北京大学数字媒体研究所
INSTITUTE OF DIGITAL MEDIA, PEKING UNIVERSITY

Digital Retina – Improvement of Cloud Vision System from Enlighten of HVS Evolution

Wen Gao

Peking University，and
Peng-Cheng Lab

Outline

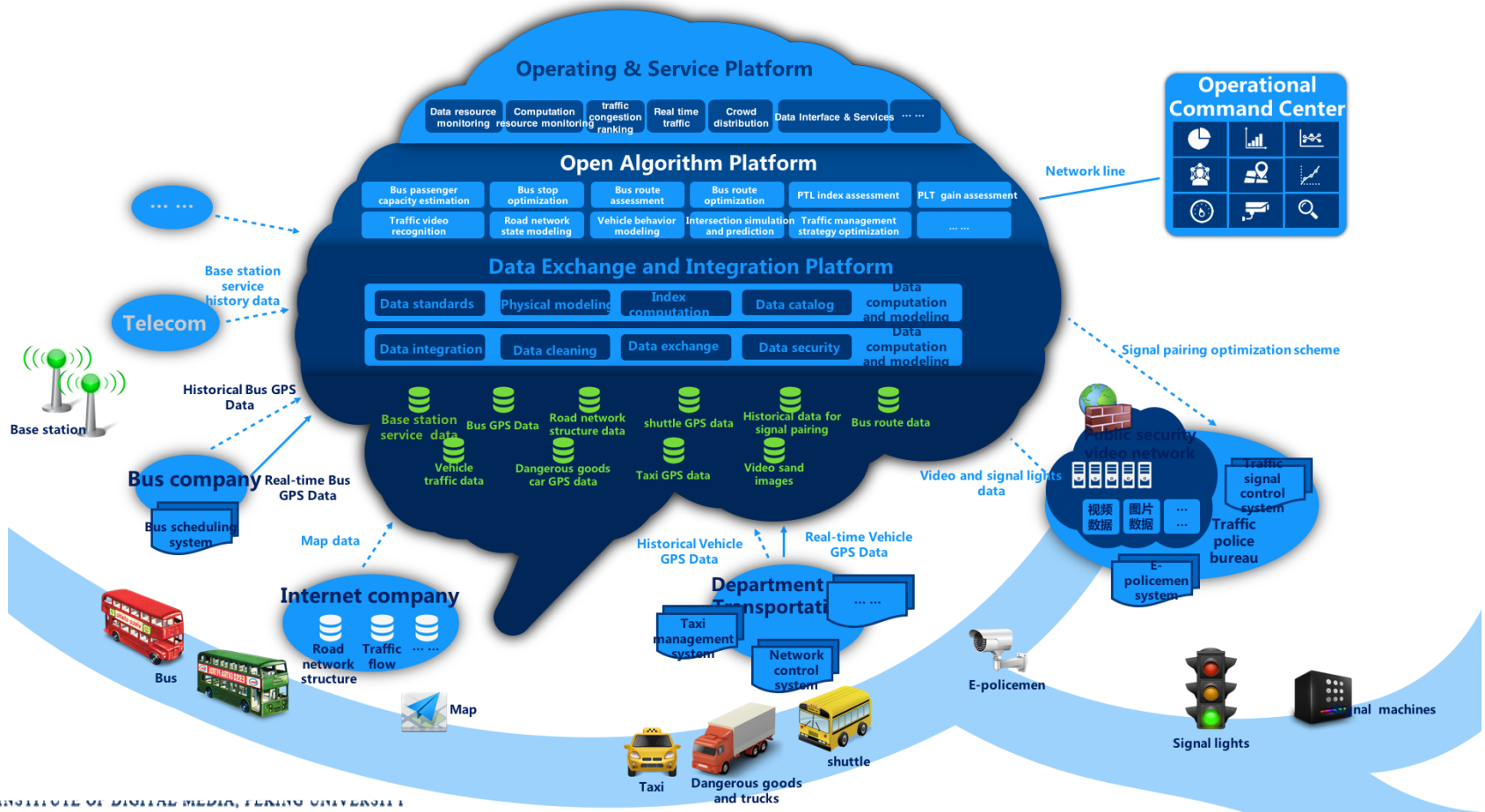
- CVS and its challenge
- What we learn from HVS?
- Digital retina – a way to make CVS more efficient
- Summary

Outline

- CVS and its challenge
- What we learn from HVS?
- Digital retina – a way to make CVS more efficient
- Summary

What is CVS

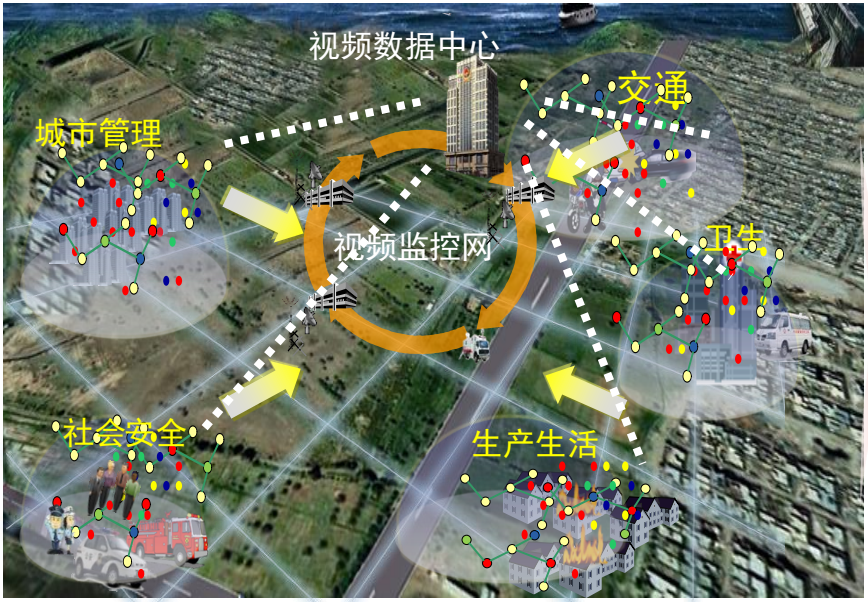
- A computer vision system in cloud, connected with a camera network system

