Differentiable Hierarchical Visual Tokenization

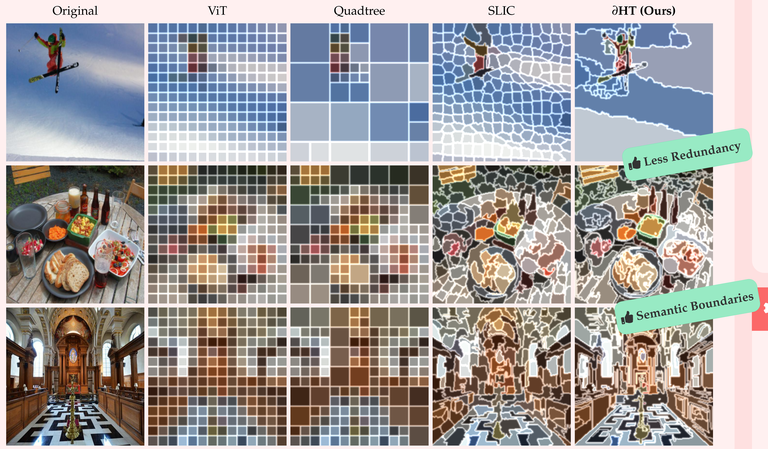
Marius Aasan, Martine Hjelkrem-Tan, Nico Catalano, Changkyu Choi and Adín Ramírez Rivera
University of Oslo, Polytechnic University of Milan, UiT The Arctic University of Norway

Motivation: Are Patch Tokens Optimal for Vision?

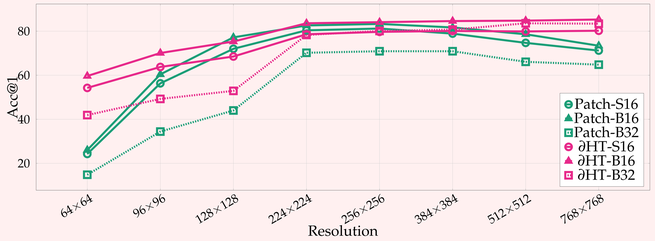
∂HT replaces fixed patches with adaptive, semantically aligned tokens.

- **Semantic Boundaries**: Patch tokens ignore object structure, fragmenting meaningful regions.
- **Redundancy**: Uniform grids waste compute on locally homogeneous areas with little signal diversity.
- **Non-adaptive**: Prior tokenizers lack end-to-end learning, fixed-budget ViTs cannot adapt token granularity.
- **Proposal**: *Use hierarchical pixel-level partitions + information-criteria pruning to select optimal tokens.*

