Initial Training Image Size Effects on CNN Accuracy-Efficiency Trade-offs Across Domains

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: How does the initial training image size affect the trade-off between accuracy and training efficiency in state-of-the-art CNNs (e.g., EfficientNet, Vision Transformers) when trained on mixed-domain. 8 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Vicinity Vision Transformer. Research question: How does the initial training image size affect the trade-off between accuracy and training efficiency in state-of-the-art CNNs (e.g., EfficientNet, Vision Transformers) when trained on mixed-domain datasets, as measured by HLCE benchmark accuracy and training time?.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.2/10.

3 Results

16 papers retrieved. 8 claims extracted; 1 independently verified. Quality review score: 5.2/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
Vicinity Attention introduces a Manhattan distance-based 2D locality to linear vision transformers.	✓	0.25
The Vicinity Attention Block contains a feature reduction attention (FRA) to improve efficiency and a feature preserving	×	0.08
VVT serves as a general-purpose vision backbone and can be easily applied to various vision tasks.	×	0.05
Extensive experiments validate the effectiveness of VVT on various computer vision benchmarks.	×	0.14
The computational complexity of linearized self-attention grows linearly with respect to the input length.	×	0.06
The computational complexity of standard self-attention grows quadratically with respect to the input length.	×	0.07
VVT-T achieves a Top-1 accuracy of 79.4% with 12.9M parameters and 3.0 GFLOPs.	×	0.03
VVT-S achieves a Top-1 accuracy of 96.1% with 12.9M parameters and 3.0 GFLOPs.	×	0.03

References

- http://arxiv.org/abs/1807.11583v1
- http://arxiv.org/abs/2206.10552v2
- http://arxiv.org/abs/2403.11999v1