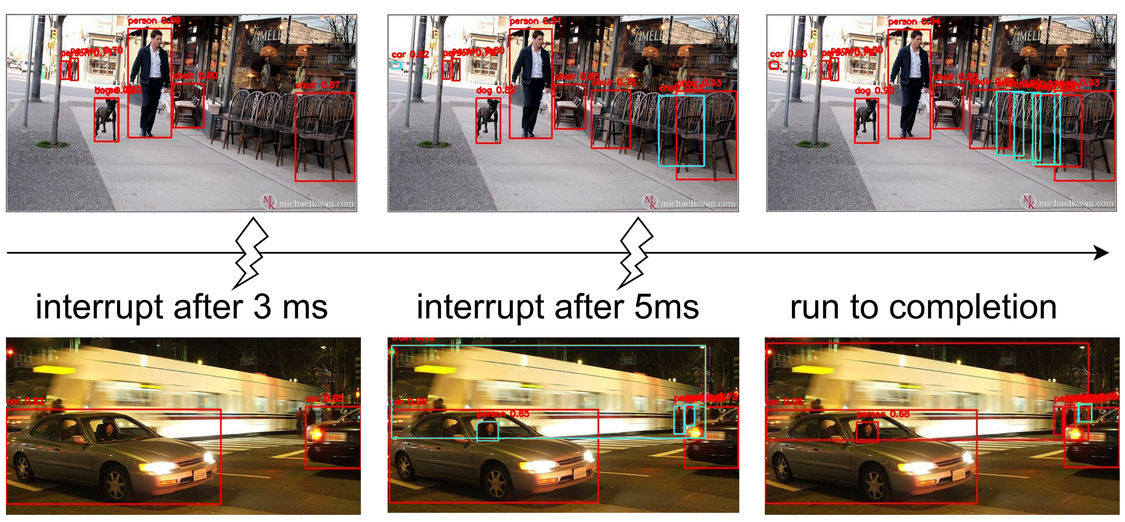
Anytime YOLO
Early exits for interruptable object detection

LAMARR INSTITUTE FOR MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE

Authors: Daniel Kuhse, Harun Teper, Sebastian Buschjäger, Chien-Yao Wang, Jian-Jia Chen

Anytime Object Detection

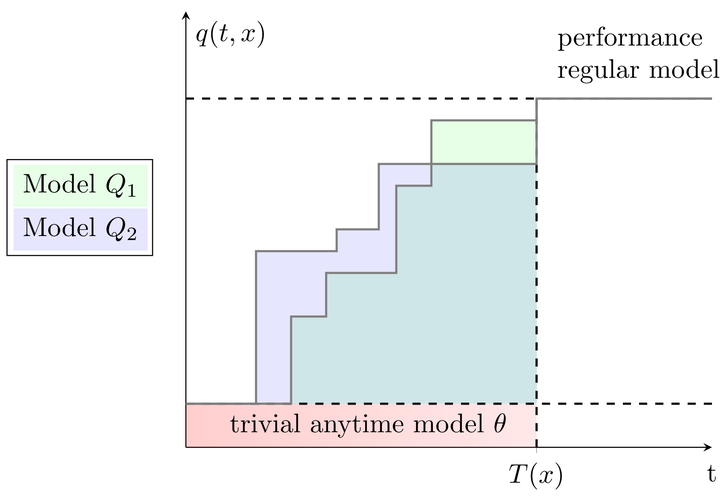
- **Interruptable** inference that still leads to results
- Interrupt time **unknown** in advance (budgeted vs anytime)
- Use **early exits** to provide results from intermediate layers



Anytime Quality

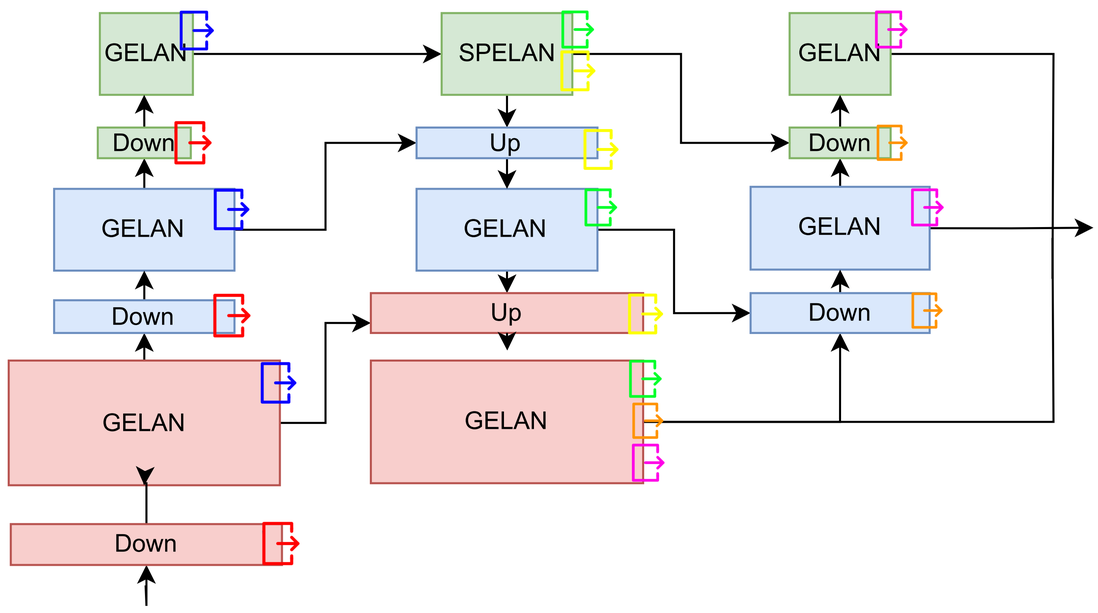
- No previous definition of anytime quality
- Suggestion: normalized,weighted area of performance plot

$$Q = \mathbb{E}_{x,y\sim\mathcal{D}}\left[\frac{1}{T(x)}\int_0^{T(x)} q(f(t,x),y)w(t)dt\right]$$