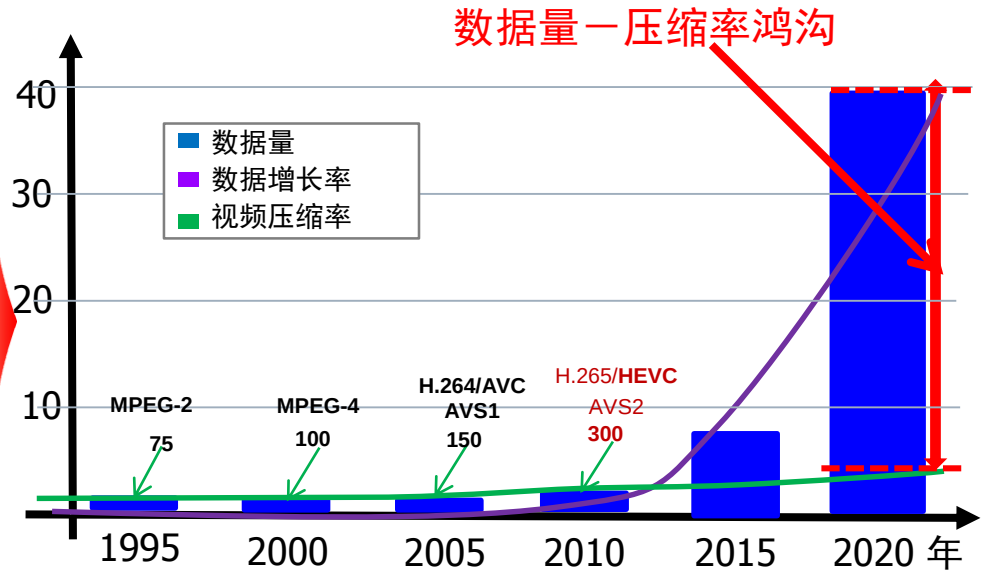


Typical processing paradigm

- ✓ 视频存储、分析与识别均在云上完成
- ✓ 分析识别的视频路数与云服务器的算力成正比

Challenge 1: large data ≠ big data





视频大数据：数据量巨大←→存储分散

数据量两年增长一倍←→压缩率十年增长一倍

一个城市
10万个
摄像头

用于存储

…×10Mbps×1Month
= 10 EB

…×5Mbps×1Month
=5EB

用于传输

…×10Mbps
= 1 Tbps

…×5Mbps
=500 Gbps

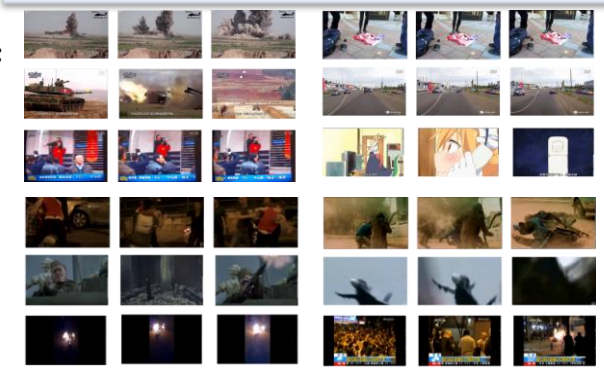
H.264/AVC

H.265/AVS2

数据大
≠
大数据

Challenge 2: Huge Volume vs Low Value

敏感视频：特殊场景或行为，涉及暴恐、突发热点、群体事件等



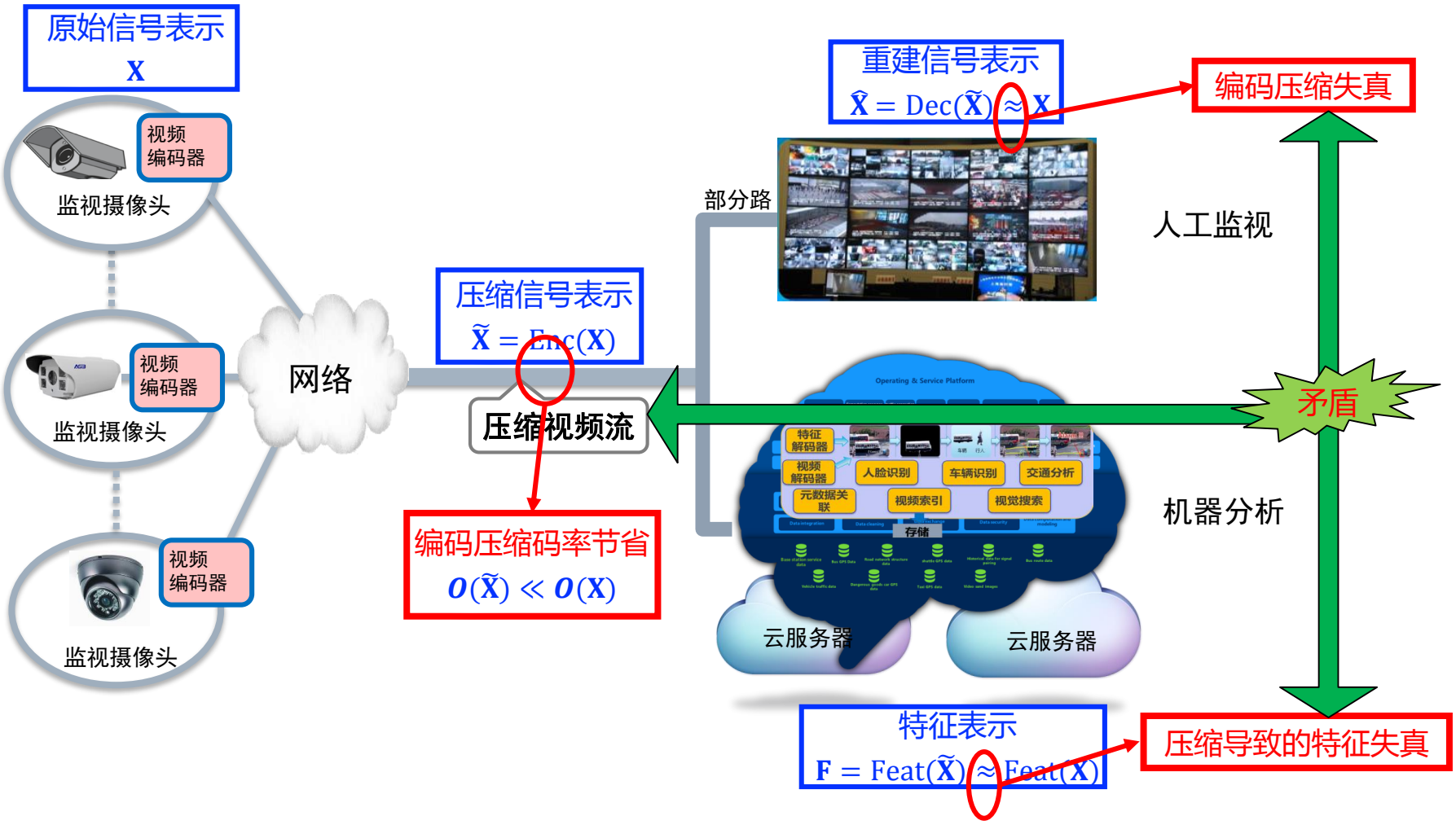
重点目标/异常事件：
价值密度高

大量正常视频→低价值密度

少量敏感视频→高价值密度

正常视频：各种日常场景，涉及对象外观、姿态、尺度、视点、复杂背景、光照条件、行为等变化

Where is the problem? Data in unstructured



Computer Vision System = VCC

- Scope of VCC(Visual Computing and Cognitive)
  - Concept
  - Sensor/Sensor net
  - Feature/Feature net
  - Modeling
  - Classification
  - Identification
  - Object Recognition
  - Understanding
  - …

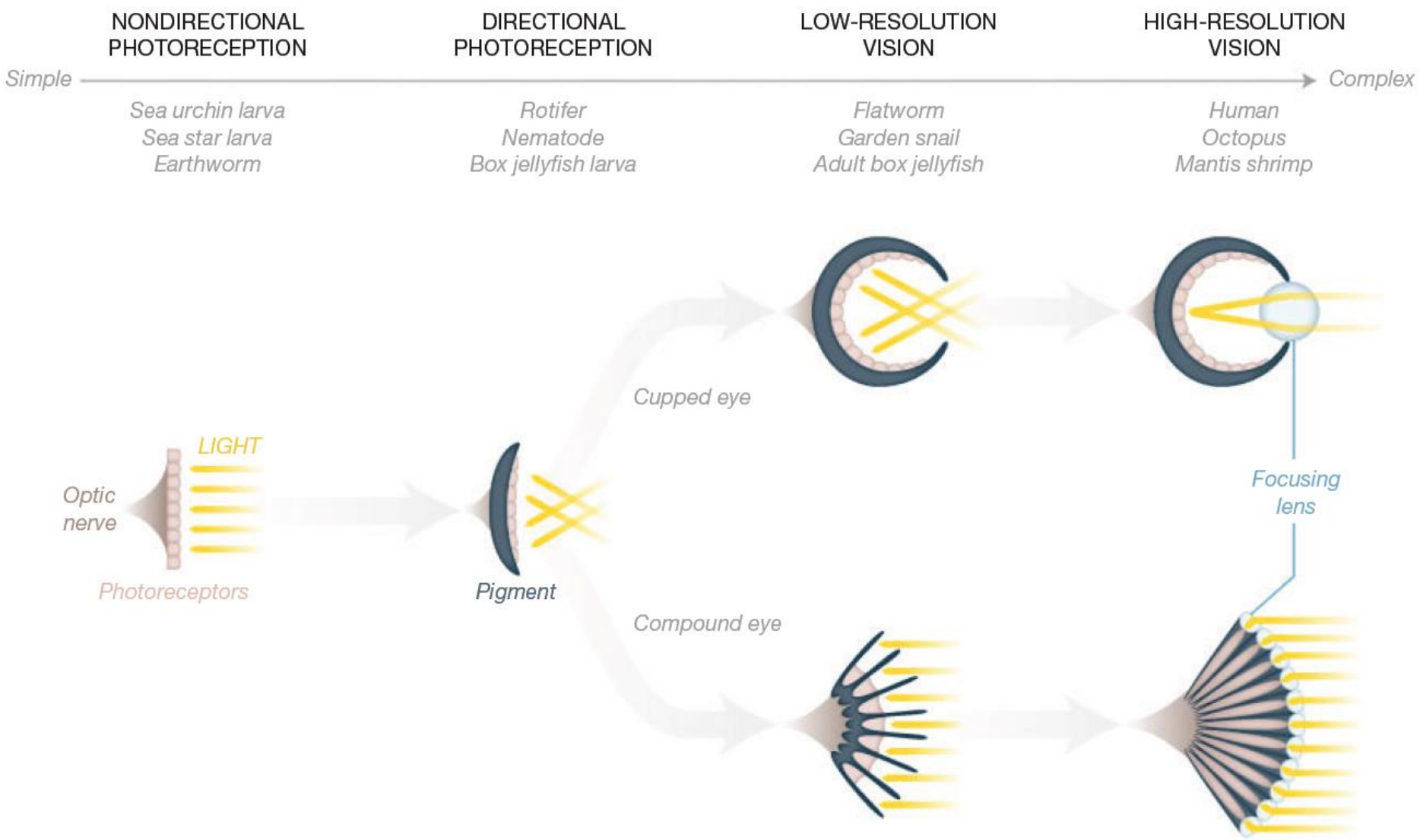
VCC vs HVS

- Engineering VCC system
  - Image caption by camera
  - Feature extraction
  - Pattern recognition
  - Scene understanding
  - …
- Human Vision System
  - Eye(Sensor) modeling
  - Visual coding/path by NN
  - Object recognition by NN
  - Scene understanding by attention/context/…

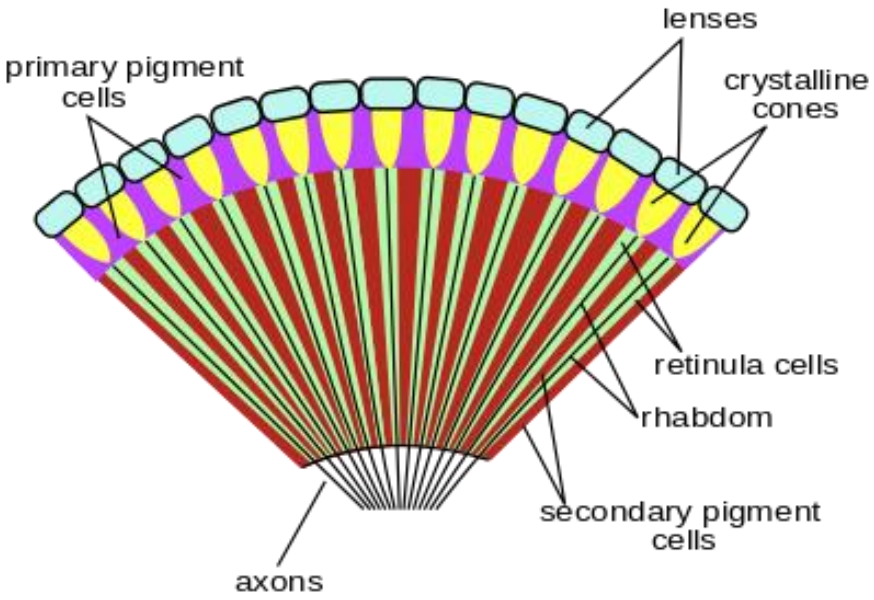
Outline

- CVS and its challenge
- What we learn from HVS?
- Digital retina – a way to make CVS more efficient
- Summary