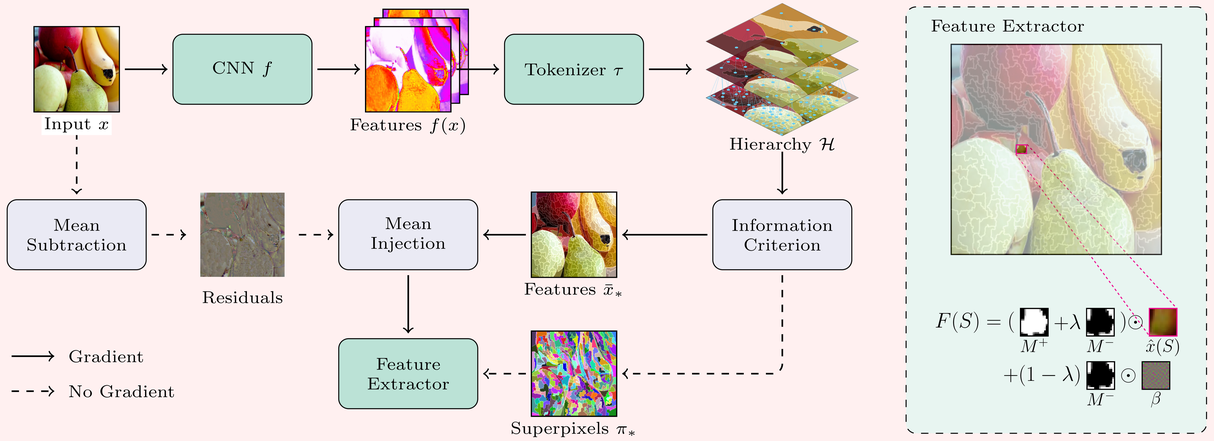
Comparing Token Granularity



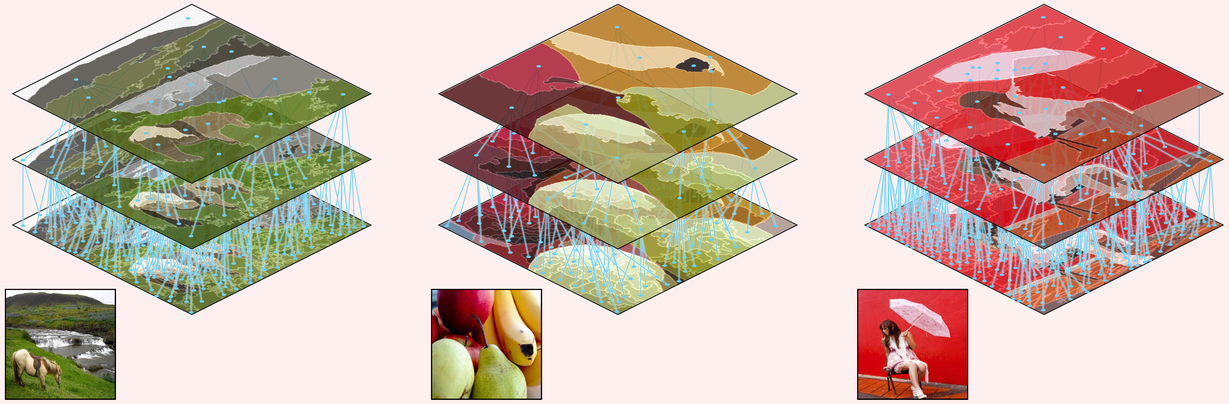
∂HT: Improved Scale Invariance



Computational Pipeline: From Image to Tokens



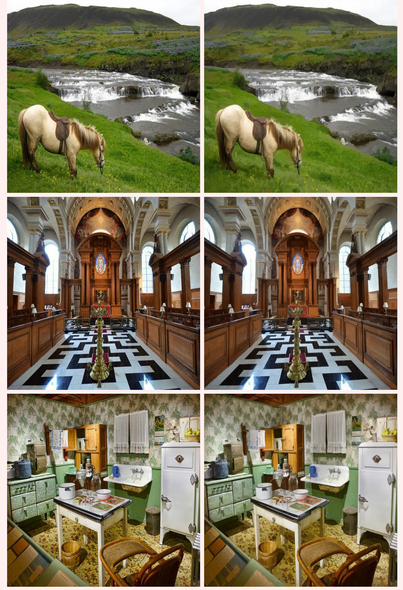
∂HT: Hierachical Token Merging and Pruning



Raster-to-Vector Graphics

Out-of-the-box!

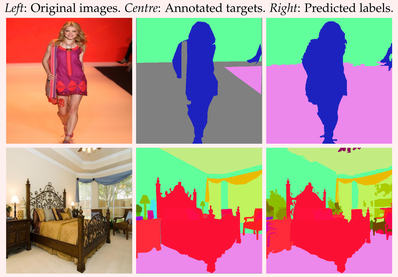
Left: Raster images. *Right*: Vectorized Images.



Semantic Segmentation

MLP Decoder

Left: Original images. *Centre*: Annotated targets. *Right*: Predicted labels.



NeurIPS 2025 — Conference and Workshops on Neural Information Processing Systems — December, 2025 — San Diego, CA, USA

ACKNOWLEDGEMENTS: EuroHPC LUMI (Sigma2 NN8104K) & EuroHPC LEONARDO (ISCRA, Cineca) & SFI Visual Intelligence (RCN 309439) & AURoRA (RCN 359216, RS-2025-03522980).