XDF XILINX DEVELOPER FORUM

Deep Learning Based Visual Perception System for Autonomous Driving

Presented By

魔视智能 MOTOVIS

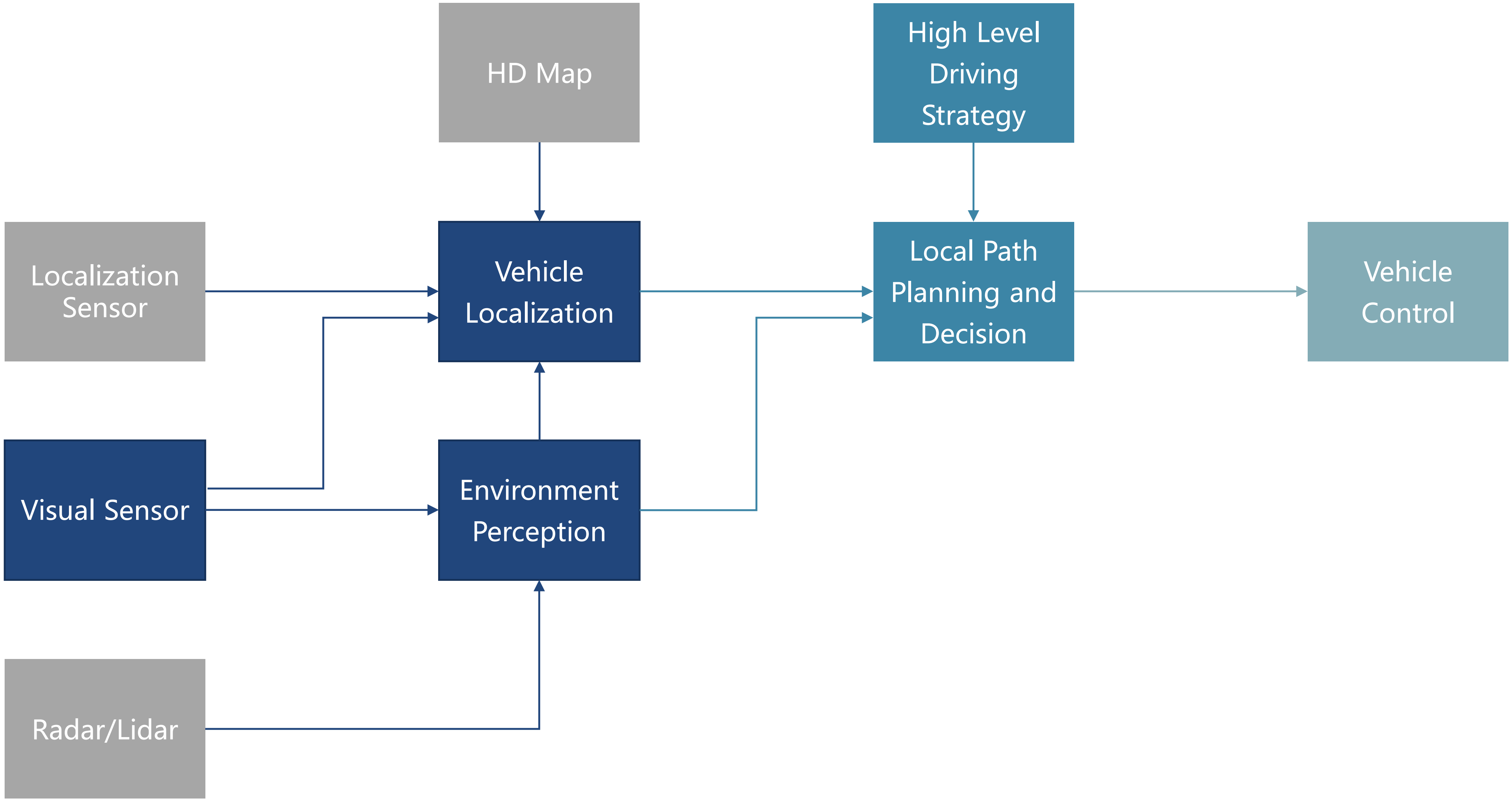
Dr. Zhenghua Yu
CEO
2018/10/16

XILINX

Autonomous Driving: From Dream to Reality

魔视智能 MOTOVIS

Environment Perception is key to Autonomous Driving



Challenges in Visual Perception


Head-on danger


Read-end danger

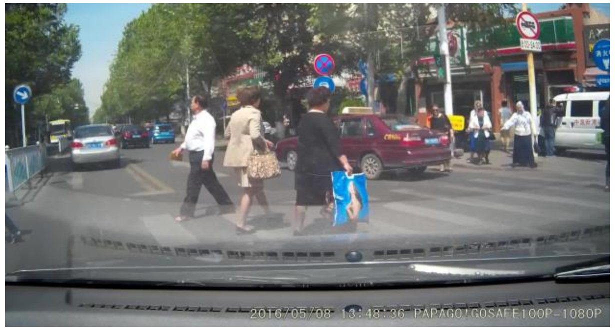

Cyclist danger


Intersection danger