Evolution of Eyes



Compound Eye

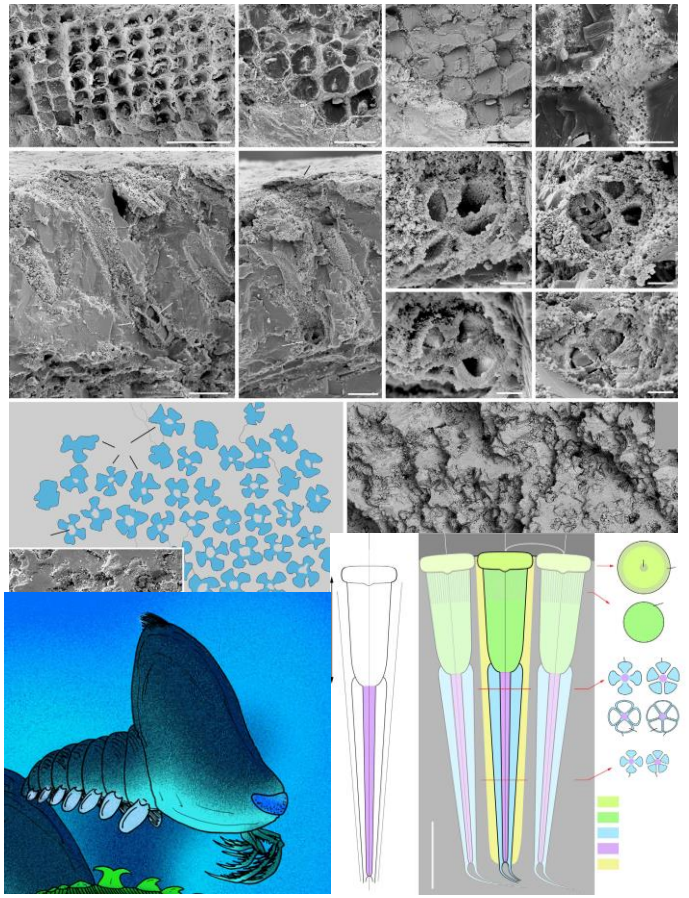


Early Compound Eye

Ed Yong. *Seeing the Light*. National Geographic. Februray, 2016

Near Modern Compound Eye

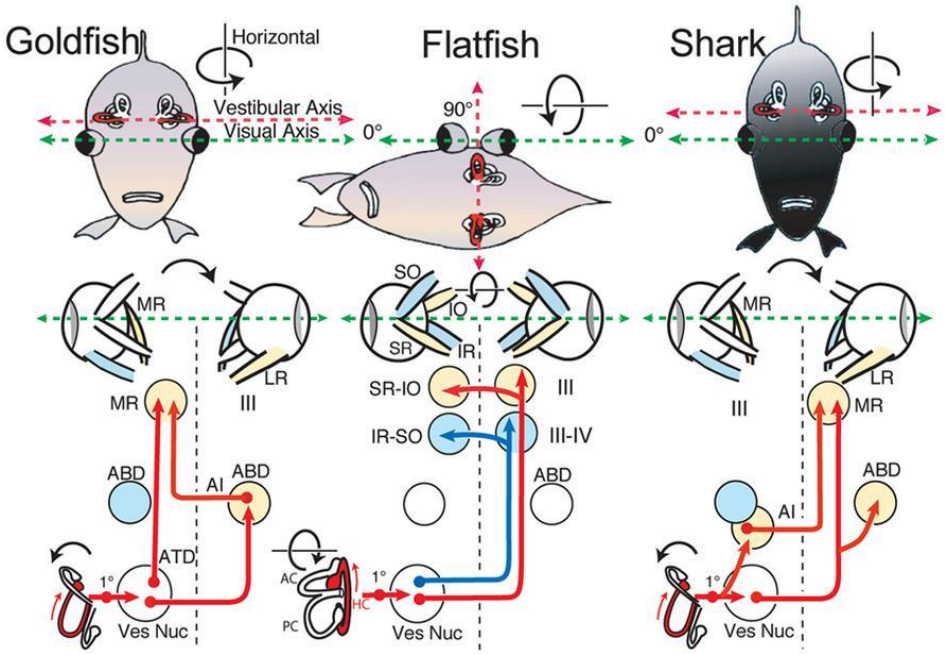
- A 160-million-year-old thylacocephalan arthropod had about 18,000 lenses on each eye, which is surpassed only by modern dragonflies.



Jean Vannier, Brigitte Schoenemann, Thomas Gillot, Sylvain Charbonnier, Euan Clarkson. Exceptional preservation of eye structure in arthropod visual predators from the Middle Jurassic. Nature Communications, Nature Publishing Group, 2016, 7, pp.10320.

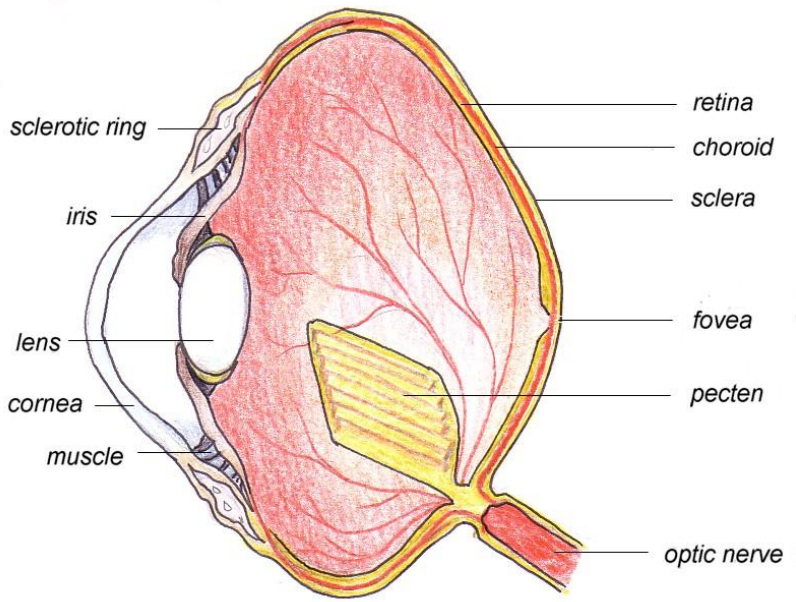
Fish eye

- Fish eyes are similar with terrestrial vertebrates, like birds and mammals, but have a more spherical lens.

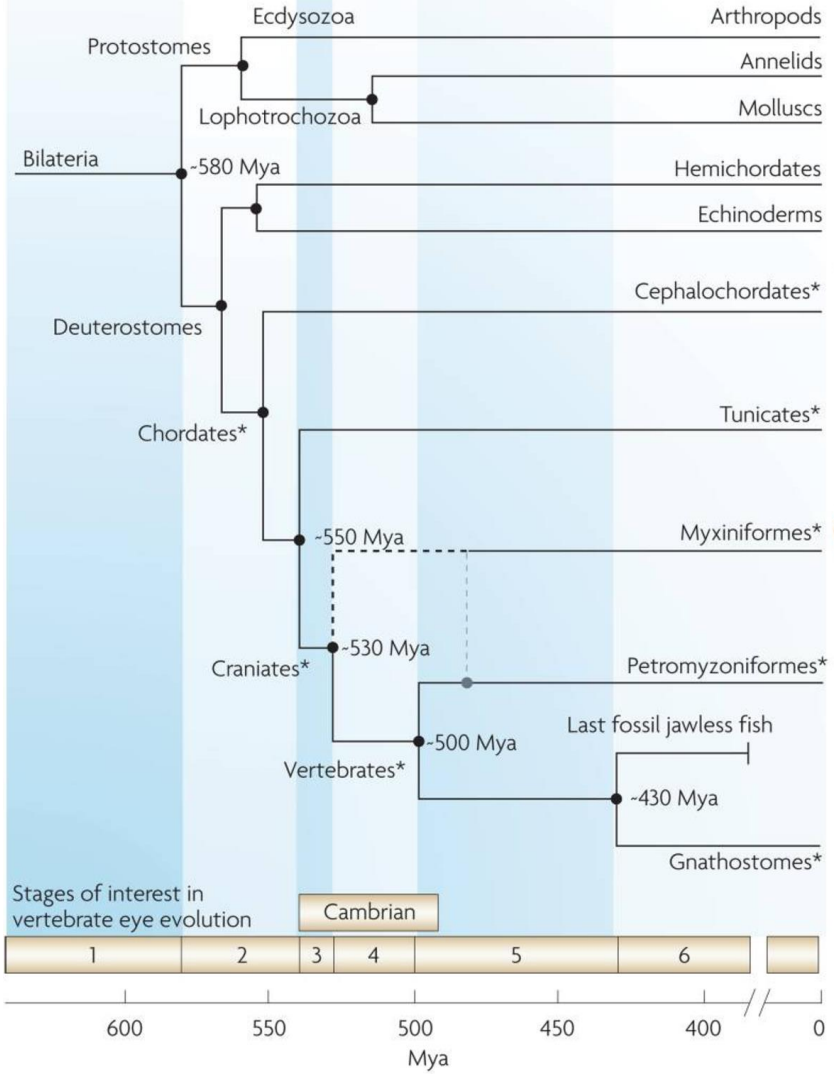


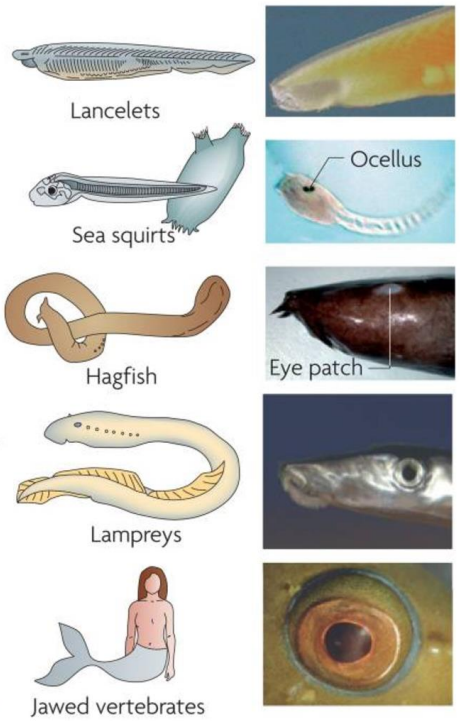
Bird eye

- The main structures of the bird eye are similar to those of other vertebrates
- The eye of a bird most close to the reptiles



Vertebrate eye evolution





1（无眼古生物）

2 Lancelet（文昌鱼）：
光感蛋白

3 Sea-squirts(喷射类)：
Ocellus（眼点）

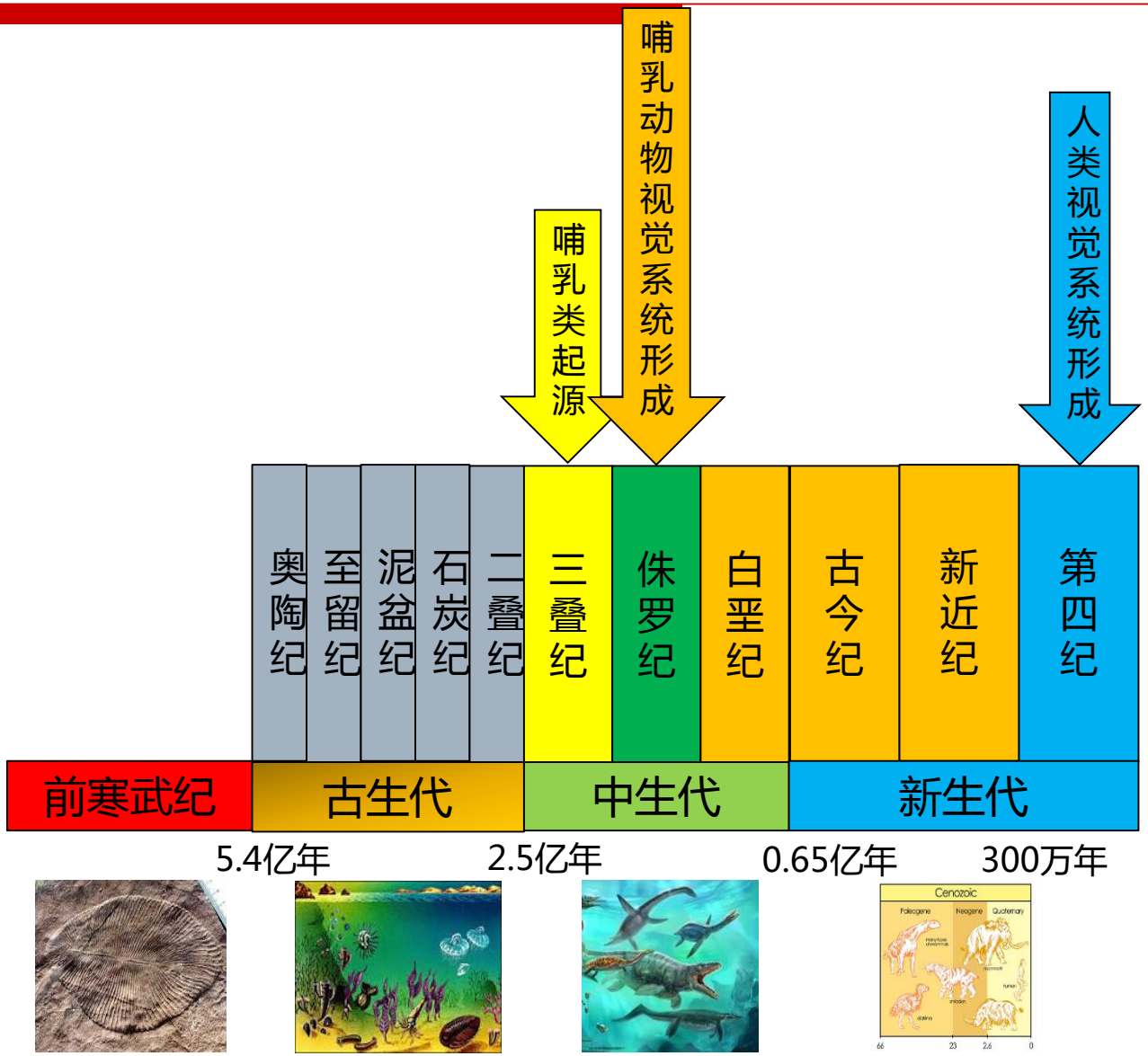
4 Hagfish
视网膜两层神经元

5 Lampreys（八目鳗）
视网膜三层神经元
眼睛基本进化完成

6 有颌脊椎动物到人

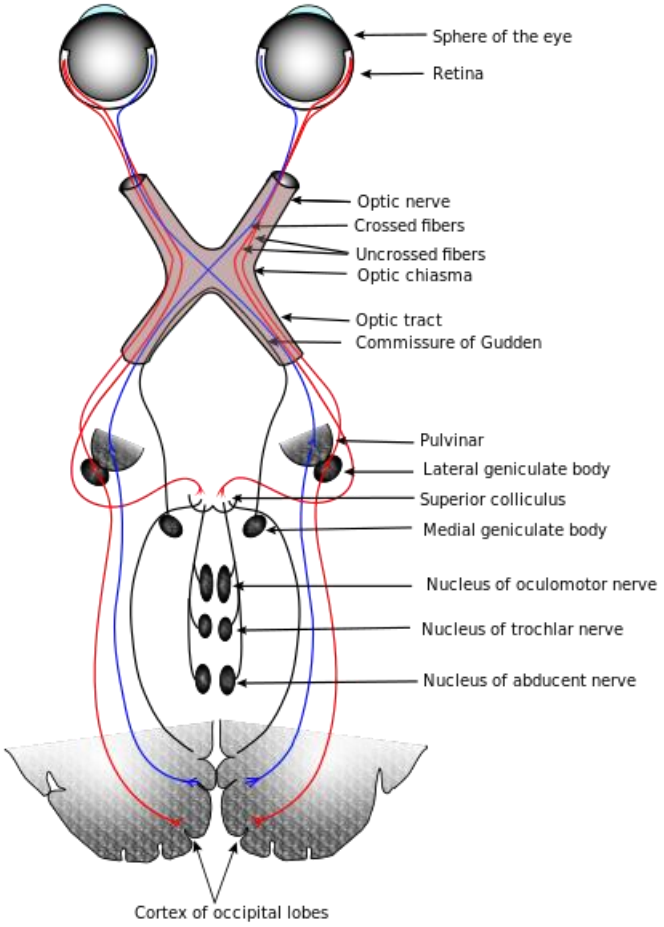
Lamb T, Collin SP, Pugh EN (2007) Evolution of the vertebrate eye: opsins, photoreceptors, retina and eye cup. Nature Rev Neurosci 8(12): 960-976

脊椎动物视觉系统进化简史

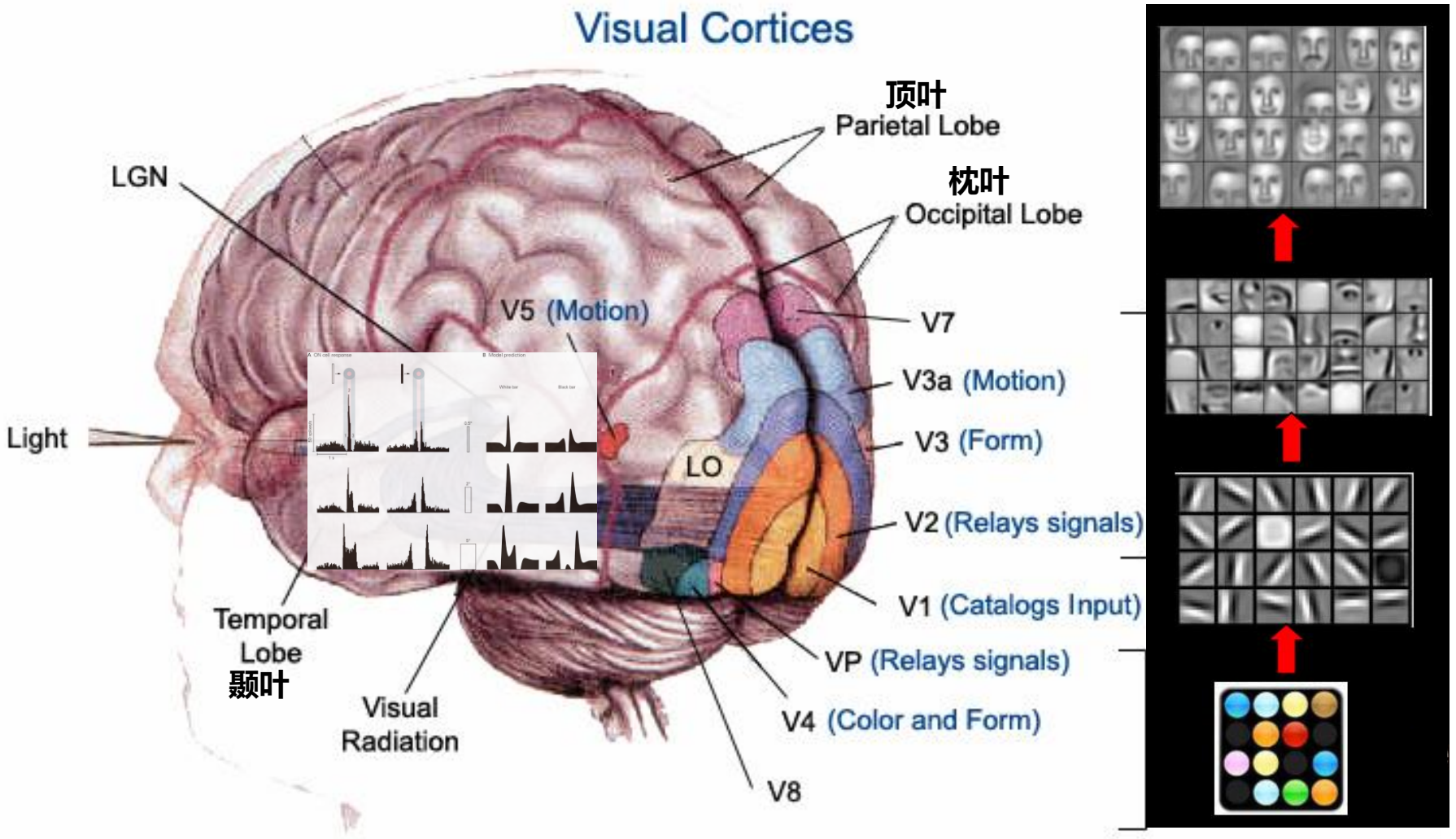


Biological Vision System

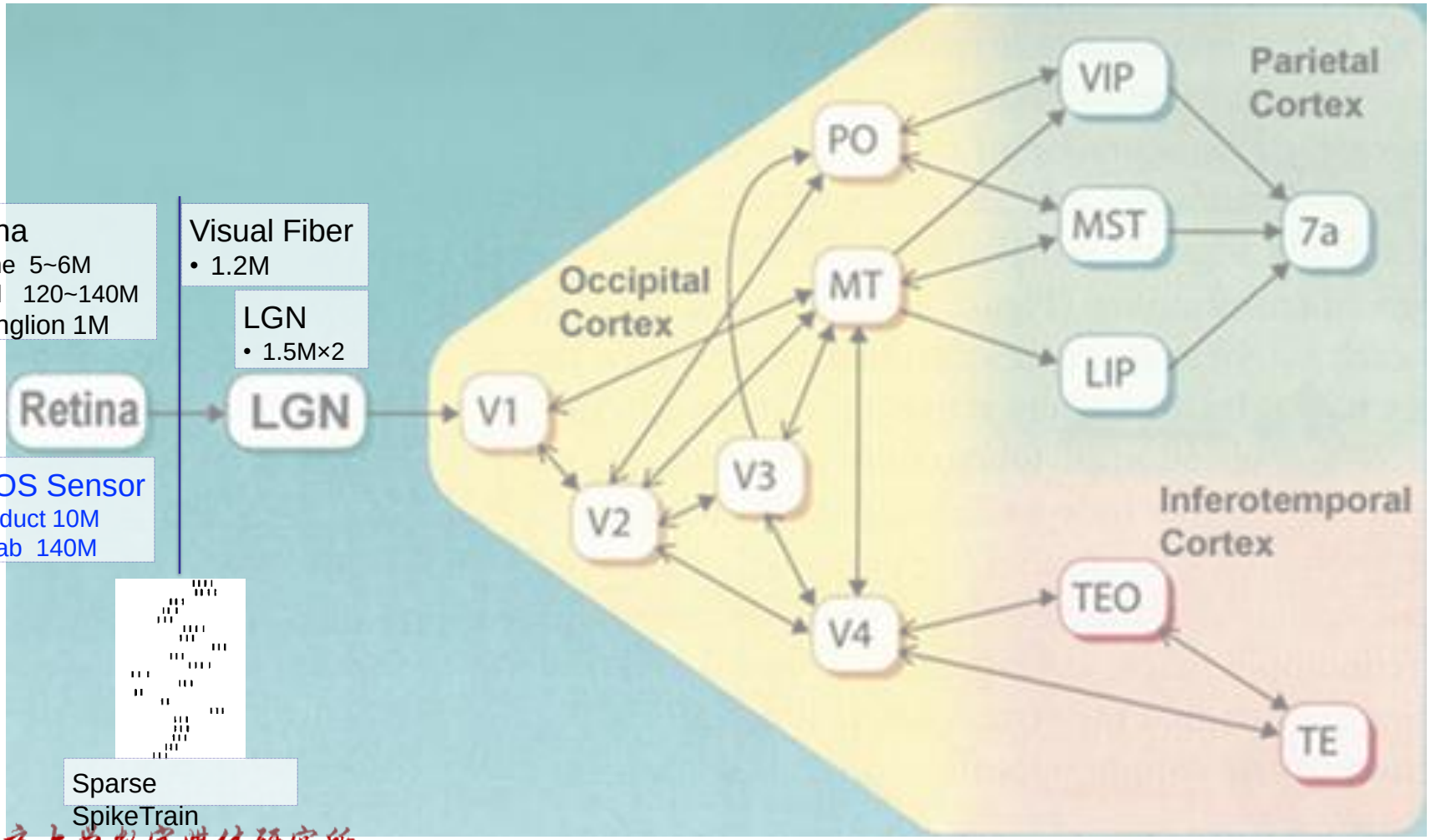
- The part of the central nervous system which gives organisms the ability to process visual detail
  - Eyes
  - Pathways
  - Visual field of the brain
- Most important sense part in the body
  - Enemy
  - Food
  - …



How the Retina Codes the Image?

