



Deploying intelligence in embedded devices with STM32 edge AI tools

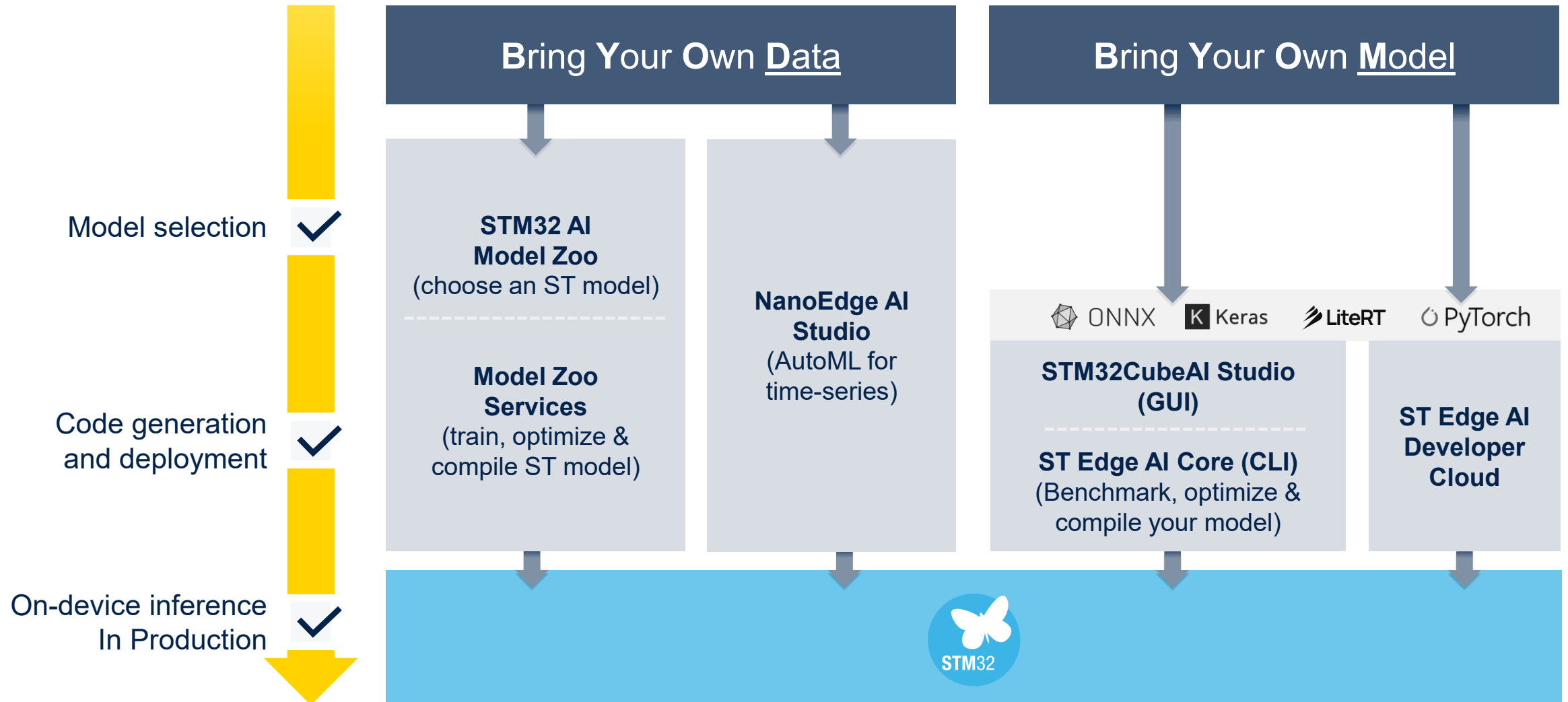
Feb 2026



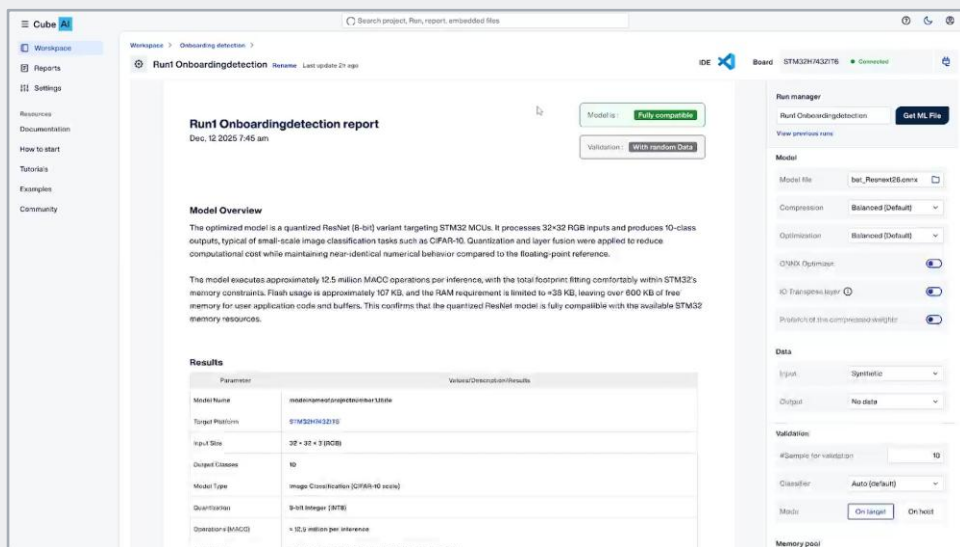
**Develop faster
with free STM32 edge AI tools**



STM32 Edge AI User Flow



STM32CubeAI Studio



[Download](#)

[Read documentation](#)

Current version: V1.0
(Feb 2026)

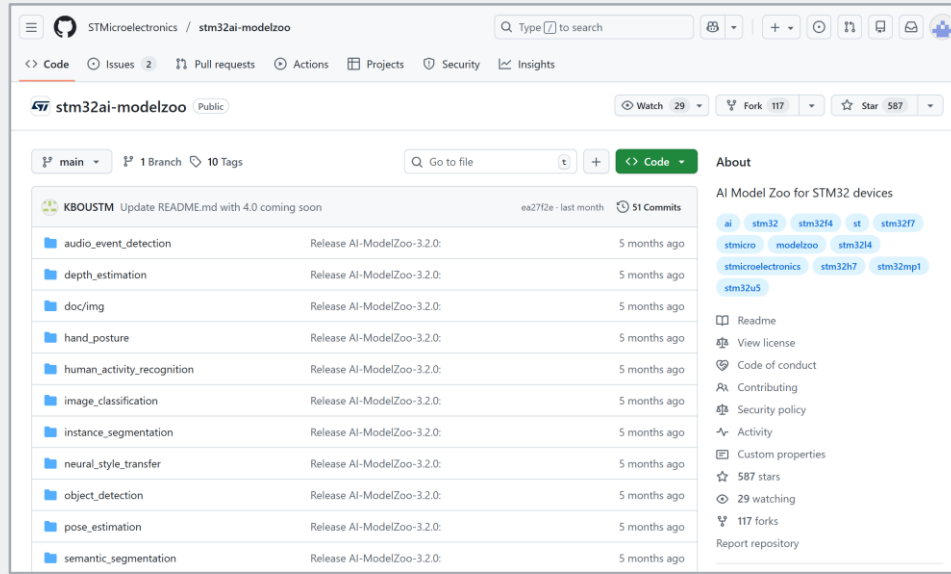


Evaluate, optimize and compile neural network models for STM32.

- Generate an STM32-optimized library from pretrained neural network and ML models
- Use detailed information about AI model RAM and flash memory sizes
- Validate optimized model against reference model on host and on target
- Support for ST Neural-ART Accelerator neural processing unit
- Support for 32-bit float and quantized neural network formats (TensorFlow™ Lite and ONNX Tensor-oriented QDQ)

An evolution of STM32Cube.AI supporting the latest generation of STM32 MCUs, and offering a smooth user experience.