Intersection danger

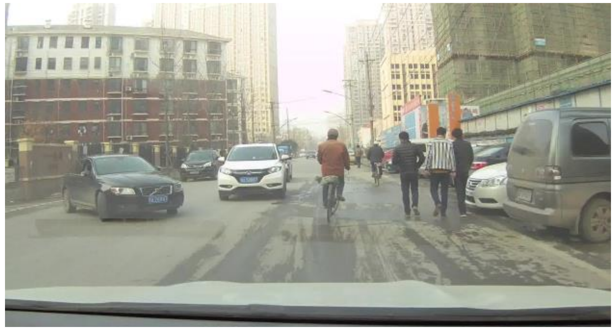

Cut-in danger


Cyclist danger


Pedestrian danger

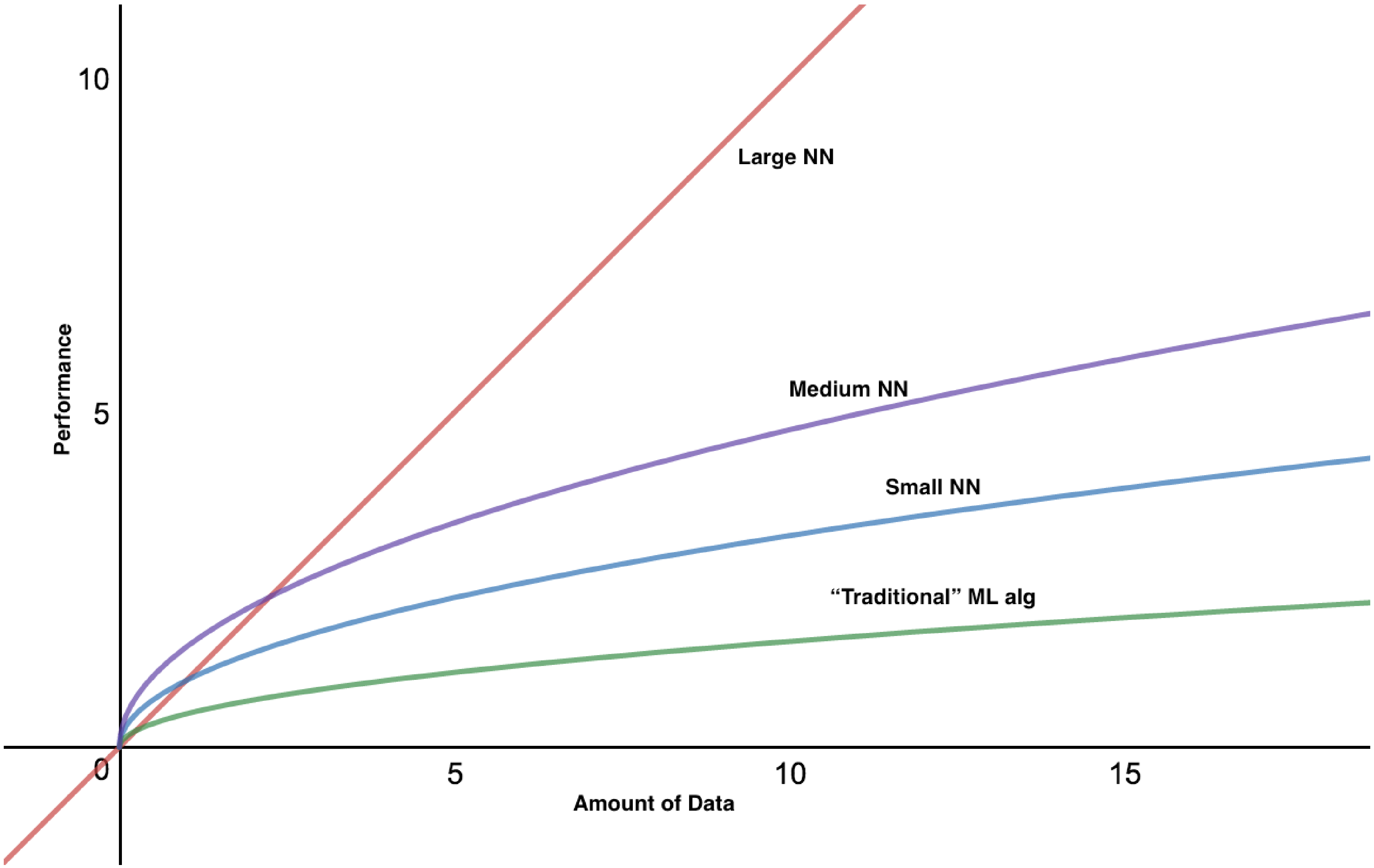

Other danger


Cyclist danger

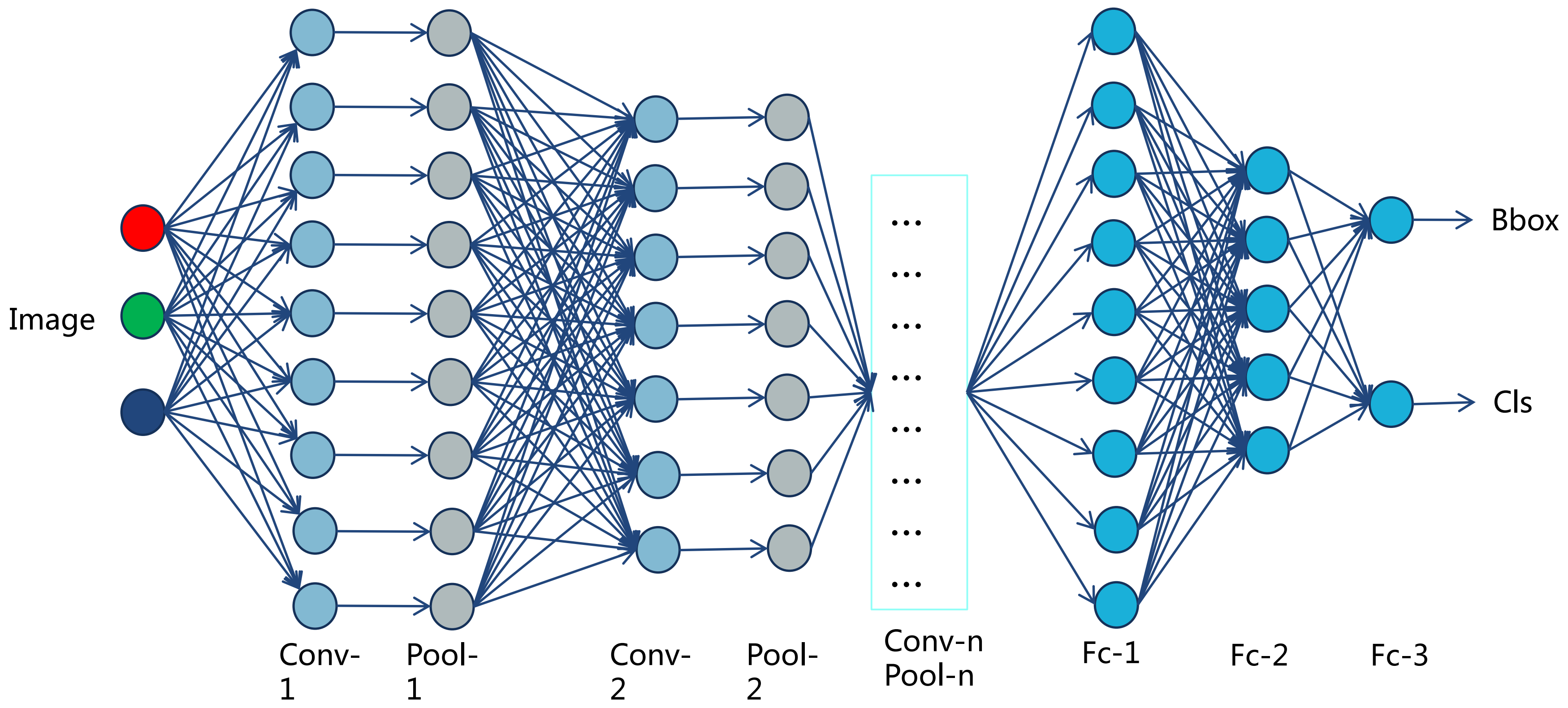

Turn around danger


Obstacle danger

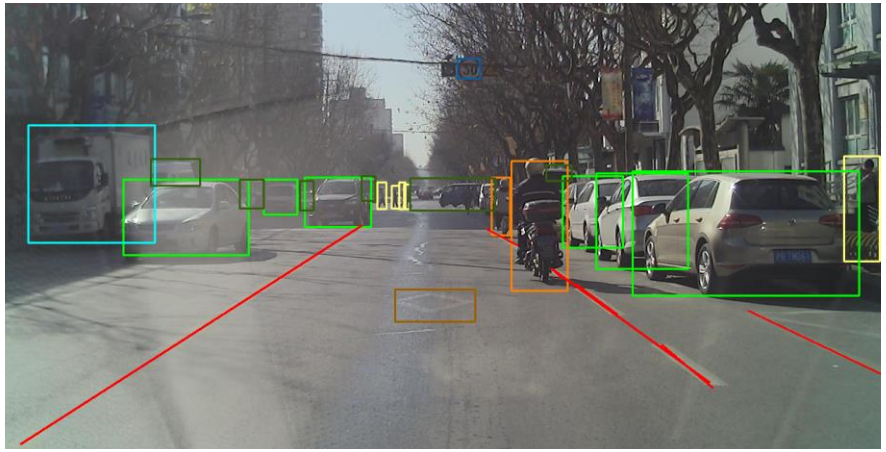
Deep Learning Comes to the Rescue



What is Deep Learning?



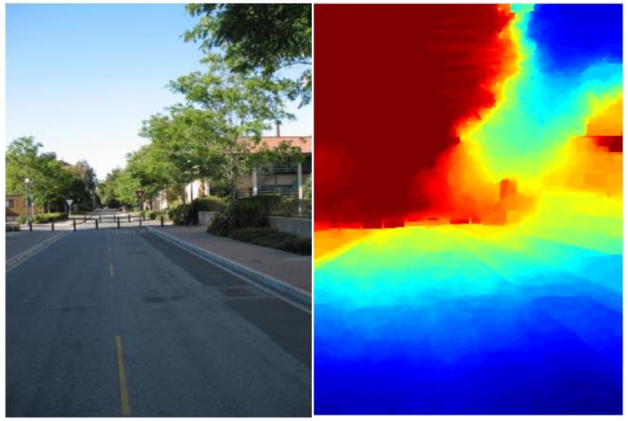
Applications of Deep Learning in Autonomous Driving



Object Detection

Object Classification

Semantic Segmentation

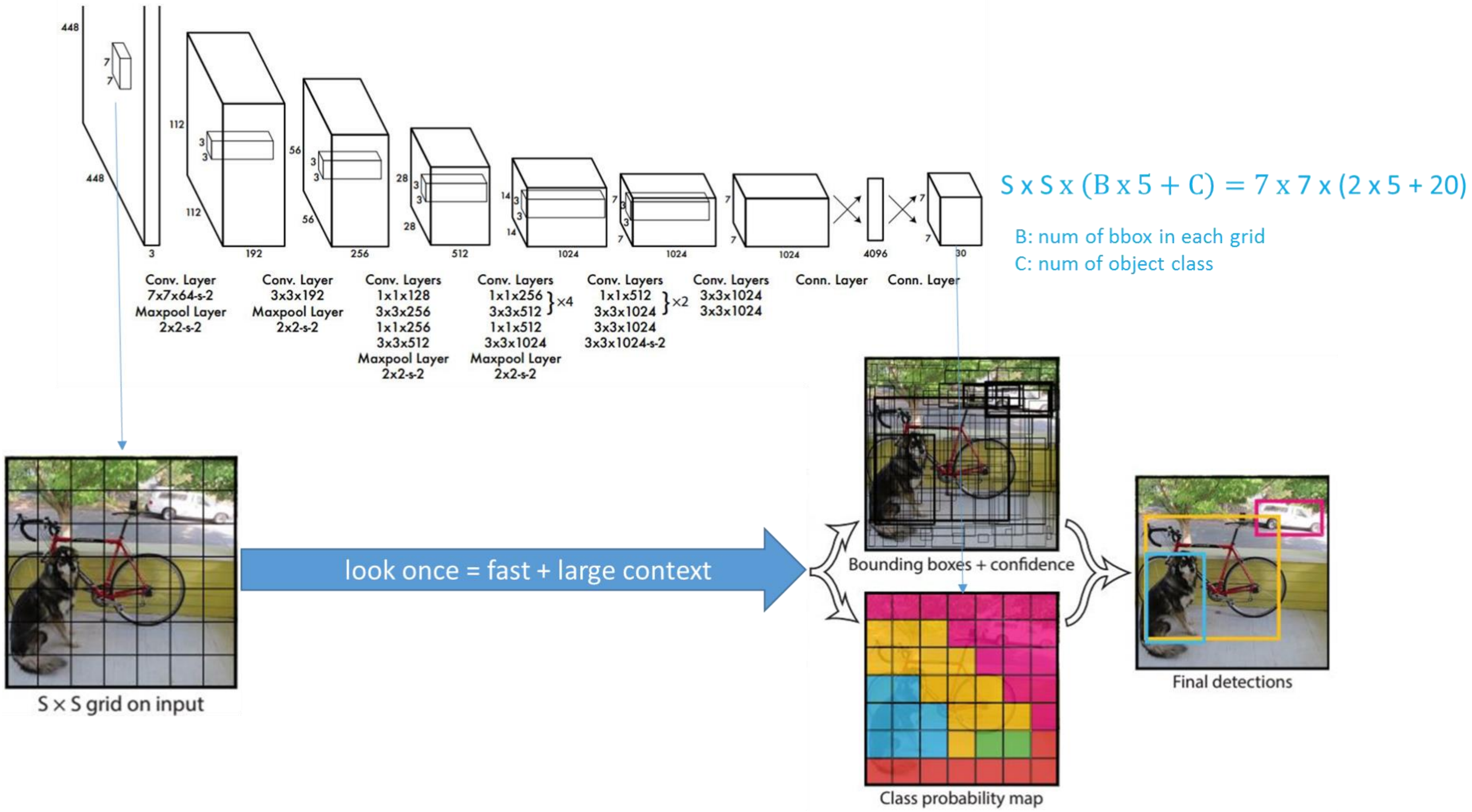


Distance Estimation

Deep Learning based Semantic Segmentation

Characteristics of Deep Learning Computation

- Single task
- High computing density
- Fast evolution of deep neural network models

