Politecnico di Torino

ScuDo
Scuola di Dottorato ~ Doctoral School
WHAT YOU ARE, TAKES YOU FAR

Doctoral Dissertation
Doctoral Program in Computer and Control Engineering (37^{th}cycle)

Efficiency Matters: Modern Techniques for Efficient Computer Vision

By

Niccolò Cavagnero

******

Supervisor:
Prof. Barbara Caputo

Doctoral Examination Committee:
Prof. Lamberto Ballan, Referee, University of Padova
Prof. Pietro Zanuttigh, Referee, University of Padova
Prof. Bastian Leibe, RWTH Aachen University
Prof. Gijs Dubbelman, Eindhoven University of Technology
Prof. Marco Mellia, Polytechnic of Turin

Politecnico di Torino
2025

Declaration

I hereby declare that, the contents and organization of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

Niccolò Cavagnero
2025

* This dissertation is presented in partial fulfillment of the requirements for **Ph.D. degree** in the Graduate School of Politecnico di Torino (ScuDo).

Efficiency Matters: Modern Techniques for Efficient Computer Vision

Niccolò Cavagnero

This thesis addresses the critical need for efficiency in deep learning, with a particular focus on computer vision applications. As the scale and complexity of AI systems continue to increase, their computational demands pose significant environmental and operational challenges. Furthermore, when deployed in safety-critical environments, these systems must not only perform accurately but also remain reliable under adverse conditions. This work presents a collection of strategies to improve the efficiency of deep neural networks, enabling their deployment in real-world settings where performance, energy consumption, and low latency are equally important.

The first part of this thesis is dedicated to the automatic design of vision backbones for image classification through Neural Architecture Search (NAS). We describe a set of training-free methods that drastically reduce search time compared to classic training-based methods, by replacing costly training evaluations with efficient heuristics. In doing so, we demonstrate that combining complementary training-free metrics leads to the discovery of high-quality architectures at a fraction of the traditional computational cost. Crucially, this combination is not naïve: we present a cyclical search algorithm that decouples the topology and size dimensions, strategically exploiting each metric where it is most informative, thus enhancing not only search efficiency but also the quality of discovered network architectures.

Shifting focus from image classification to more complex downstream tasks, the second part of the thesis investigates efficient architectures for image segmentation. In particular, we present two efficient segmentation models: *PEM* and *EoMT*. In PEM, we revisit task-specific components found in modern segmentation networks and design a lightweight convolutional decoder alongside a more efficient cross-attention mechanism. Our decoder replicates the benefits of transformer-based alternatives while substantially reducing computational overhead, and our efficient cross-attention design accelerates inference by exploiting the inherent redundancy of visual features. Building upon this, EoMT

takes a step further toward simplicity and efficiency by leveraging foundation models to remove these task-specific components entirely. This results in a conceptually simple architecture that exhibits not only high performance but also impressive speed, achieving state-of-the-art accuracy with significantly reduced complexity.

In the final part of this thesis, we address the challenge of improving model reliability in the presence of transient hardware faults caused by neutron-induced radiation, while preserving the computational efficiency critical for real-time and resource-constrained applications. Traditional fault-tolerance strategies, including hardware redundancy and Error-Correcting Codes (ECC), typically come with significant area, power, or latency overheads, severely limiting their practical deployment in real-world scenarios. In contrast, we present an efficient, deep learning-based solution that incorporates fault mitigation directly into the network architecture and training procedure. This approach, termed *DieHardNet*, achieves robust inference without introducing any runtime overhead. The effectiveness of DieHardNet is rigorously validated through a comprehensive experimental campaign, combining not only application-level and instruction-level fault injections but also real-world neutron beam exposure across multiple commercial GPU platforms.

Altogether, this thesis contributes a set of modern techniques for designing deep neural networks that are not only accurate but also efficient, reliable, and deployable in real-world environments.