STM32 Model Zoo



Access now

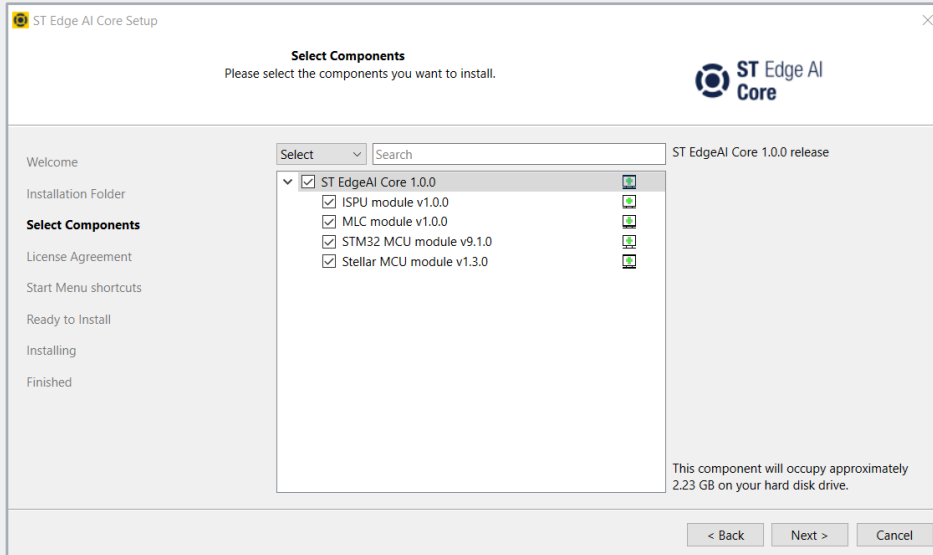
Current version: V4.0
(Feb 2026)



The largest collection of free reference edge AI models optimized for STM32

- Choose from a collection of AI models optimized for STM32 devices
- Retrain any model from datasets using the scripts
- Deploy AI models on the application using the scripts
- Bring your own model (BYOM) or bring your own data (BYOD)
- Available on GitHub
- Free-of-charge, user-friendly license terms

ST Edge AI Core



Download

[Read documentation](#)

Current version: V3.0
(Dec 2025)



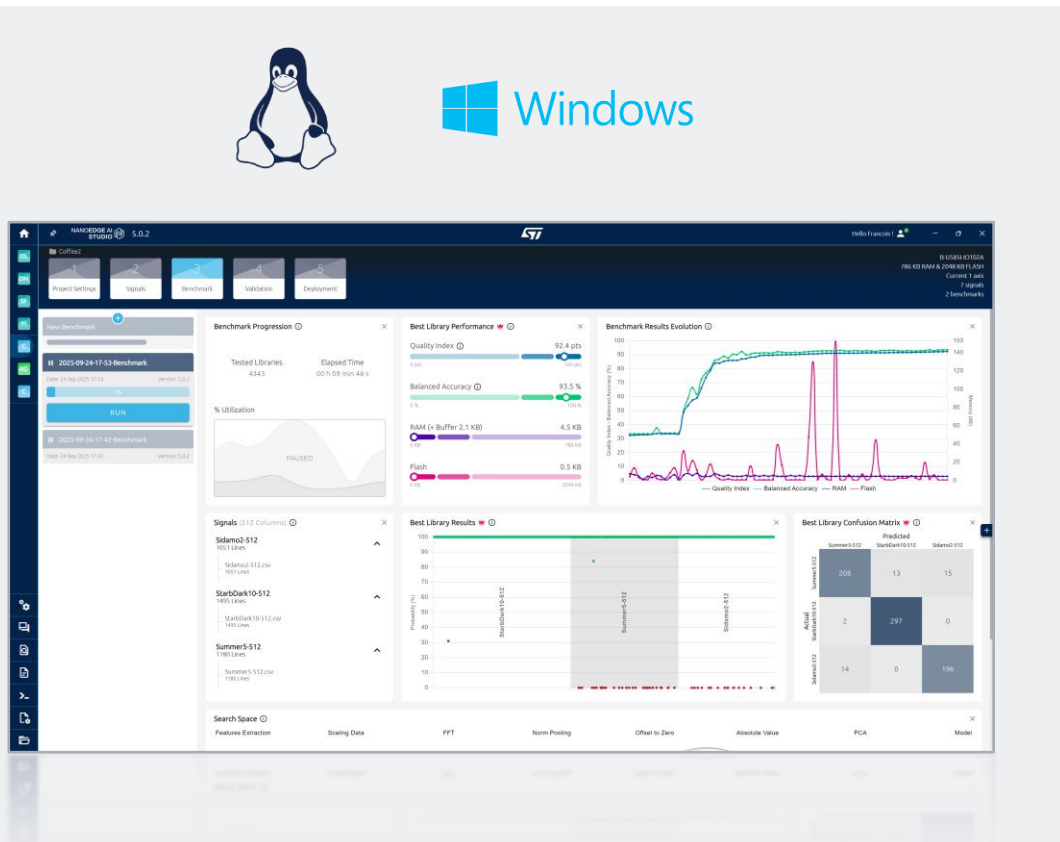
Optimize and compile edge AI models for STM32 with this command-line interface tool (CLI)

- Supports multiple ST devices, including microcontrollers, microprocessors, and MEMS sensors
- Import AI models from the most popular ML frameworks
- Review the detailed analysis of the selected algorithms
- Validate and optimize the models for different devices
- Generate executable code for the selected ST hardware
- Support ST Neural-ART NPU and HSP

NanoEdge AI Studio

Free AutoML tool to develop your time series ML library in a fast and efficient way.