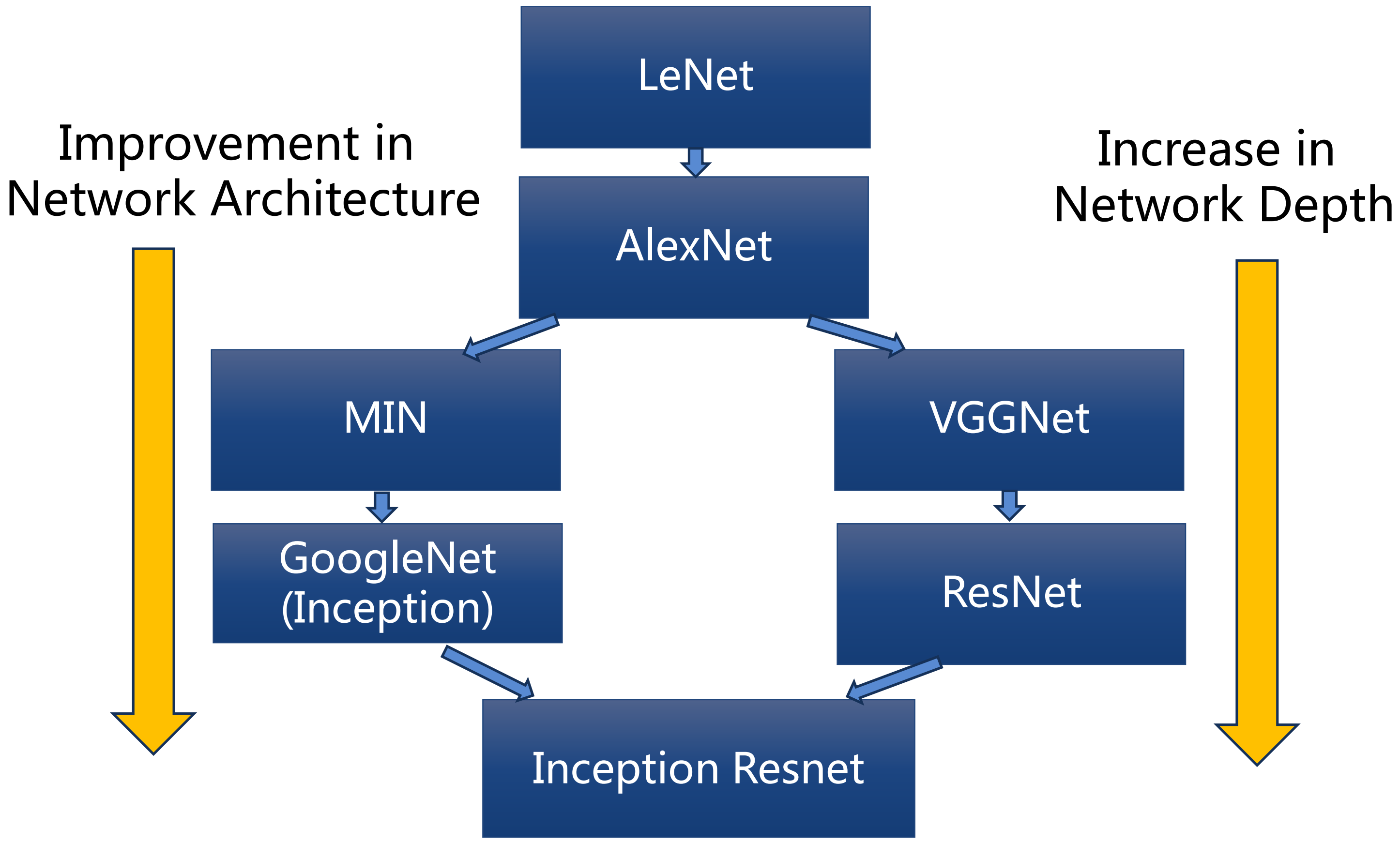


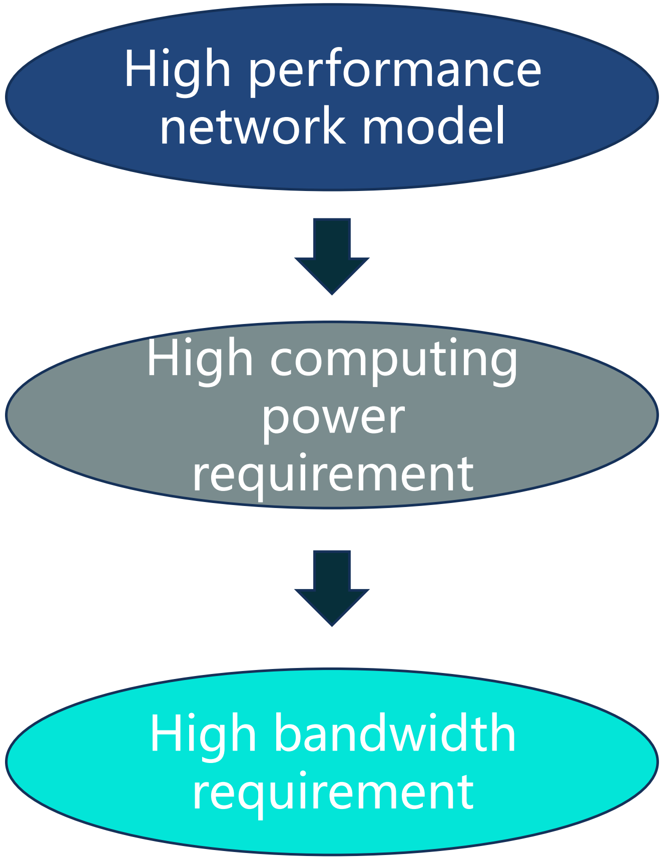
yolo

The basic module of CNN is a convolution stream, which consists of four parts: convolution, pooling, nonlinear, and batch normalization.

Challenges in Selecting Deep Learning Computing Platform

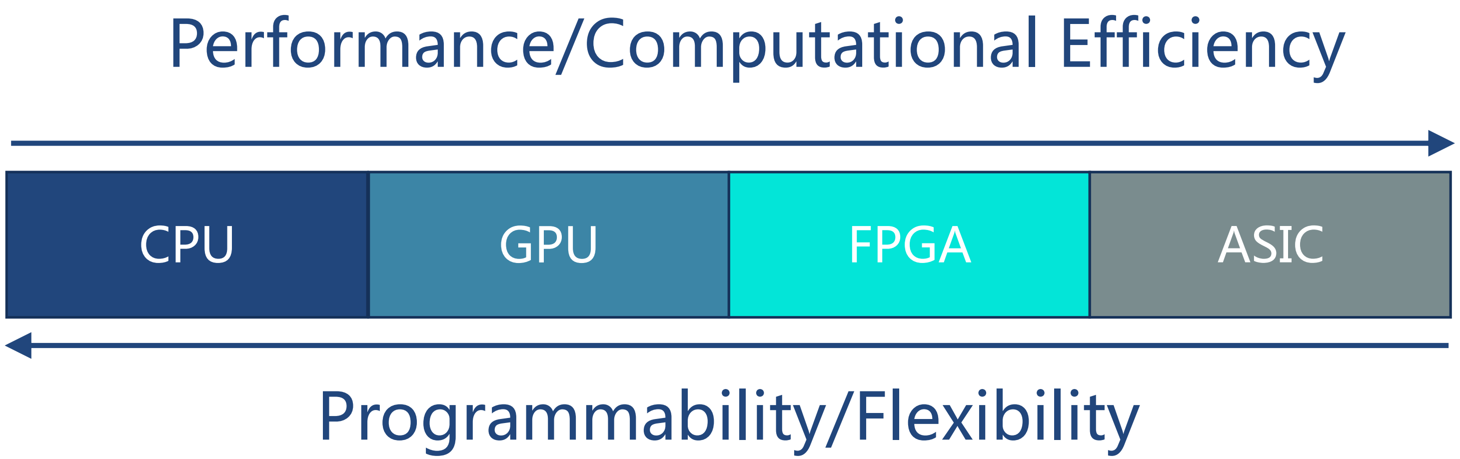