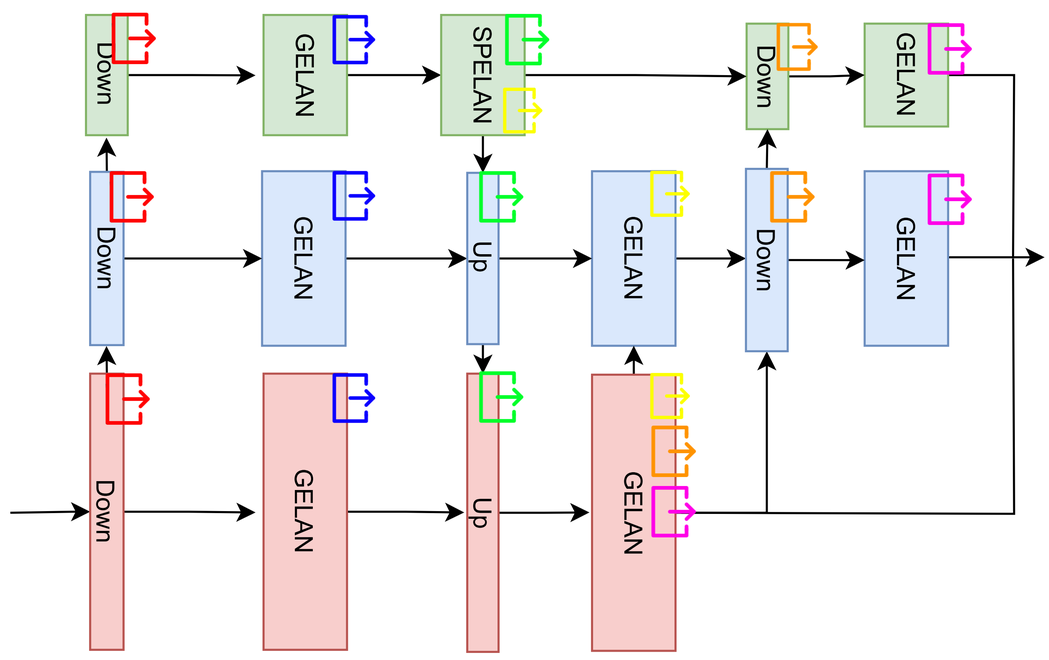


Early Exit Architecture enabling Anytime

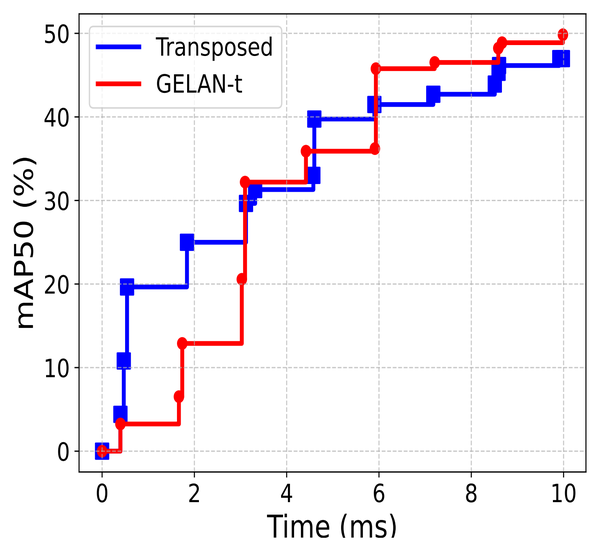
- YOLO architecture pyramidal **–** 3 **subexits** per exit
- Subexits operate at different resolutions
- Subexits can be mixed

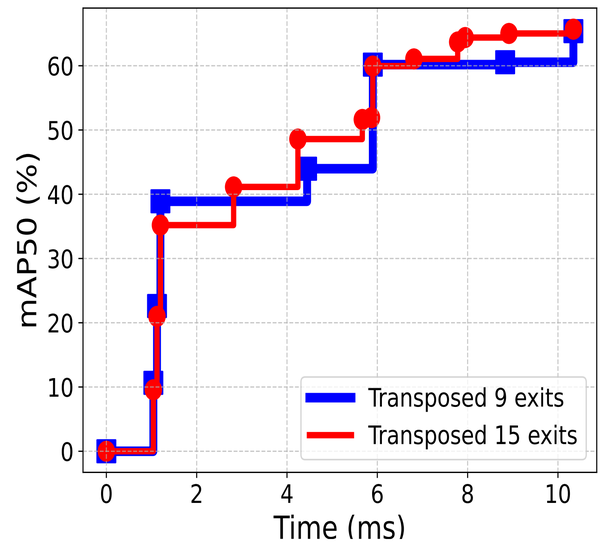


- Pyramidal architecture has strict order
- Idea: transpose it!
- Earlier exit with 3 resolutions, can optimize order



Performance





Challenges

- GPUs currently **uninterruptable** (chunked execution)
- Better target MCUs?
- **Soft** anytime – compute final result after interrupt (delay)
- **Hard** anytime – return immediately (redundant results)
- YOLO has **high exit cost** combined with **post-processing**
- Future: NMS-less? Cheaper heads? Stitching together models?
- For which applications is overhead worth it?

Partner institutions:
technische universität dortmund
Fraunhofer IAIS
UNIVERSITÄT BONN
Fraunhofer IML

Institutionally funded by:
Bundesministerium für Forschung, Technologie und Raumfahrt
Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen