16T Large Model AI Box

- AIBOX-1688
- AIBOX-186

V1.0 2024-9-12

T-CHIP INTELLIGENCE TECHNOLOGY

Product features

A new generation of AIoT processors

Equipped with SOPHON computing AI processor BM1688/CV186AH, it has an octa-core/hexa-core ARM Cortex-A53, with a maximum frequency of 1.6GHz, and a built-in neural network acceleration engine TPU, AIBOX-1688 has a computing power of 16T@INT8, and AIBOX-186 has a computing power of 6T@INT8.

Powerful multi-channel video AI performance

The AI box supports up to 32 channels of 1080P H.265/H.264 video decoding and 32 channels of 1080P HD video processing (decoding + AI analysis), making it ideal for various AI applications such as face detection and license plate recognition on video streaming.

The private deployment of large language models

Support the private deployment of ultra-large-scale parameter models under the Transformer architecture, including large language models such as Gemma-2B, LlaMa2-7B, ChatGLM3-6B, Qwen1.5-1.8B.

Multiple deep learning frameworks

Support traditional network architectures such as CNN, RNN, and LSTM;
a variety of deep learning frameworks, including TensorFlow, PyTorch, TensorRT, TFLite, PaddlePaddle, Caffe, ONNX, as well as custom operator development.

Product features

Strong network communication capability

With dual Gigabit Ethernet (1000Mbps/RJ45), the AI box ensures high-speed and stable network communication, meeting the needs of various application scenarios.

All-aluminum alloy enclosure for heat dissipation

The industrial-grade all-metal enclosure with aluminum alloy structure for thermal conduction. The side of the top cover features a grille design for efficient heat dissipation. Its top cover is a porous hexagonal design, combining elegance with high efficiency. The compact, exquisite device operates stably and meets the needs of various industrial-grade applications.

Abundant expansion interfaces

It has 2 × Gigabit Ethernet, 2 × USB3.0, 1 × TF Card, 1 × Type-C, 1 × HDMI2.0, 1 × Console and other expansion interfaces, which are convenient to connect various peripherals and realize multi-field product applications.

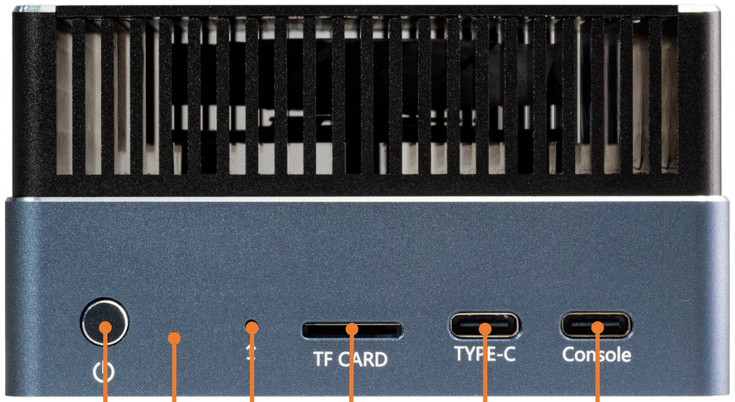
A wide range of applications

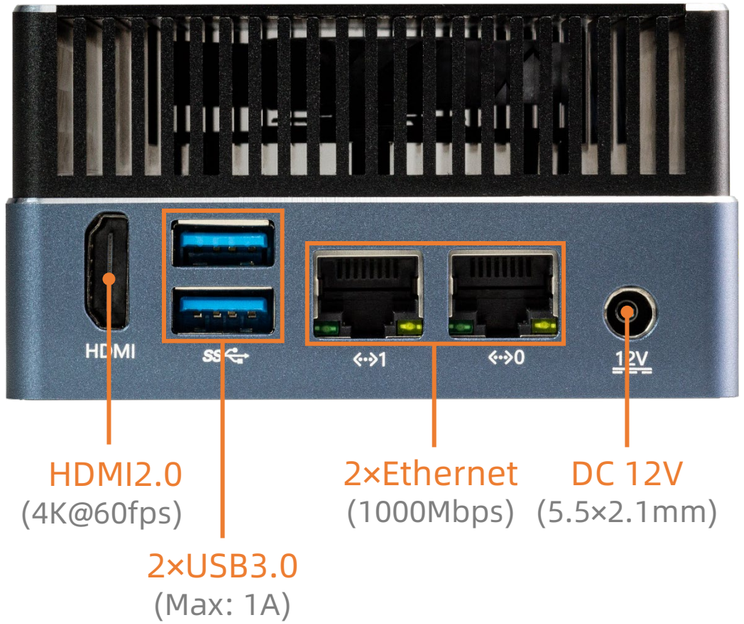
The device is widely used in intelligent surveillance, AI education, services based on computing power, edge computing, private deployment of large models, and data security and privacy protection.