- High-performance network model -> increase in computing power
- Increase in computing power -> increase in bandwidth demand





Embedded Computing Platform for Deep Learning

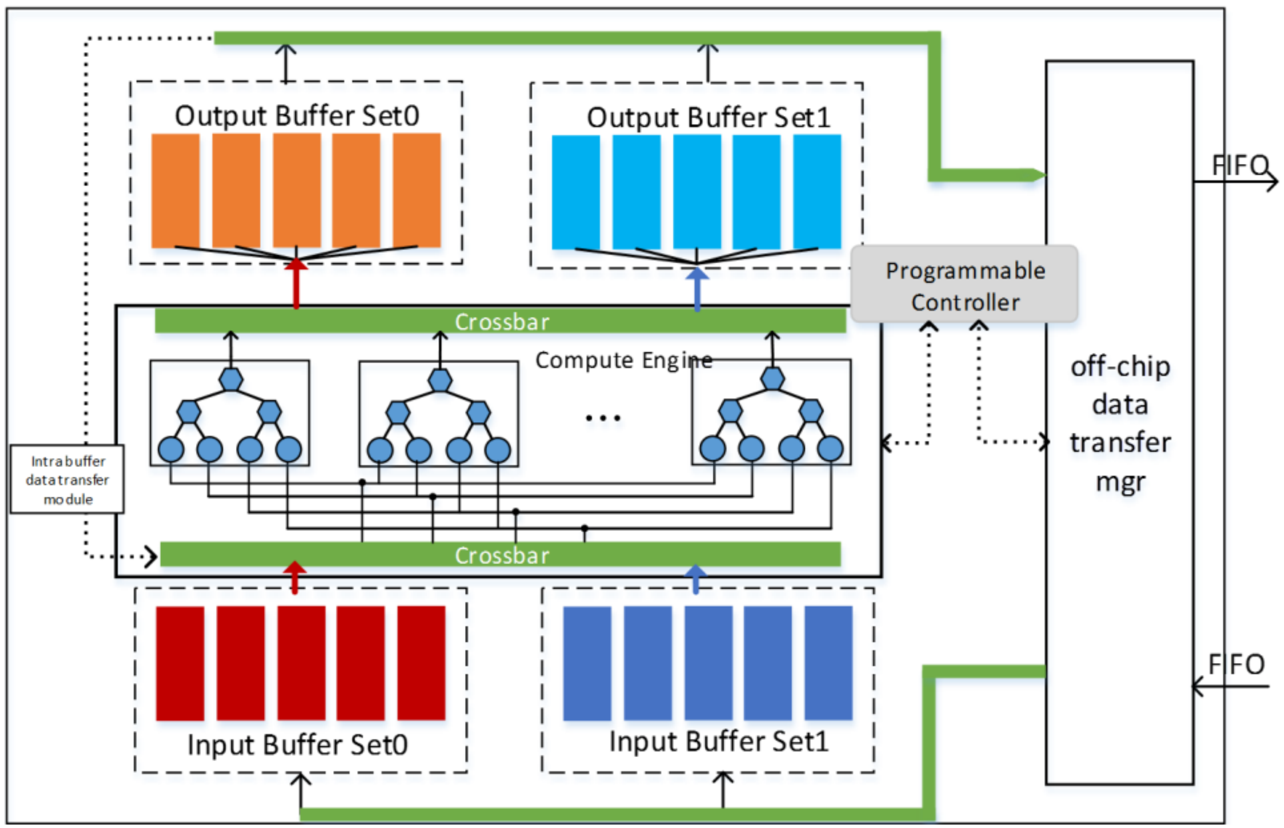
- CPU/GPU：Low power efficiency
- ASIC：High development cycle, low flexibility
- FPGA：Parallel Computing, high power efficiency and flexibility, low development cycle