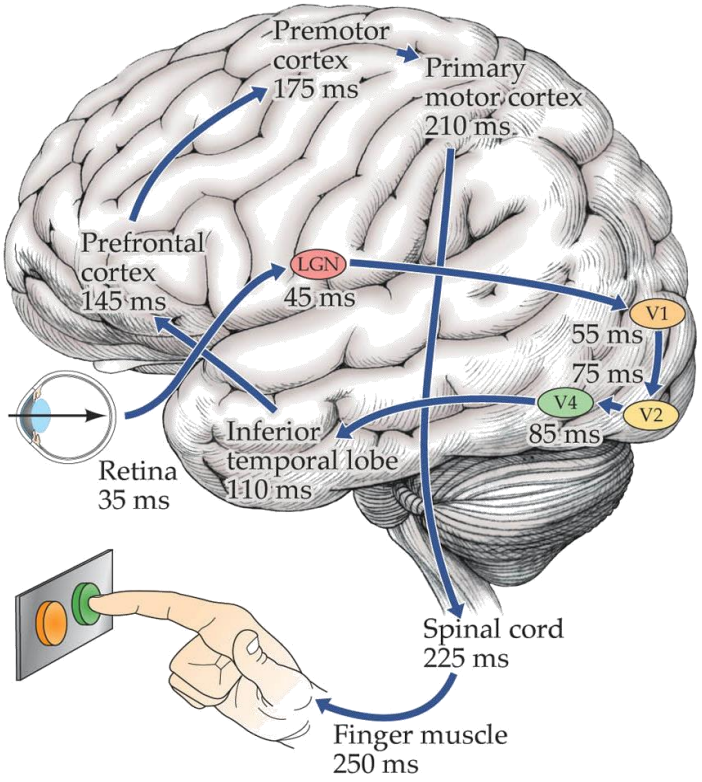


How We See?


Retina
• Cone 5~6M
• Rod 120~140M
• Ganglion 1M
Visual Fiber
• 1.2M
LGN
• 1.5M×2
Retina
LGN
CMOS Sensor
• product 10M
• in lab 140M
Sparse
SpikeTrain
Occipital
Cortex
V1
V2
V3
V4
PO
MT
VIP
MST
LIP
Parietal
Cortex
7a
Inferotemporal
Cortex
TEO
TE

Where is the Image/Picture?

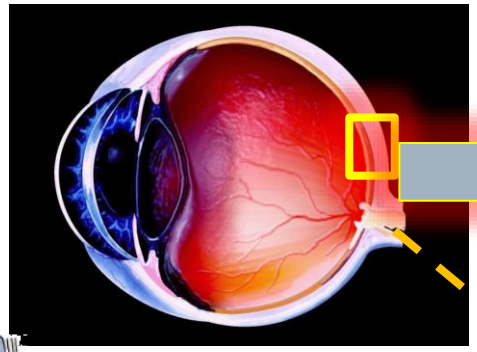
- It is believed the clear image on the outside world is reconstructed in the first 50ms after the optical stimulus
  - 0ms: photoreceptors output
  - 20ms: Retina
  - 30ms: LGN
  - 40ms: V1 (orientation-selective response)
  - 50ms: V1 (temporary memory)
  - 80-100ms: IT (Face-selective response)
  - 160-220ms: objects recognition(animal, food,…in category)
- More and more details known on what happens in the retina and primary visual system, but a whole picture and model is absent (what happens?)

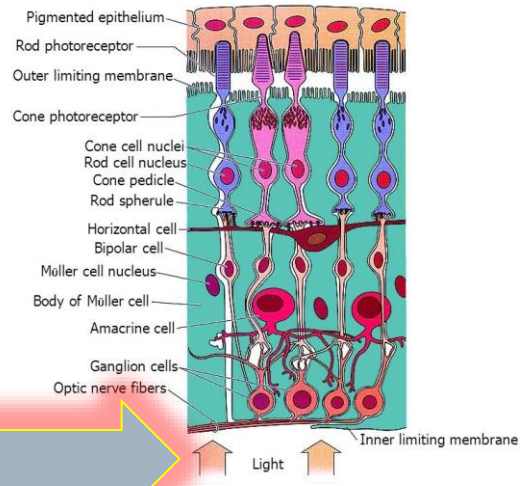


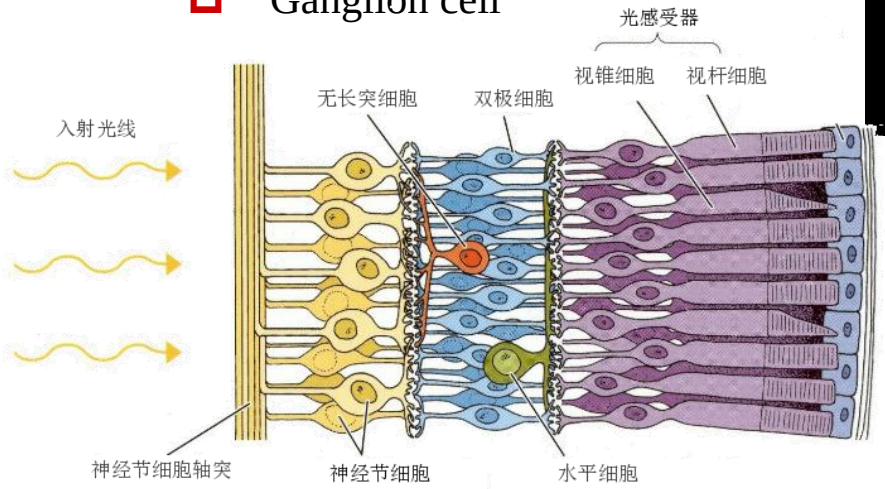
Maunsell and Gibson 1992; Raiguel et al. 1989; Nowak et al. 1995; Schmolesky et al. 1998; Thorpe, Fize& Marlot 1996

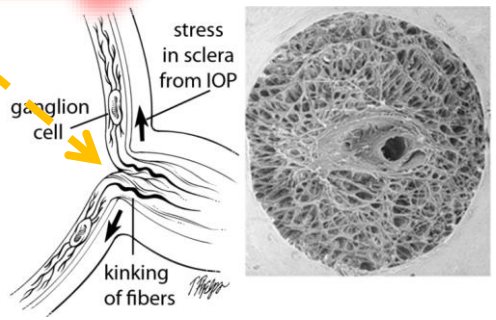
Retina: a high resolution sensory system

- 视网膜结构(Retina Structure)
  - 感光细胞(126M photoreceptor cell)
    - 锥状细胞Cone Cell，600多万个（RGB各1/3）
    - 杆状细胞Rod Cell，1.2亿个
  - 双极细胞
    - Bipolar cell
  - 神经节细胞
    - Ganglion cell

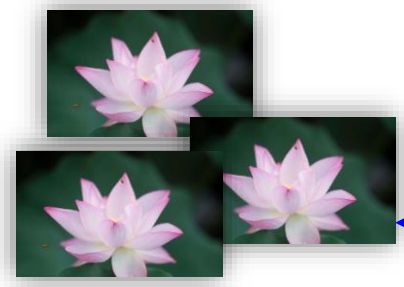




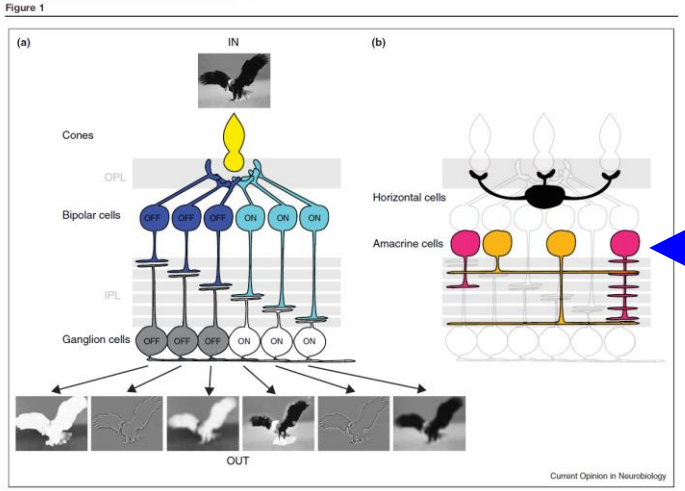




Retina: also feature extraction



视网膜中央凹具有高的空间分辨率，可捕捉细节



ELSEVIER
Available online at www.sciencedirect.com
ScienceDirect
Current Opinion in Neurobiology

Cell Types, Circuits, Computation
Rava Azeredo da Silveira[1,2] and Botond Roska[3]

视网膜外周具有高的时间分辨率，可捕捉快速运动信息，可提取并编码场景或物体的特征，如纹理、轮廓等

视网膜外周

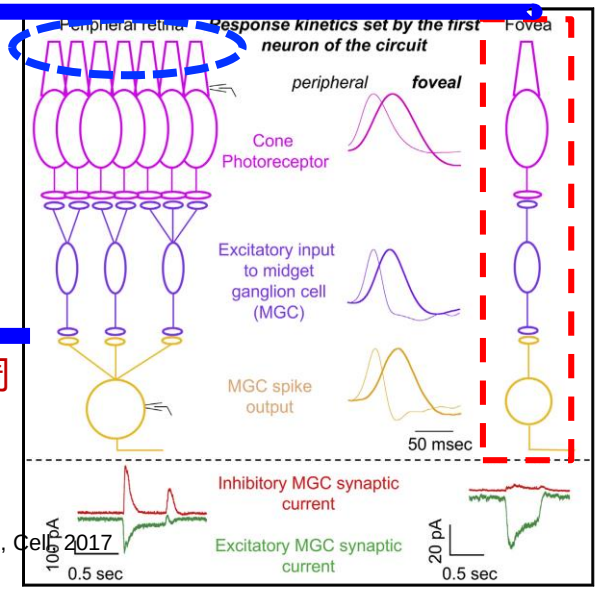
Article

Cell

Cellular and Circuit Mechanisms Shaping the Perceptual Properties of the Primate Fovea

中央凹

Graphical Abstract



Sinha,et.al., Cell 2017

Authors
Raunak Sinha, Mrinalini Hoon, Jacob Baudin, Haruhisa Okawa, Rachel O.L. Wong, Fred Rieke

Correspondence
rsinha@uw.edu (R.S.), mhoon@uw.edu (M.H.), rieke@uw.edu (F.R.)

