



#### INTRODUCTION

**Background:** Transformers are the SOTA models for image classification and other computer vision tasks. Their  $O(N^2)$ computational complexity makes it hard to work with long sequences/high resolution images. Therefore, a wide variety of architectural modifications have been proposed to make transformers more efficient.

Question: Which transformer variant is the most efficient one and under what circumstances?

**Problem:** There is a wide diversity of training and evaluation conditions, so papers are not comparable on a fair basis.

Solution: We conduct a comprehensive large-scale analysis of the efficiency of 45+ transformer-like models for image classification on ImageNet. We train every model from scratch. We compare model efficiency using the Pareto front.



Authors: Tobias Nauen (tobias\_christian.nauen@dfki.de) (RPTU, DFKI), Sebastian Palacio (ABB), Federico Raue (DFKI), Andreas Dengel (DFKI, RPTU).

# Which Transformer to Favor? A Comparative Analysis of Efficiency in Vision Transformers

### The baseline ViT model is still Pareto optimal for image classification.

## Larger models are more efficient than higher resolution images.







#### **OBSERVATIONS**



- TokenLearner, ToMe.



![](_page_0_Picture_23.jpeg)

![](_page_0_Picture_26.jpeg)

• Using high-resolution (384<sup>2</sup> px) images is not Pareto optimal (dashed vs. dotted lines). • ViT is pareto optimal at all sizes for 3 out of 4 metrics.

• Token sequence reduction models offer a good tradeoff in speed vs. accuracy; especially

RPTU: University of Kaiserslautern-Landau, Kaiserslautern DFKI: German Research Center for Artificial Intelligence, Kaiserslautern ABB: ABB Germany, Mannheim