Deep Learning on FPGA

魔视智能
MOTOVIS

- Each compute engine includes convolution calculation, addition tree, non-linear calculation, and pooling calculation
- Pipeline design, input and output using ping-pong buffer cache mechanism to reduce DDR access latency
- Supports dynamic configuration of fixed point precision



Model Optimization

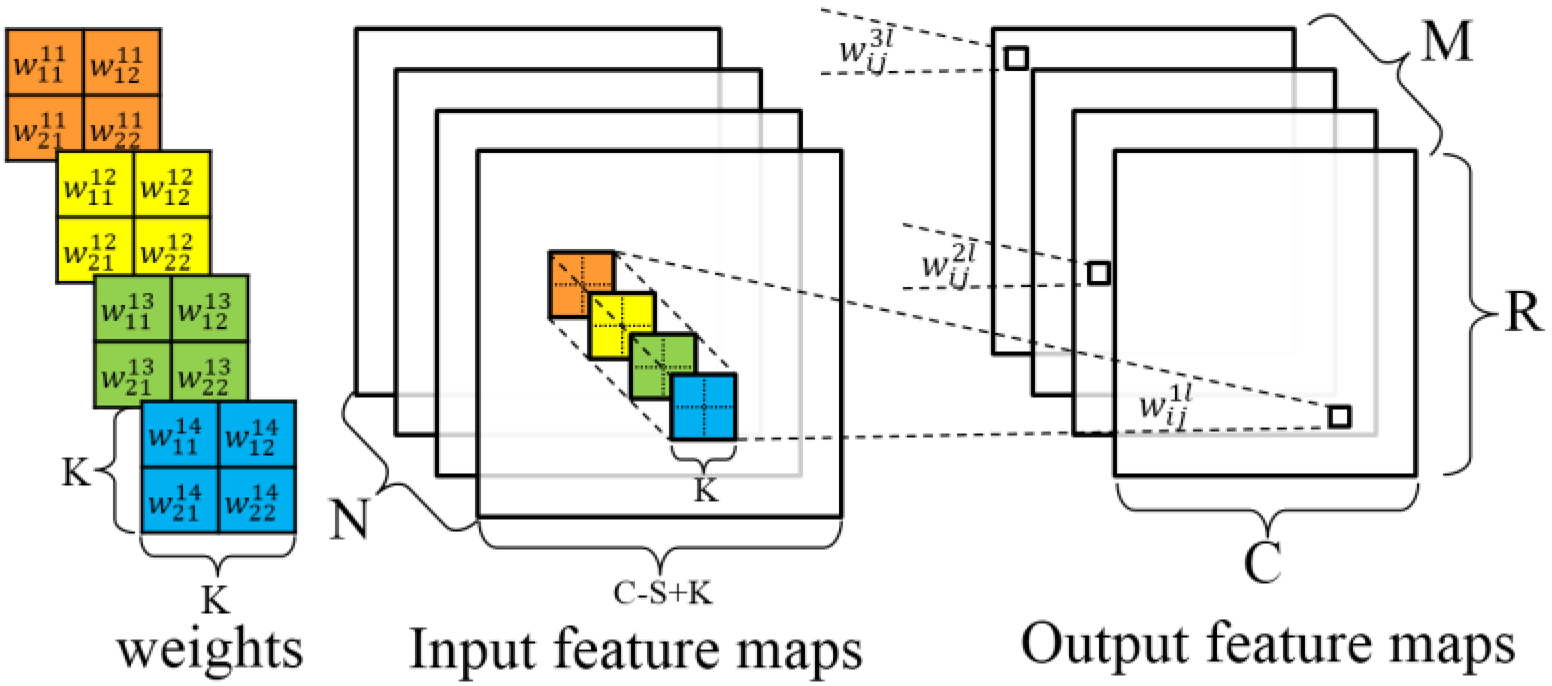
- Group convolution： Grouped convolution, the amount of calculation is reduced to 1/groups
- 3x3 kernel size： Use a combination of two 3 × 3 convolution kernels instead of one 5 × 5 convolution kernel for better results
- Insert a 1×1 convolution kernel: reduce the amount of convolutional layer parameters
- Dilated convolution： Allows a fixed-size convolution kernel to see a larger area

Model Compression

- Pruning
  - Remove low-contributing neurons or connections
- Sparse Optimization
  - Ensure that all data read into and into the compute module from the cache is valid data to avoid a large number of useless zero elements occupying storage bandwidth and computing resources.
- Distillation compression
  - Use the output of the pre-trained complex model (teacher model) as a supervisory signal to train another simple model (student model)

Data Reuse

- CNN reuse: the entire Feature Map reuse of a set of convolution kernels, and the reuse of multiple sets of convolution kernels by a set of Feature Maps. When the above methods are used in combination, the data reuse rate can be greatly improved and the storage access bandwidth requirement can be reduced.