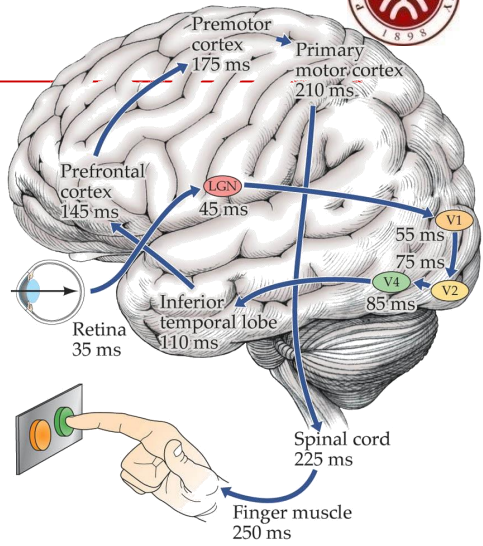
In Brief

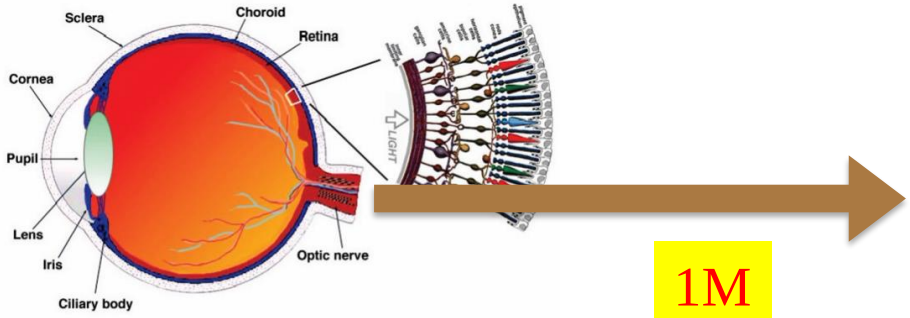
The unique perceptual features of the primate fovea are shaped by the properties of cone photoreceptors, rather than retinal circuit computations.

现有数字成像系统仅模拟了中央凹的功能

Pathway: feature compression

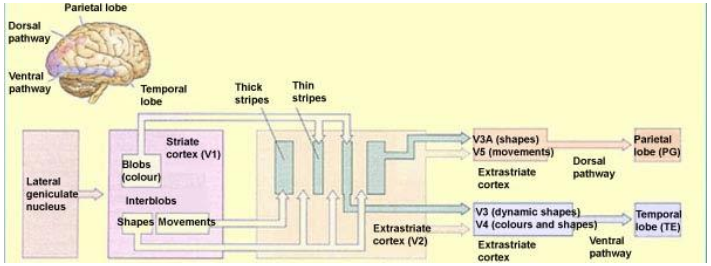
- 视网膜，由1.26亿个（126M）光感细胞构成
  - 锥状Cone细胞：600多万个（RGB各1/3）
  - 杆状Rod细胞：1.2亿个
  - 视神经网络：100万条神经
- 大脑，1300g
  - 860亿（约10^{11}）神经元（neurons），140亿个神经细胞
- 二者连接：100万突触





126M

1M



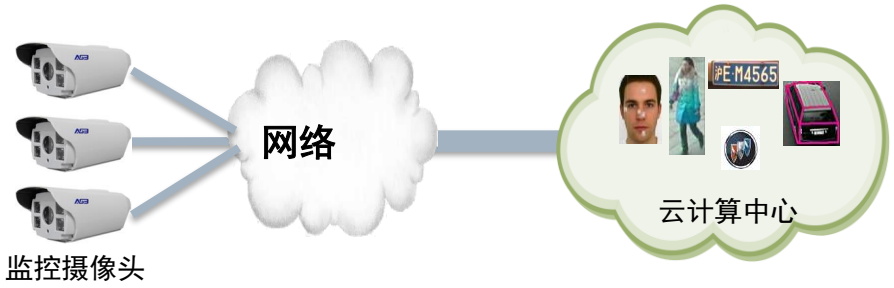
140亿

Maunsell and Gibson 1992; Raiguel et al. 1989; Nowak et al. 1995; Schmolesky et al. 1998; Thorpe, Fize& Marlot 1996

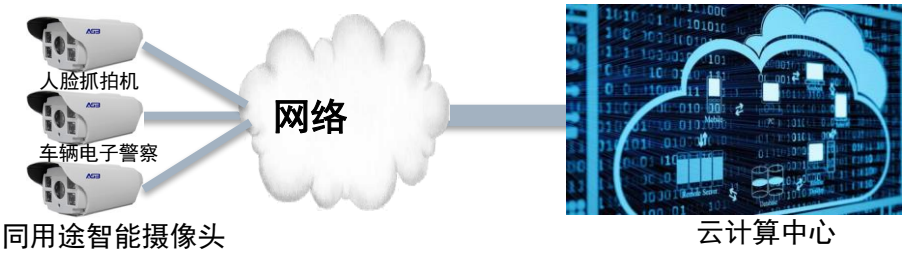
眼睛到大脑之间，编码压缩了多少？

Compress Ratio, 126 : 1

Rethinking: current CVS is weak in feature



Mode 1：video compression only



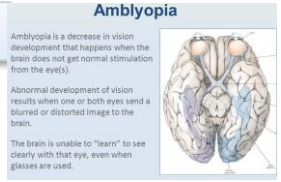
Mode 2：recognition result only

类比

自闭症：不能有效过滤掉无关的感觉刺激信号的输入,以至于被过多的无关刺激信息所淹没,干扰大脑对有用信息的加工,导致了选择注意、认知及信息处理异常等各种相关症状

工程学缺陷：成本高、浪费能源

类比



癫痫症：带来推理 泛化缺陷
弱视症：带来认知平衡缺陷

工程学缺陷：系统适应性差、升级困难

如何改进？

类“视网膜-大脑”的工作方式——分工协调

What we learn from human-like vision?

- Biological system
  - Three key parts
    - Eye(Retina), image and feature coding
    - Pathways, transmission
    - Brain(visual field), object recognition
- Artificial system
  - Three key parts
    - Digital Retina(image and feature coding)
    - Networking
    - Cloud Computing

Outline

- CVS and its challenge
- What we learn from HVS?
- Digital retina – a way to make CVS more efficient
- Summary

数字视网膜：仿生物视网膜的视觉计算架构

中国科学：信息科学 2018 年 第 48 卷 第 8 期：1076 -1082
SCIENTIA SINICA Informationis
《中国科学》杂志社
SCIENCE CHINA PRESS
观点与争鸣

数字视网膜：智慧城市系统演进的关键环节

高文[1]，田永鸿[1*]，王坚[2]

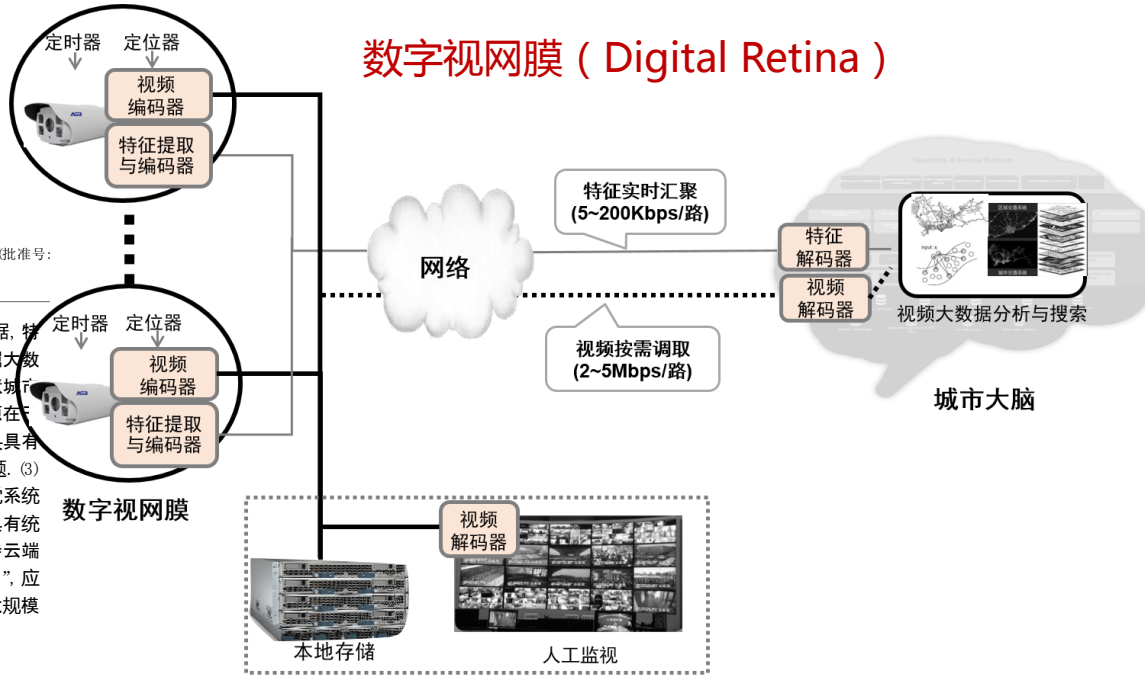
1. 北京大学信息科学技术学院，北京 100871
2. 阿里巴巴集团，杭州 311121
* 通信作者. E-mail: yhtian@pku.edu.cn

收稿日期：2018 -01 -31；接受日期：2018 -03 -03；网络出版日期：2018 -05 -21

国家重点研发计划 “云计算与大数据”重点专项（批准号：2017YFB1002400）、国家重点基础研究发展计划（973）（批准号：2015CB351800）和国家自然科学基金大数据科学中心项目（批准号：U1611461）资助

摘要 本文阐述了作者对智慧城市建设和发展的主要观点：(1) 如何实时聚合各类城市大数据，特别是来自视频监控网络的图像视频数据，并通过构建基于云计算的 “城市大脑”来分析和挖掘大数据价值并服务于城市运营与管理，是智慧城市发展中亟待解决的一个关键问题. (2) 现阶段智慧城市建设的现状是 “有眼、有脑”，但作为 “眼睛”的摄像头功能过于单一使得 “脑强眼弱”，其根源在于传统监控摄像机网络所采用的技术体系是为存储而不是分析设计的. 尽管近期有些智能摄像头具有车牌或人脸识别功能，但是这种单纯强调 “边缘计算”的方案仍然无法解决 “眼脑合一”的问题. (3) 为了解决目前阻碍智慧城市系统功能快速演进的难题，我们应借鉴人类进化了数十万年的视觉系统之 “人类视网膜同时具有影像编码与特征编码功能”这一特性，研究与设计数字视网膜，使之具有统一时间戳和精确地理位置，能同时进行高效视频编码和紧凑特征表达的联合优化，并有效支持云端大规模监控视频分析与快速视觉搜索等功能. (4) 为利用数字视网膜来构筑智慧城市的 “慧眼”，应积极布局与推进相关标准制定、芯片与硬件实现、支撑软件开发与软硬件开源社区，并开展大规模测试与应用.

