuracy can be destabilized by small perturbations, which explains why adversarial and noisy conditions create such disproportionate performance degradation.

4. Solutions and strategies

4.1. Improving robustness

Improving the robustness of deep learning models has become a central research focus, as fragile models limit the reliability of image recognition in safety-critical contexts. A key line of defense is adversarial training, where models are exposed to adversarially perturbed examples during training. Wang et al. report that such training paradigms substantially enhance model resilience against gradient-based adversarial attacks [6]. However, this improvement is achieved at the expense of prolonged training duration and elevated computational requirements. Despite these inherent trade-offs, adversarial training remains among the most effective strategies for mitigating deliberate input perturbations.

Beyond adversarial defenses, data-centric approaches play an equally important role. Augmentation methods such as Cut Mix, Mix Up, and Auto Augment artificially expand training sets by creating diverse variations of the data. Javed et al. [1] show that these techniques enhance

DOI: 10.54254/2755-2721/2025.LD28611

robustness not only against synthetic perturbations but also against natural variations like lighting changes or occlusions. More advanced strategies leverage generative adversarial networks (GANs) to create realistic synthetic samples, which help models generalize better in domains where annotated data is scarce.

Another promising development is domain adaptation and domain generalization, which reduce sensitivity to dataset shifts. Balendran et al. [2] highlight how multi-site training in medical imaging —combining data from different hospitals—improves generalization across institutions. Additionally, robust optimization techniques such as randomized smoothing and certified defenses aim to provide formal guarantees on model behavior under perturbations.

In practice, no single method is sufficient on its own. Combining adversarial training with data augmentation and domain adaptation offers a more comprehensive path forward. However, these improvements often come with increased training cost, leaving open the question of how to balance robustness with efficiency-an issue further explored in Section 4.2.

4.2. Enhancing efficiency

Improving efficiency is crucial for deploying deep learning models in real-world environments where computational resources and energy budgets are limited. A primary strategy is model compression, which reduces redundancy in neural networks. Network pruning techniques involve the removal of non-critical weights or neurons from deep neural networksstudies have demonstrated that convolutional neural networks (CNNs) can undergo pruning ratios of up to 90%while incurring only minimal degradation in model accuracy [3]. Quantization is another common method, lowering memory and computational requirements by converting 32-bit floating-point parameters into 8- or even 4-bit integers. Together, pruning and quantization can dramatically reduce inference time and storage needs.

In parallel, lightweight architectures has been designed specifically for constrained devices. Models such as MobileNet [10], ShuffleNet, and EfficientNet provide strong accuracy while reducing the number of parameters and floating-point operations (FLOPs) by orders of magnitude. Liu et al. [4] emphasize that these networks achieve competitive ImageNet performance with far fewer resources than traditional ResNets, making them practical for mobile applications and embedded systems.

Another promising method is knowledge distillation, where a large “teacher” model guides a smaller “student” model to mimic its outputs. Paula et al. [7] reports that this approach maintains much of the teacher’s accuracy while drastically lowering computational demands. More recently, hardware-aware neural architecture search (NAS) has gained traction, automatically designing models optimized for specific hardware platforms.

Despite these advances, trade-offs remain. Compressed or lightweight models may lose fine-grained recognition ability, and training with distillation adds complexity. Combining compression, lightweight design, and hardware-aware optimization offers a practical path to balance accuracy and efficiency.

4.3. Reducing data dependence

Another major line of research seeks to reduce the heavy reliance of deep learning models on large labeled datasets. Transfer learning is a widely adopted strategy, where models pretrained on large datasets such as ImageNet are fine-tuned on smaller domain-specific datasets. This method has

DOI: 10.54254/2755-2721/2025.LD28611

proven effective in fields like medical imaging, where annotated data is scarce but pretrained weights provide strong initialization [9].