Bit-width Compression

- Low bit width means less power, bandwidth, and power consumption when dealing with the same task.
- When the high bit width is converted to the low bit width quantization, the loss of precision is inevitable. In this regard, the impact on accuracy can be reduced by quantization mode, adjustment of the representative range, encoding, and even increasing the depth of the model (binary network).

	Layer outputs	CONV parameters	FC parameters	32-bit floating point baseline	Fixed point accuracy
LeNet (Exp 1)	4-bit	4-bit	4-bit	99.1%	99.0% (98.7%)
LeNet (Exp 2)	4-bit	2-bit	2-bit	99.1%	98.8% (98.0%)
Full CIFAR-10	8-bit	8-bit	8-bit	81.7%	81.4% (80.6%)
SqueezeNet top-1	8-bit	8-bit	8-bit	57.7%	57.1% (55.2%)
CaffeNet top-1	8-bit	8-bit	8-bit	56.9%	56.0% (55.8%)
GoogLeNet top-1	8-bit	8-bit	8-bit	68.9%	66.6% (66.1%)

Fine-tuned networks with dynamic fixed point parameters and outputs for convolutional and fully connected layers. The numbers in brackets indicate accuracy without fine-tuning

Source: Gysel et al, HARDWARE-ORIENTED APPROXIMATION OF CONVOLUTIONAL NEURAL NETWORKS, ICLR 2016

Winograd Transform

魔视智能 MOTOVIS

- Reducing the number of multiplications by Winograd Transform
- But the larger the kernel size, the more complex the transformation. 3x3 convolution kernel is more suitable

F (2x2, 3x3), Direct CNN

Functions	Multiplier	Adders
Matrix mult	589824	393216

F (2x2, 3x3), Winograd CNN

	Functions	Multiplier	Adders
	Data Trans $[B^T dB]$		4096
	Filter Trans $[GgG^T]$		458752
$[U \odot V]$	element-wise mult	262144	
	Result Trans $A^T[U \odot V]A$		3072
	Total	262144	465920

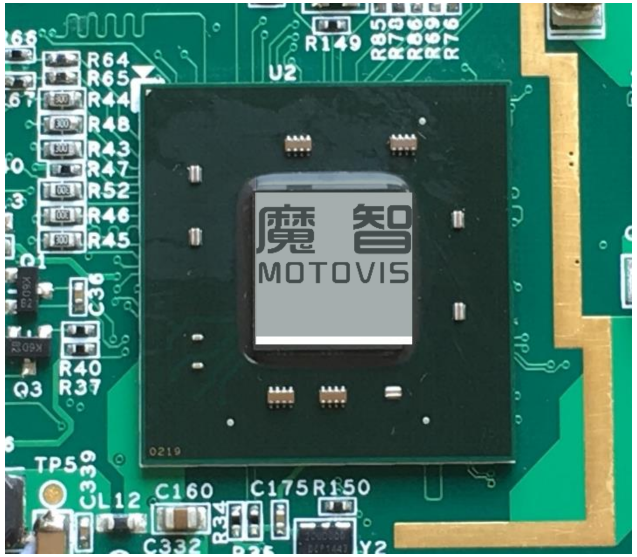
Winograd v.s. Direct:
2.1x improvement

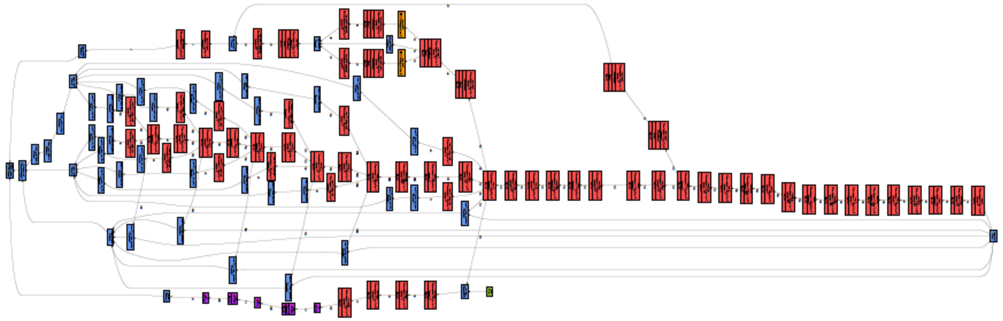
Automotive Grade Embedded Deep Learning Chip

魔视智能 MOTOVIS

- **Motovis's deep learning and FPGA based ADAS solution has entered mass production**
- **The product follows the automotive industry standard development process and the hardware is fully compliant with the ISO 26262 standard**

- Highly programmable, suitable for current and future deep learning networks
- High computing power, supports >100 layer deep learning network
- Highly optimized deep learning engine achieves 2.8X computing performance
- Compliance to ISO 26262 and other automotive industry functional safety regulations
- Low power consumption, low cost, high reliability