- Find and configure the best AI library for your embedded project.
- Supports STM32 microcontrollers and smart sensors
- Use synthetic data for creating anomaly detection models
- On-device learning for optimized anomaly detection
- Simplified data extraction from sensors
- Free-of-charge, user-friendly license terms



Download

[Read documentation](#)

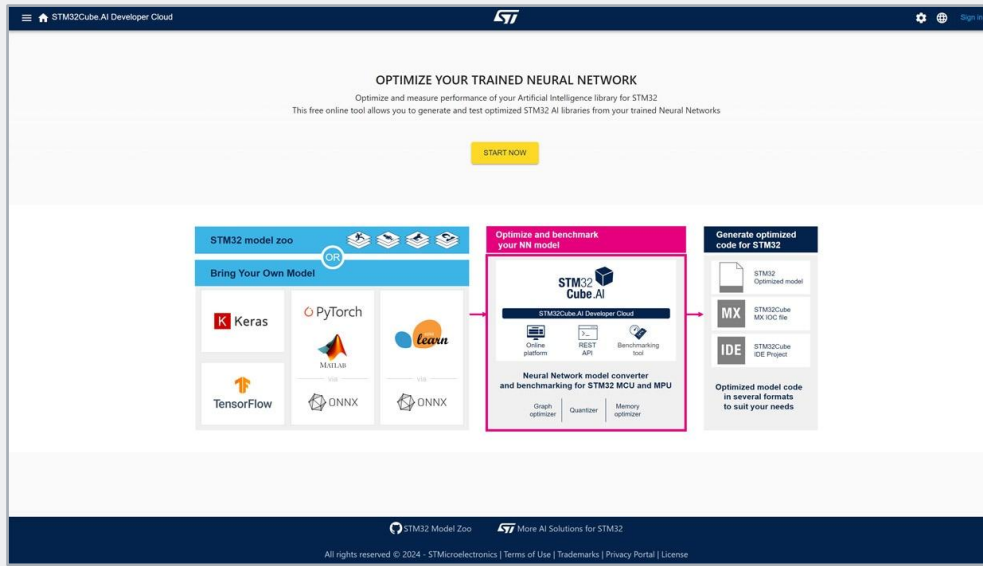
Current version: V5.1.0
(Feb 2026)



ST Edge AI Developer Cloud

Evaluate and benchmark your edge AI models in the cloud

- Built on ST Edge AI Core to perform AI model optimizations and validations on multiple ST devices
- Online AI benchmark service on multiple ST devices
- Optimize and profile your AI model
- Run your AI models on the ST board farm



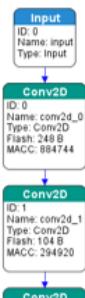
Access now

[Read documentation](#)

The 3 pillars of STM32Cube.AI

Graph optimizer

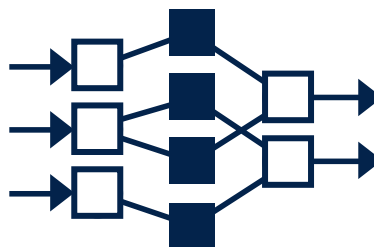
Automatically improve performance through graph simplifications & optimizations that benefit STM32 target HW architectures



- Auto graph rewrite
- Node/operator fusion
- Layout optimization
- Constant-folding...
- Operator-level info to finetune memory footprint and computation

Quantized model support

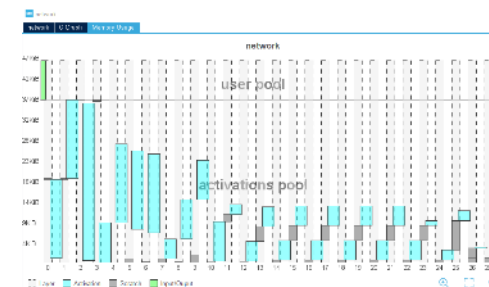
Import your quantized ANN to be compatible with STM32 embedded architectures while keeping their performance



- From FP32 to Int8 or mixed-precision
- Minimum loss of accuracy
- Code validation on target
 - Latency
 - Accuracy
 - Memory footprint