Beyond transfer learning, self-supervised learning has emerged as a powerful alternative to supervised training. Techniques such as SimCLR, BYOL, and MoCo learn representations from unlabeled data by solving pretext tasks, such as predicting whether two image augmentations come from the same original image. These methods allow models to leverage massive collections of unlabeled images, significantly reducing annotation requirements. Trigka [9] notes that self-supervised frameworks have shown promise in healthcare and satellite imagery, where labeled data is particularly limited.

A third promising approach is multimodal learning, where models jointly process image and text data. CLIP, for example, is trained on hundreds of millions of image–text pairs and generalizes well across tasks without task-specific supervision. Such models exploit the abundance of text data available on the web, indirectly reducing dependence on labeled images.

5. Conclusion

Deep learning has transformed image recognition, delivering state-of-the-art performance across diverse domains such as healthcare, autonomous driving, and security. However, this review has shown that significant challenges remain for practical deployment. Three primary bottlenecks—limited robustness, high computational and storage costs, and strong dependence on large labeled datasets—continue to restrict the reliability and scalability of current models. These limitations arise from increasing model complexity, domain distribution shifts, and the absence of robustness- or security-aware design.

A wide range of strategies have been proposed to address these issues. Robustness can be improved through adversarial training, advanced data augmentation, and domain adaptation, although these methods often increase training complexity. Efficiency has been enhanced by model compression, lightweight architecture, and knowledge distillation, enabling deployment on mobile and embedded devices while also reducing the environmental cost of training. Finally, dependence on labeled data is being mitigated through transfer learning, self-supervised methods, and multimodal learning, which exploit vast pools of unlabeled or weakly labeled data.

Despite this progress, trade-offs remain. Improving robustness often comes at the expense of computational cost, while efficiency-oriented techniques may reduce accuracy. Similarly, self-supervised methods demand substantial pretraining resources. Looking ahead, integrating these strategies into unified frameworks will be essential for balancing accuracy, robustness, and efficiency. Moreover, advances in multimodal learning, privacy-preserving computation, and explainable AI are likely to shape the next generation of image recognition systems.

In summary, deep learning has advanced, but building models that are accurate, efficient, and robust remains a challenge. Continued research will be critical for safe and scalable deployment.

References

[1] Javed H, El-Sappagh S, Abuhmed T. (2024) Robustness in deep learning models for medical diagnostics: security and adversarial challenges towards robust AI applications. Artif Intell Rev

[2] Balendran A, Beji C, Bouvier F, Khalifa O, Evgeniou T, Ravaud P, et al. (2025) A scoping review of robustness concepts for machine learning in healthcare. npj Digit Med, 8: 38

[3] Dantas PV, da Silva Jr W, Carvalho LC, Carvalho CB. (2024) A comprehensive review of model compression techniques in machine learning. Appl Intell, 54: 11804–11844

DOI: 10.54254/2755-2721/2025.LD28611

[4] Liu D, Zhu Y, Liu Z, Liu Y, Han C, Tian J, et al. (2025) A survey of model compression techniques: past, present, and future. Front Robot AI

[5] Liu J, Zhang K, Wang H, Chen L. (2023) A comprehensive survey of robust deep learning in computer vision. Comput Vis Image Underst, 229: 1036576.

[6] Wang J, Ai J, Lu M, Su H, Yu D, Zhang Y, et al. (2024) A survey of neural network robustness assessment in image recognition. arXiv [Internet]. 2024 Apr 12 [cited 2025 Aug 17]. Available from: https: //arxiv.org/abs/2404.08285

[7] Paula E, Soni J, Upadhyay H, Lagos L. (2025) Comparative analysis of model compression techniques for achieving carbon efficient AI. Sci Rep, 15(1): 23461. DOI: 10.1038/s41598-025-07821-w

[8] Khan Z, Ali S, Rehman S, Zhang T, Hussain F. (2025) Deep learning model compression and hardware acceleration for efficient deployment. Sensors, 25(3): 970

[9] Trigka M. (2025) A comprehensive survey of deep learning approaches in image analysis. Sensors, 25(2): 531. DOI: 10.3390/s25020531

[10] Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H. (2017) MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv: 1704.04861 [Internet]. Available from: https: //arxiv.org/abs/1704.04861