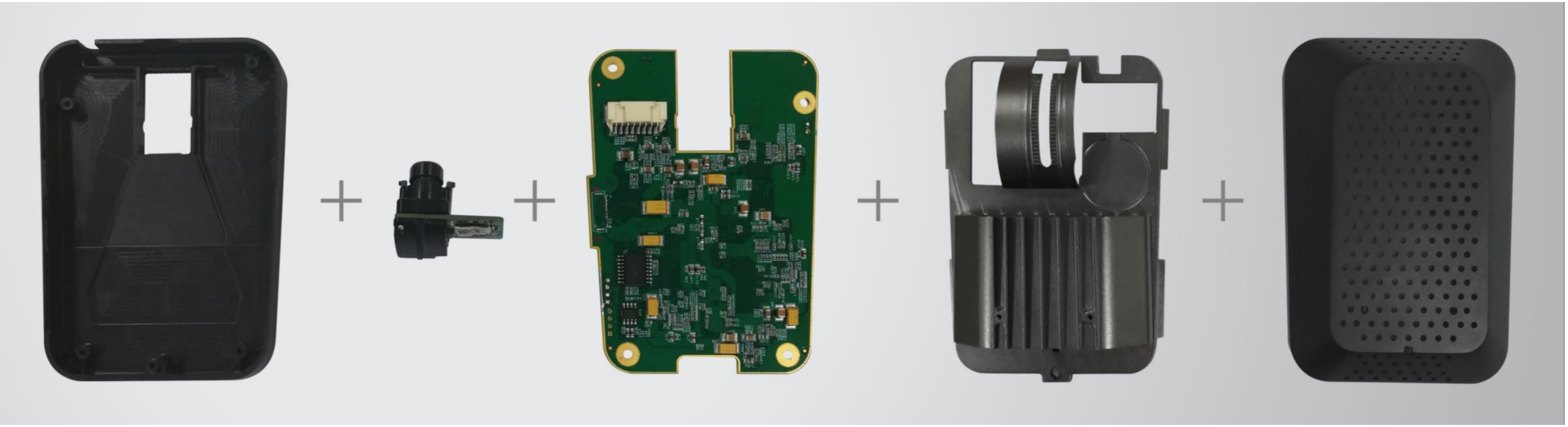




Certification Mark:

Product: Software Tool for Safety Related Development

Embedded Deep Learning based ADAS Product Design



System Performance - Highway, Daytime (Car, Bus, Traffic Lane, Curb, Freespace)

魔视智能
MOTOVIS

Validation of System - Functions

车道偏移预警
Lane Departure Warning (LDW)

前向碰撞预警
Forward Collision Warning (FCW)

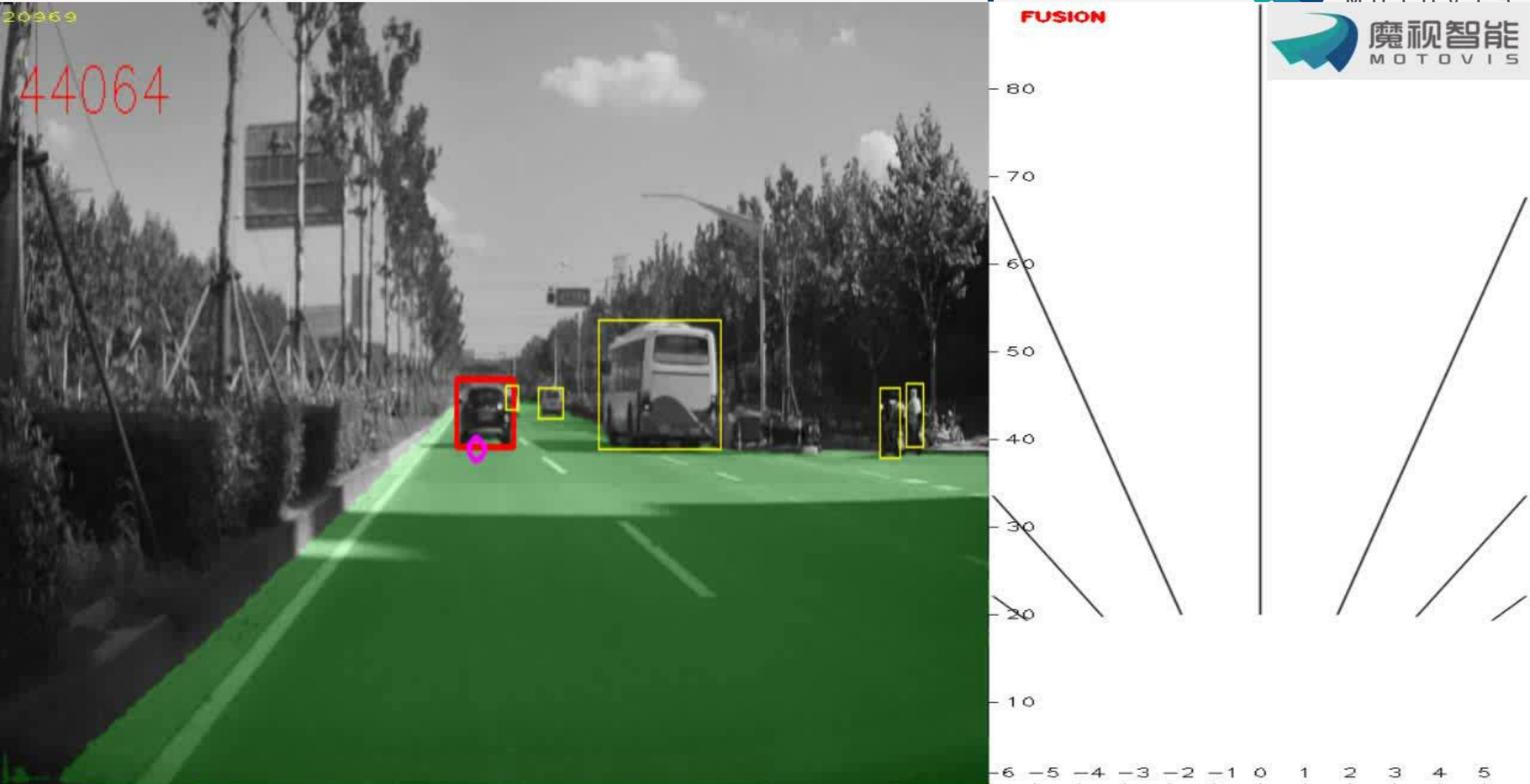
行人碰撞预警
Pedestrian Collision Warning (PCW)

前车距离监控
Headway Monitoring and Warning (HMW)

自动紧急制动
Automatic Emergency Brake (AEB)

Validation of Sensor Fusion for AEB: Radar Only

Sensor Fusion for AEB: Radar & Video (with Freespace)



Embedded Deep Learning for Autonomous Driving

Automotive Standard — Following Automotive Standard & Process

Computing Chip — Embedded Automotive High Performance Computing Platform

Algorithm — High Accuracy Deep Learning Algorithm

Data — Hugh Amount of Training & Testing Data

Adaptable.
Intelligent.

魔视智能
MOTOVIS