关键词 智慧城市，城市大脑，数字视网膜



引用格式：高文，田永鸿，王坚. 数字视网膜：智慧城市系统演进的关键环节. 中国科学：信息科学，2018，48：1076–1082，doi：10.1360/N112018-00025
Gao W, Tian Y H, Wang J. Digital retina: revolutionizing camera systems for the smart city (in Chinese). Sci Sin Inform, 2018, 48: 1076–1082, doi: 10.1360/N112018-00025

Definition of a Digital Retina

- A digital retina, should be satisfied as at least below items,
  1. Unified time stamp
  2. The geographical location
  3. High efficient video coding
  4. High efficient feature coding
  5. Joint optimization between video coding and feature coding
  6. High efficient model updating
  7. Top-down attention
  8. Software defined function X

Definition of a Digital Retina

- A digital retina, should be satisfied as at least below items,
  1. Unified time stamp
  2. The geographical location
  3. High efficient video coding
  4. High efficient feature coding
  5. Joint optimization between video coding and feature coding
  6. High efficient model updating
  7. Top-down attention
  8. Software defined function X

Feature group 1: item 1 & 2

- 特征1：全局统一的时空ID，包括
  - 使用全网统一的时间
  - 提供精确地理位置，如GPS, Baidou
  - (可扩展)提供摄像机视角、参数等信息

从摄像图像视频中场景、对象：可定位、可标识

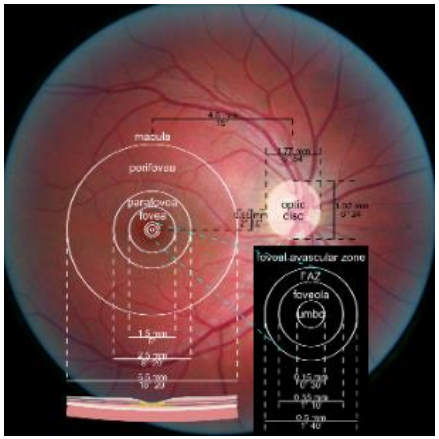
引用格式：高文, 田永鸿, 王坚. 数字视网膜：智慧城市系统演进的关键环节. 中国科学: 信息科学, 2018, 48: 1076–1082, doi: 10.1360/N112018-00025
Gao W, Tian Y H, Wang J. Digital retina: revolutionizing camera systems for the smart city (in Chinese). Sci Sin Inform, 2018, 48: 1076–1082, doi: 10.1360/N112018-00025

Definition of a Digital Retina

- A digital retina, should be satisfied as at least 8 features,
  1. Unified time stamp
  2. The exact geographical location
  3. High efficient video coding
  4. High efficient feature coding
  5. Joint optimization between video coding and feature coding
  6. High efficient model updating
  7. Top-down attention
  8. Software defined function X

Feature group 2: item 3 & 4 & 5

- 特征2：多层次视网膜表示: 视频编码 + 特征编码 + 联合优化
  - 视频编码：为了存储和离线观看的影像重构
  - 特征编码：为了模式识别和场景理解的紧凑特征表示
  - (可扩展)联合优化：模拟生物视网膜，支持视频码流与特征码流联合编码优化



视网膜

仿视网膜的多层视觉信息表示

引用格式：高文, 田永鸿, 王坚. 数字视网膜: 智慧城市系统演进的关键环节. 中国科学: 信息科学, 2018, 48: 1076–1082, doi: 10.1360/N112018-00025
Gao W, Tian Y H, Wang J. Digital retina: revolutionizing camera systems for the smart city (in Chinese). Sci Sin Inform, 2018, 48: 1076–1082, doi: 10.1360/N112018-00025

Definition of a Digital Retina

- A digital retina, should be satisfied as at least 8 features,
  1. Unified time stamp
  2. The exact geographical location
  3. High efficient video coding
  4. High efficient feature coding
  5. Joint optimization between video coding and feature coding
  6. High efficient model updating
  7. Top-down attention
  8. Software defined function X

Feature group 3: item 6 & 7 & 8

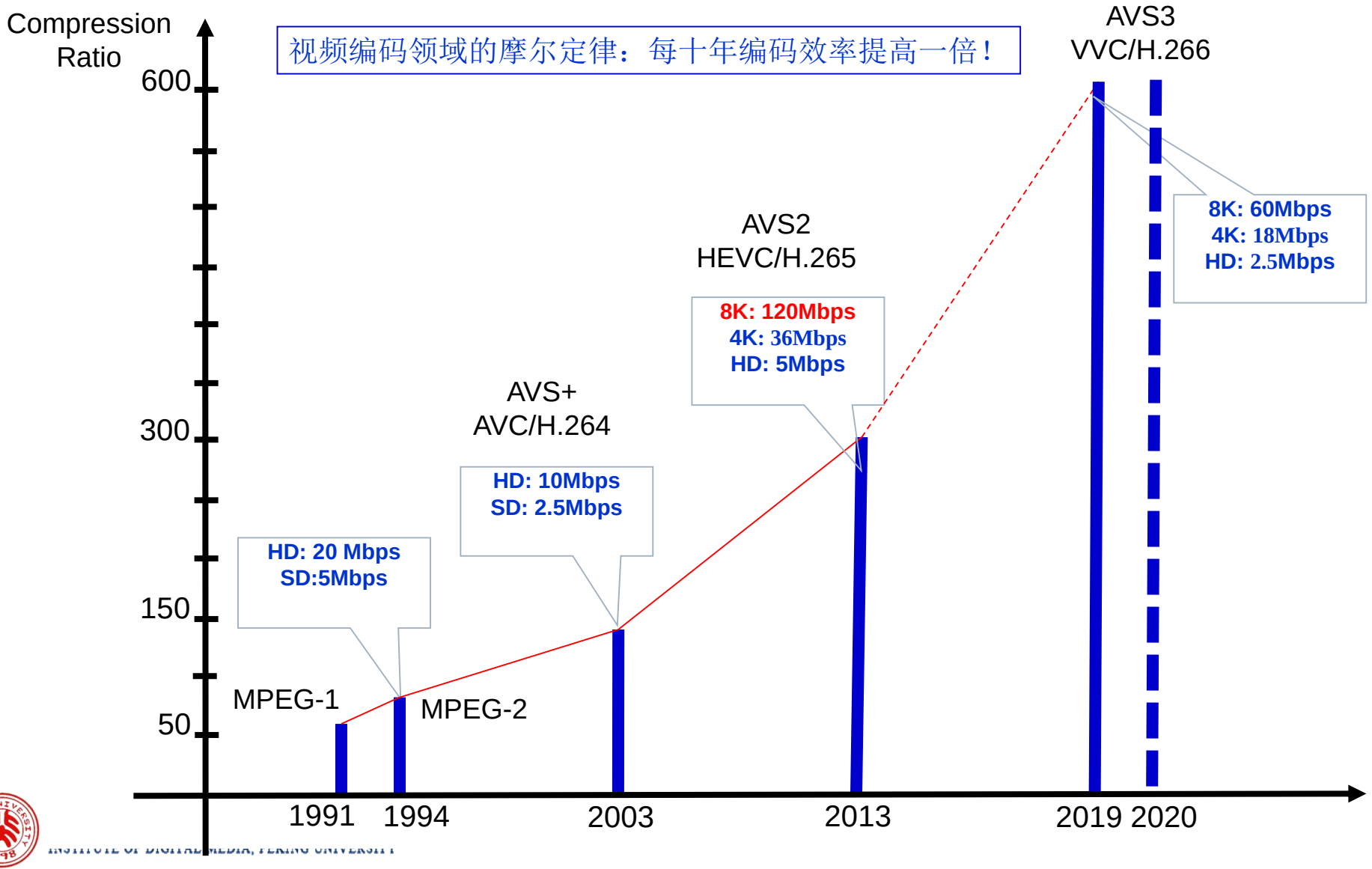
- 特征3：模型可更新 + 注意可调节 + 软件可定义
  - 模型可更新：支持端/边深度学习模型的自适应迁移、压缩、更新与转换
  - 注意可调节：模拟视觉注意机制，在端设备、感知网络等层面实现动态注意调节
  - 软件可定义：支持端边云协同计算与推理，实现特征实时汇聚与视频按需调取

引用格式：高文, 田永鸿, 王坚. 数字视网膜: 智慧城市系统演进的关键环节. 中国科学: 信息科学, 2018, 48: 1076–1082, doi: 10.1360/N112018-00025
Gao W, Tian Y H, Wang J. Digital retina: revolutionizing camera systems for the smart city (in Chinese). Sci Sin Inform, 2018, 48: 1076–1082, doi: 10.1360/N112018-00025

Enable Tech 1: video coding

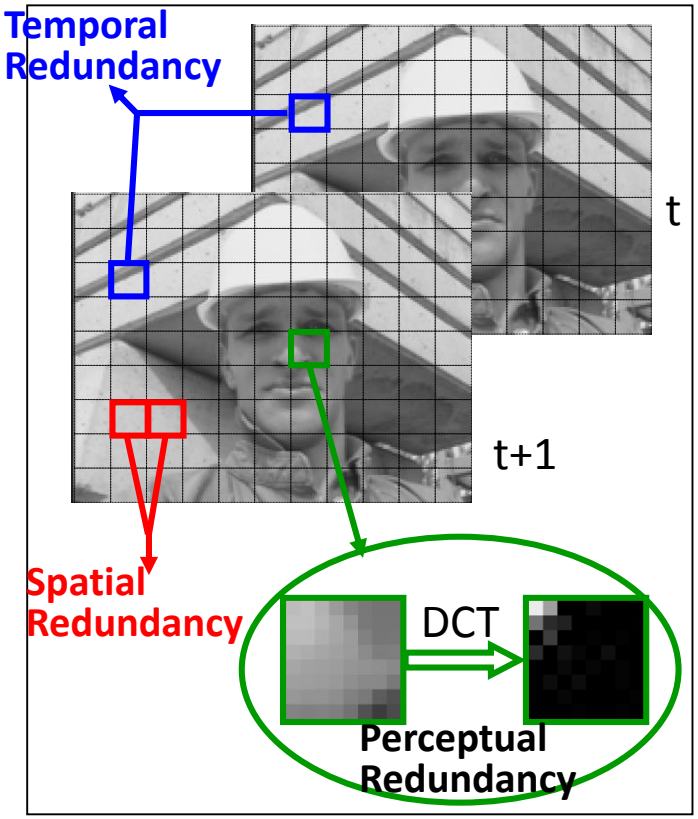
- Our key contribution:
  - Background Modeling, best for Surveillance Video Coding