Memory optimizer

Optimize memory allocation to get the best performance while respecting the constraints of your embedded design

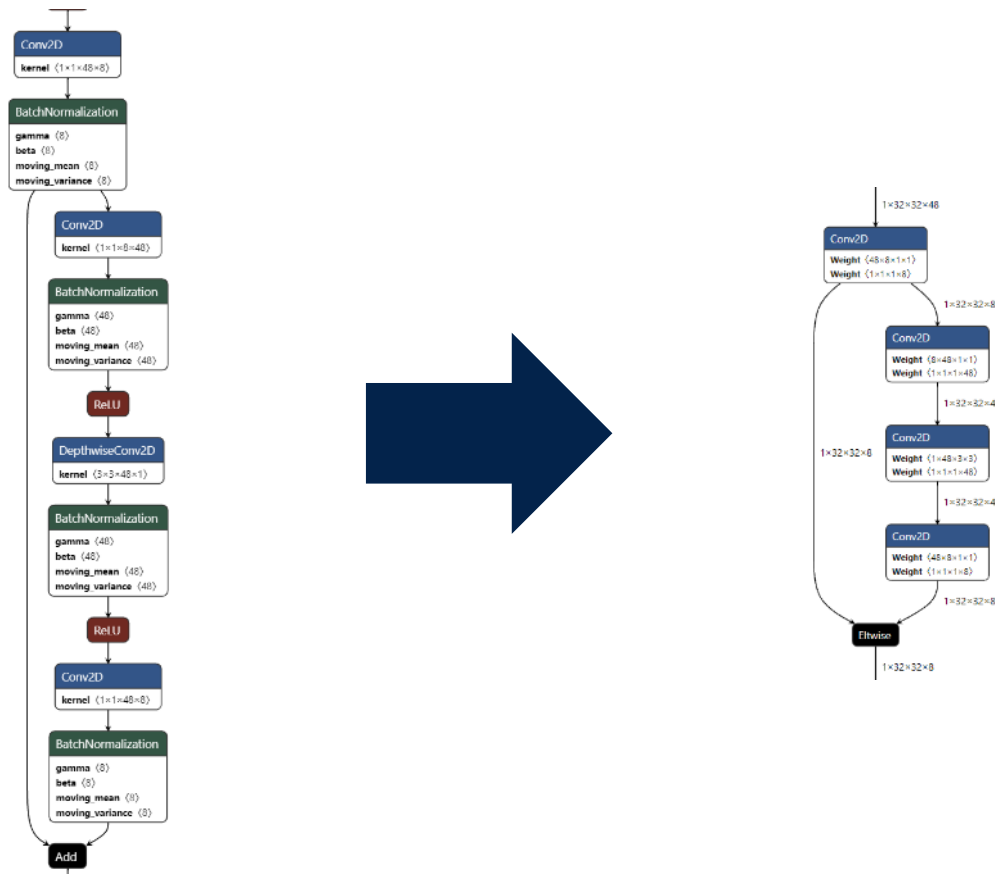


- Memory allocation
- Internal/external memory repartition
- Model-only update option

STM32Cube.AI is **free of charge**, available both in graphical interface and in command line.

Graph optimizer

Squeeze your graph to fit into an MCU!



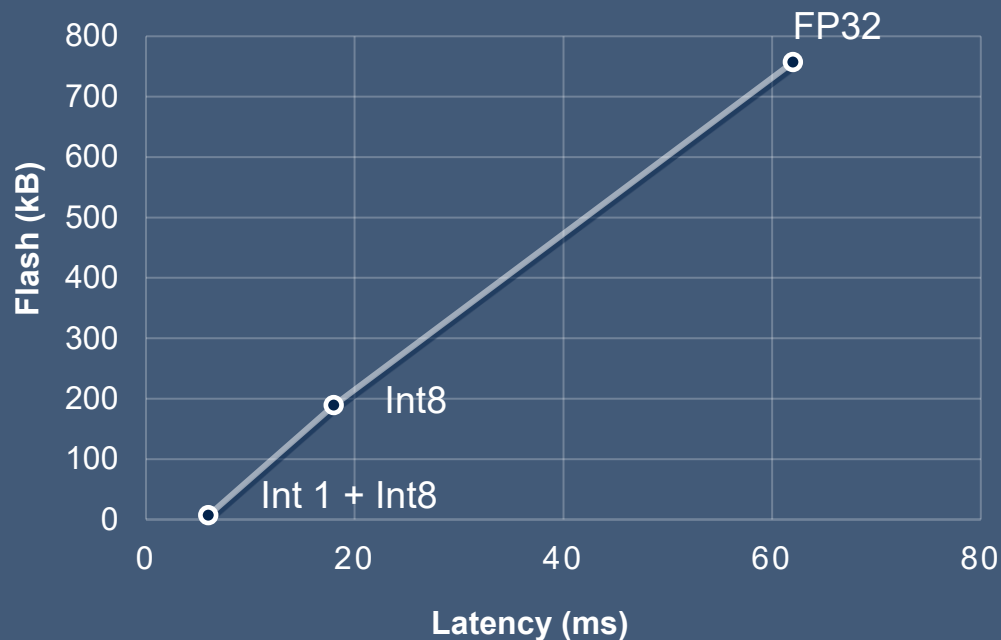
Fully automated process in the STM32Cube.AI workflow