Specifications

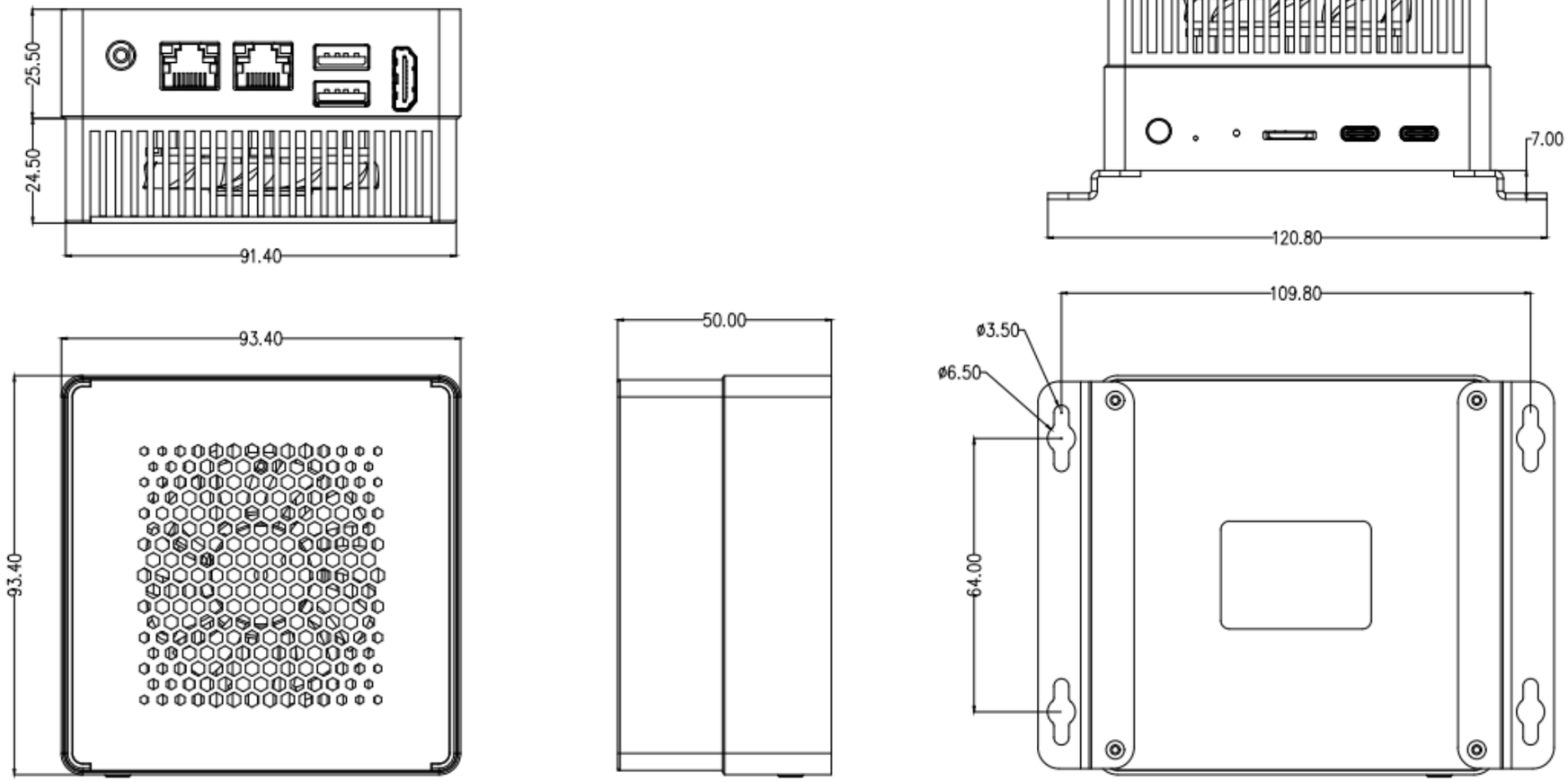
		AIBOX-1688	AIBOX-186
Basic Specifications	SOC	SOPHON BM1688	SOPHON CV186AH
	CPU	Octa-core ARM Cortex-A53 @ 1.6GHz	Hexa-core ARM Cortex-A53 @ 1.6GHz
	TPU	32T@INT4, 16T@INT8, 4T@FP16/BF16, and 0.5T@FP32 computing power	6T@INT8, 12T@INT4, and 1.5T@FP16/BF16 computing power
	Decoding/ Encoding	Video decoding: H.265/H.264 decoding (Max performance: 1920×1080@480FPS or 3840×2160@120FPS) Video encoding: H.265/H.264 encoding (Max performance: 1920×1080@300FPS or 3840×2160@75FPS) Image codec: JPEG/MJPEG Baseline codec (JPEG codec with a maximum resolution of 1080P@480FPS)	
	RAM	8GB LPDDR4 (4GB/8GB/16GB optional)	4GB LPDDR4 (4GB/8GB/16GB optional)
	Storage	32GB eMMC (32GB/64GB/128GB/256GB optional)	
	Storage Expansion	1 × M.2 (Expandable PCIe NVMe SSD(default support)/ SATA SSD(supported after software update), supports 2242/2260/2280) (inside the device), 1 × TF Card	
	Power	DC 12V/3A (5.5 × 2.1mm)	
	Power consumption	Normal: 7.2W(12V/600mA), Max: 14.4W(12V/1200mA)	Normal: 6W(12V/500mA), Max: 10.8W(12V/900mA)
	OS	Linux	
	Software support	・ The private deployment of ultra-large-scale parameter models under the Transformer architecture, including large language models such as Gemma-2B, LlaMa2-7B, ChatGLM3-6B, Qwen1.5-1.8B. ・ Traditional network architectures such as CNN, RNN, and LSTM; a variety of deep learning frameworks, including