- Your original graph is optimized at the very early stage for optimal integration into the STM32 MCU/MPU
- Loss-less conversion

Quantized model support

Simply use quantized networks to reduce memory footprint and inference time

LATENCY & MEMORY COMPARISON FOR QUANTIZED MODELS



STM32Cube.AI supports quantized neural network models with **all parameter formats**:

- FP32
- Int8
- Mixed binary Int1 to Int8 (Qkeras*, Larq.dev*)

**Please contact edge.ai@st.com to request the relevant version of STM32Cube.AI*



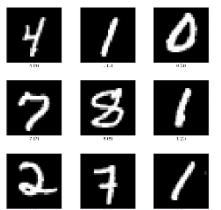
HW Target: NUCLEO-STM32H743ZI2

Model: Low complexity handwritten digit reading

Freq: 480 MHz

Accuracy: >97% for all quantized models

Tested database: MNIST dataset



MNIST dataset

Memory optimizer

Optimize performance easily with the memory allocation tool



Model RAM consumption per layer

- Easily identify the most critical layers

Model memory allocation

- Set your external memory
- Map in non-contiguous internal flash section
- Partition internal vs external flash memories

Re-use model input buffer to store activation data*

- Minimize RAM requirements

Relocatable network

- A separate binary is generated for the library and the network to enable standalone model upgrade

☒ Use external flash Memory: Custom

Split weights between internal and external flash using a linker script

Start Address: 0x00000000 Size (Mbytes)

Tensor	Size	Internal 440KB	External 0KB
conv1_weights	864	☒	☐
conv1_bias	32	☒	☐
conv_dw_1_weights	288	☒	☐
conv_dw_1_bias	32	☒	☐
conv_pw_1_weights	512	☒	☐

☐ Use external RAM Memory: Custom

Start Address: 0x00000000

☒ Use activation buffer

Start Address: 0x00000000 Act. size (by... 752712

☐ Copy weight to RAM

Start Address: Weight size: 451496

☒ Use activation buffer for input buffer (--allocate-inputs) ☐ Force classifier validation output (--classifier)

☒ Use activation buffer for the output buffer (--allocate-outputs)

☒ Split weights during code generation (--split-weights)

☒ Generate relocatable network (--relocatable)

Report's output directory

C:\Users\richard\stm32cube\mx Browse...

☐ Enable custom layer support

Custom Layer JSON File: Browse...



* Requires input and activation buffers in same memory

OK

Cancel

ST edge AI model partners



Facilitating the design of custom-made solutions

ST and Authorized Partners deliver ready-to-use solutions and speed up customer application development.



Irida Labs



EmbedUR



Ultralytics



Siana Systems



ClearML



id3 Technologies



OpenMV



Rtone



Edge impulse

[See all partners](#)

Browse our case studies and find inspiration for your next project

ST EDGE AI SUITE ▾ [Home](#) [Case studies](#) [Tools](#) [Edge AI explained](#) [Contact](#)

Home > ST Edge AI Suite > Case studies

Case studies

Discover how to improve embedded solutions using edge AI.