TensorFlow, PyTorch, TensorRT, TFLite, PaddlePaddle, Caffe and ONNX, enabling pedestrian detection, face detection, face recognition, liveness detection, and other video structured applications, as well as custom operator development ・ Docker container management technology	・ The private deployment of ultra-large-scale parameter models under the Transformer architecture, including large language models such as Gemma-2B, LlaMa2-7B, ChatGLM3-6B, Qwen1.5-1.8B. ・ Supports a variety of deep learning frameworks, including TensorFlow, PyTorch, TensorRT, TFLite, PaddlePaddle, Caffe and ONNX ・ Docker container management technology
	Size	93.4mm × 93.4mm × 50.0mm	
	Weight	≈ 500g	
	Environment	Operating Temperature: -20℃ ~ 60℃, Storage Temperature: -20℃ ~ 70℃, Storage Humidity: 10% ~ 90%RH (non-condensing)	
Interface Specifications	Ethernet	2 × Gigabit Ethernet (1000Mbps/RJ45)	
	Video output	1 × HDMI2.0 (4K@60fps)	
	USB	2 × USB3.0 (Max: 1A), 1 × Type-C (USB 2.0 OTG)	
	Other interfaces	1 × Console (Debug serial)	
	Button	1 × Power button, 1 × Recovery	

Interface description





Dimension


25.50
24.50
91.40
91.40
7.00
120.80
93.40
93.40
50.00
109.80
ø3.50
ø6.50
64.00

T-CHIP INTELLIGENCE TECHNOLOGY

Contact Us
(+86)18688117175

E-mail
global@t-firefly.com

Website
https://en.t-firefly.com/

Address
Room 2101, Hongyu Building, #57 Zhongshan 4Rd, East District, Zhongshan, Guangdong, China.