TYPES

- ☐ Customer
- ☐ Idea
- ☐ Partner
- ☐ Demo

RESOURCES

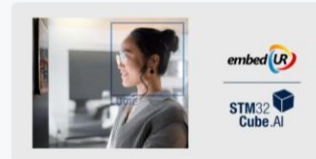
- ☐ Tutorial
- ☐ Function-Pack
- ☐ GitHub
- ☐ Video
- ☐ Dataset
- ☐ Model zoo

INDUSTRY

- ☐ Entertainment
- ☐ Wearables
- ☐ Smart city
- ☐ Smart home
- ☐ Smart building
- ☐ Smart office
- ☐ Industry

HARDWARE

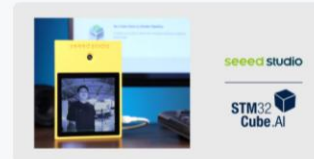
- ☐ STM32 MCU
- ☐ MEMS MLC
- ☐ STM32 AI MCU



VISION | STM32CUBE.AI | STM32 AI MCU | PARTNER | VIDEO

How to personalize smart home with familiar face identification

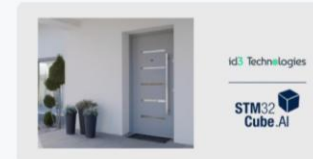
embedUR's on-device face authentication embedded on STM32N6 with easy mobile enrollment



VISION | STM32CUBE.AI | STM32 AI MCU | PARTNER | SMART HOME

Handheld development platform for real-time vision, motion, and voice at the edge

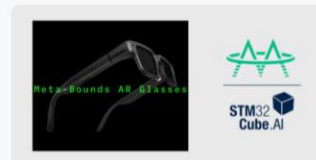
All-in-one STM32N6-based platform for on-device edge AI with NPU acceleration



INDUSTRIAL | SMART CITY | VISION | STM32CUBE.AI | STM32 AI MCU

Secure entry systems using id3 face recognition with liveness detection

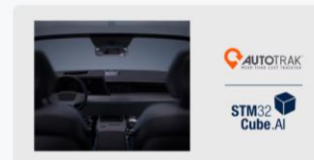
Edge processing with RGB camera and ToF sensor ensures rapid and secure anti-spoofing for access control using the...



VISION | STM32CUBE.AI | STM32 AI MCU | VIDEO | CUSTOMER

How ultralight AR glasses are redefined with Edge AI

Meta-Bounds elevates AR glasses with STM32N6 MCU edge AI and computer vision technology



VISION | STM32CUBE.AI | STM32 AI MCU | VIDEO | CUSTOMER

How Autotrak enhances in-cabin monitoring with cost-effective edge AI

Prevent distracted driving with real-time edge AI camera technology powered by the STM32N6 microcontroller

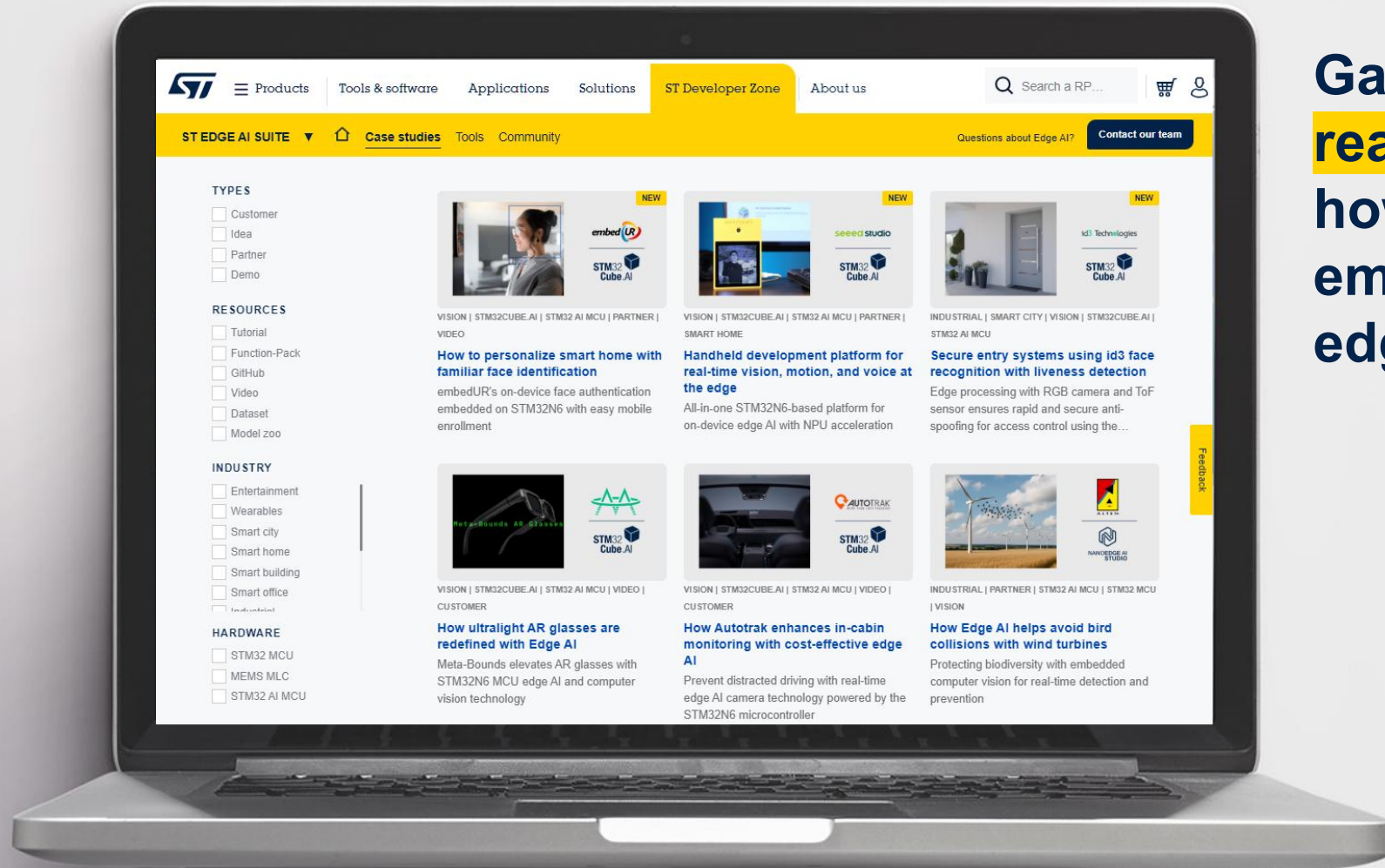


INDUSTRIAL | PARTNER | STM32 AI MCU | STM32 MCU | VISION

How Edge AI helps avoid bird collisions with wind turbines

Protecting biodiversity with embedded computer vision for real-time detection and prevention

Gain **valuable insights from real-world case studies** on how to improve your embedded solutions using edge AI.



Edge AI case studies

AR-Glasses
10 hours
On single-charge

Meta-Bounds AR Glasses



Meta-Bounds 莫界

Panasonic



Virtual sensor

Tire pressure
measured through the
e-motor current
consumption

~15-40%
Energy saving per
washing cycle



**Predictive
maintenance**

Multisensors and
learning on device

oxytronic



**Electrical arc
detection**

Enhanced accuracy to
avoid false detection



**Computer vision
on MCU**

Cost-effective, compact,
and low-power solution

AUTOTRAK
MORE THAN JUST TRACKING



Virtual sensor

Measuring motor
internal temperature
based on external data

HPE GROUP



**Smart context
detection**

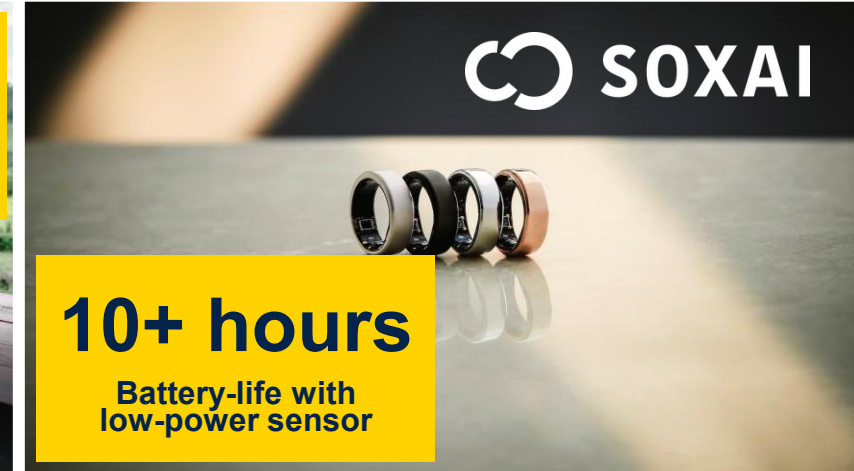
In/out bag - on table



SOXAI

10+ hours

Battery-life with
low-power sensor



Our technology starts with You



Find out more at www.st.com/st-edge-ai-suite

© STMicroelectronics - All rights reserved.

ST logo is a trademark or a registered trademark of STMicroelectronics International NV or its affiliates in the EU and/or other countries.

For additional information about ST trademarks, please refer to www.st.com/trademarks.

All other product or service names are the property of their respective owners.