Standards in video coding

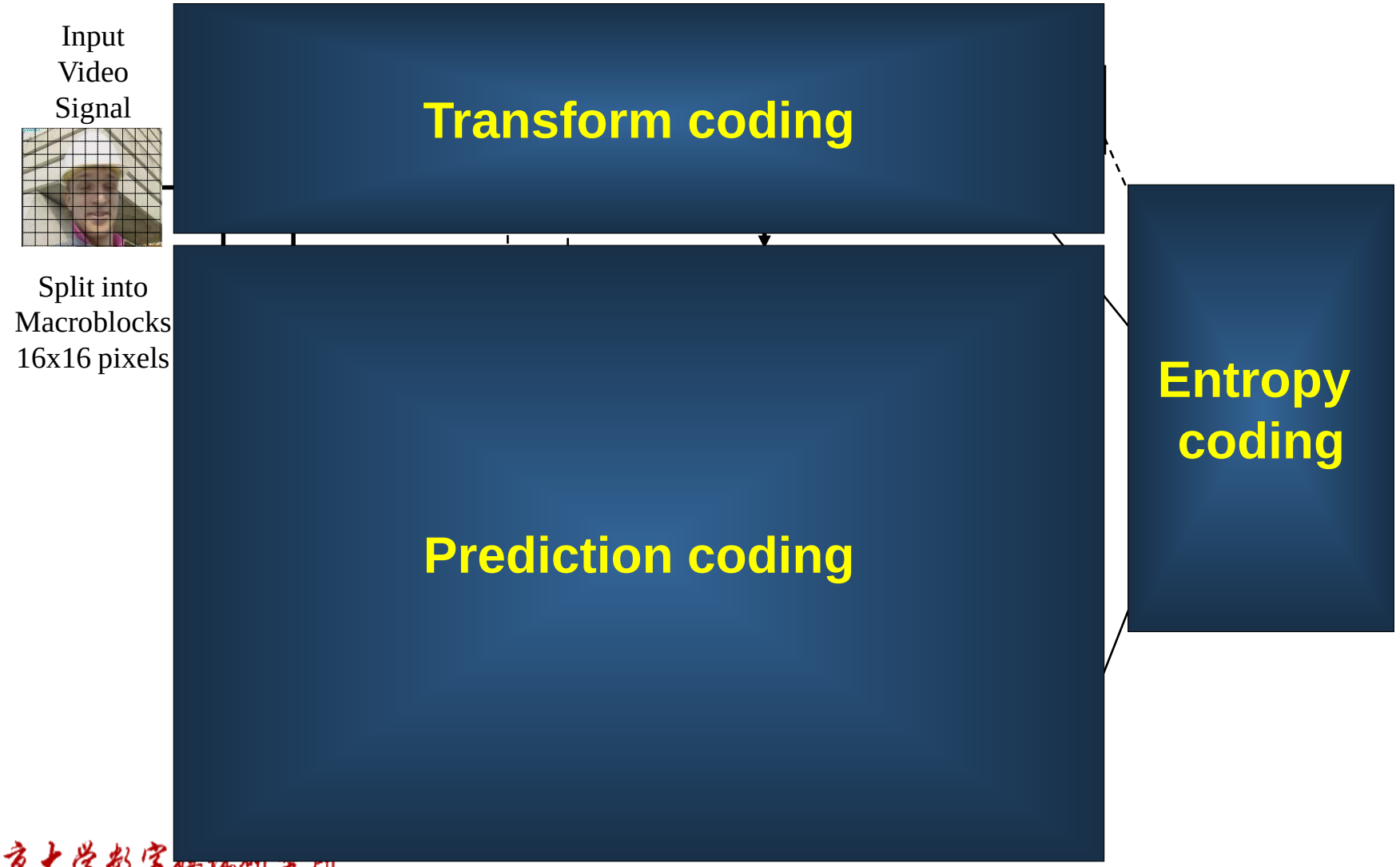


Video Coding: Fundamentals

- Key Issue: Redundancy Removal
  - Utilize signal processing tools on pixels or blocks
  - Based on Shannon Information Theory

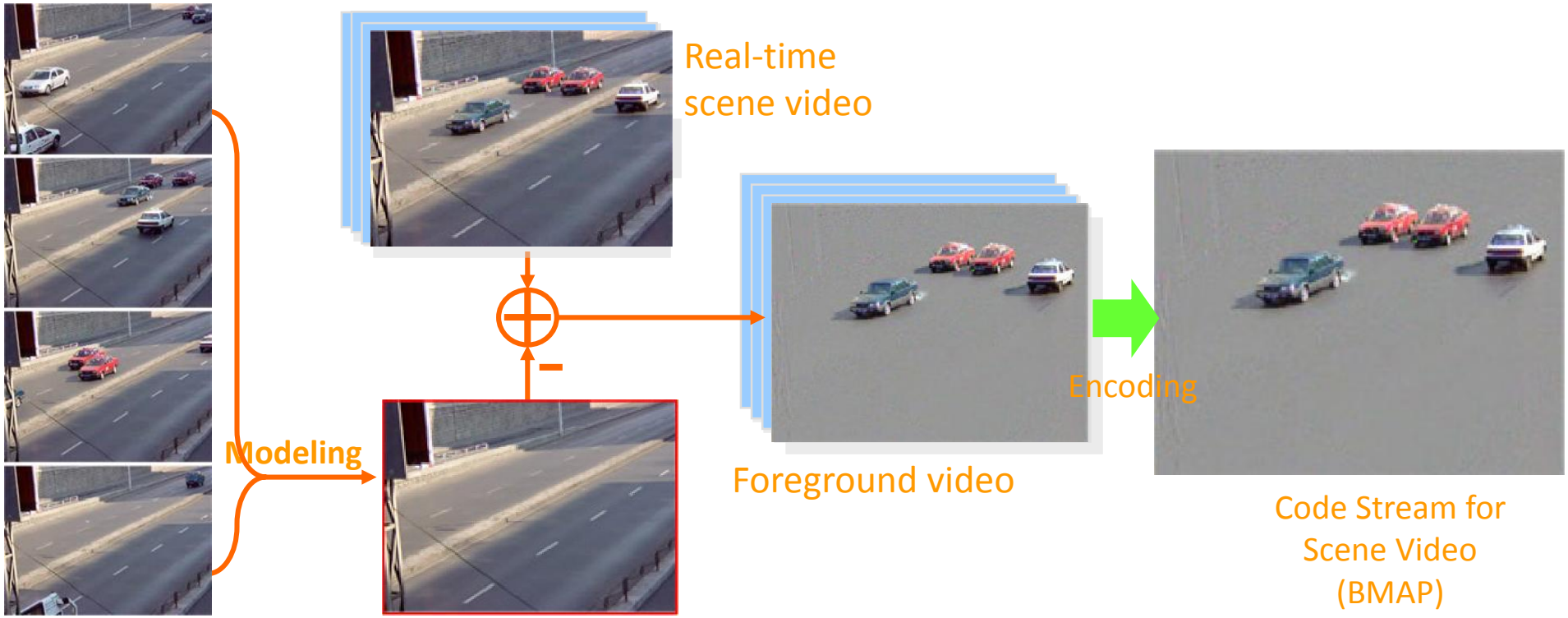


Hybrid Video Coding

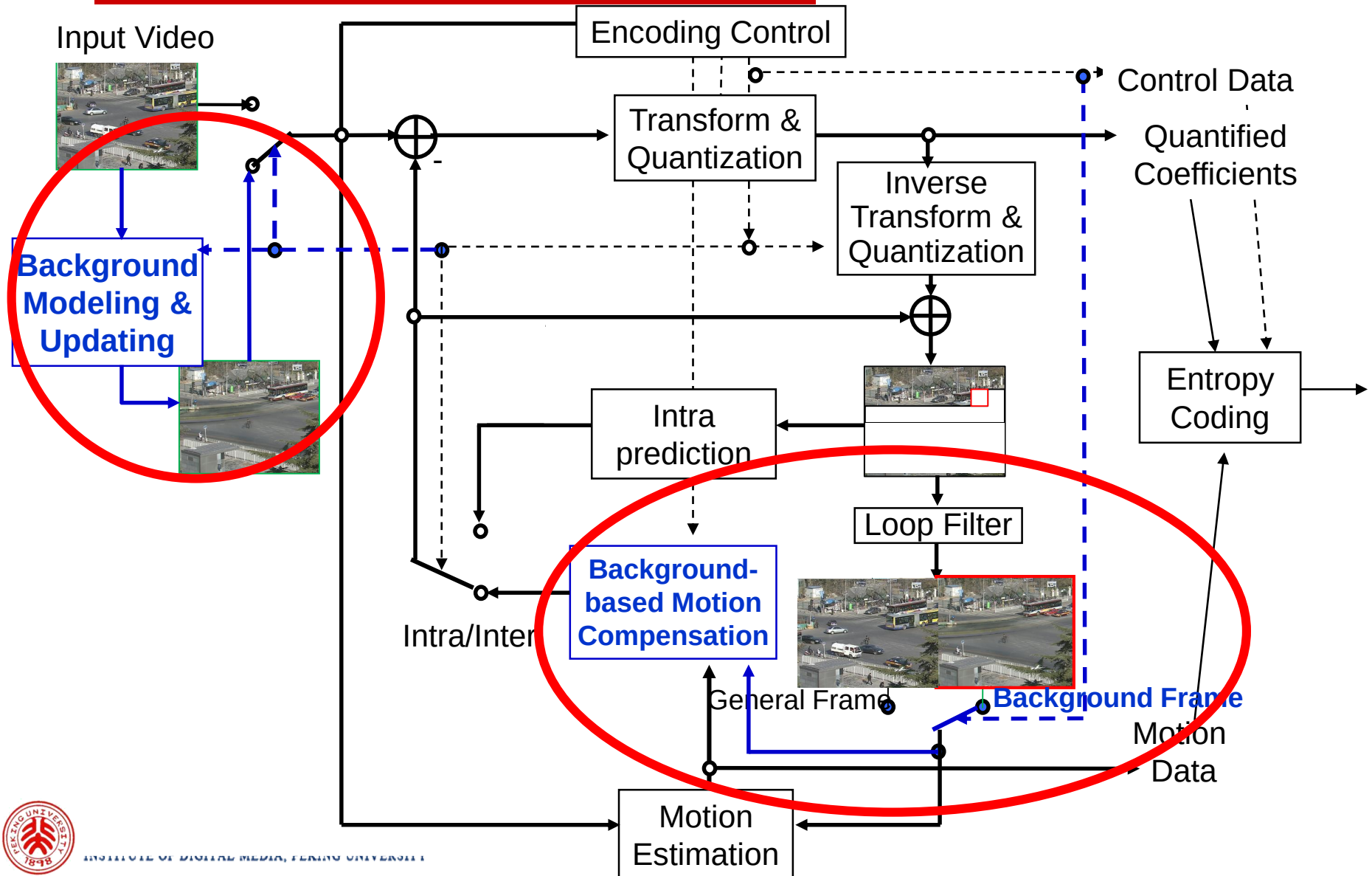


Modeling Background for Video Compression

- The background redundency – A new kind of redundency that can be captured by the background model more or less
- Generating a virtual picture based on background model and embedding into the code stream



Background-Modeling based Surveillance Video Coding



40% gain in coding efficiency using our model up to HEVC

- HEVC HM12.0 vs. BHO

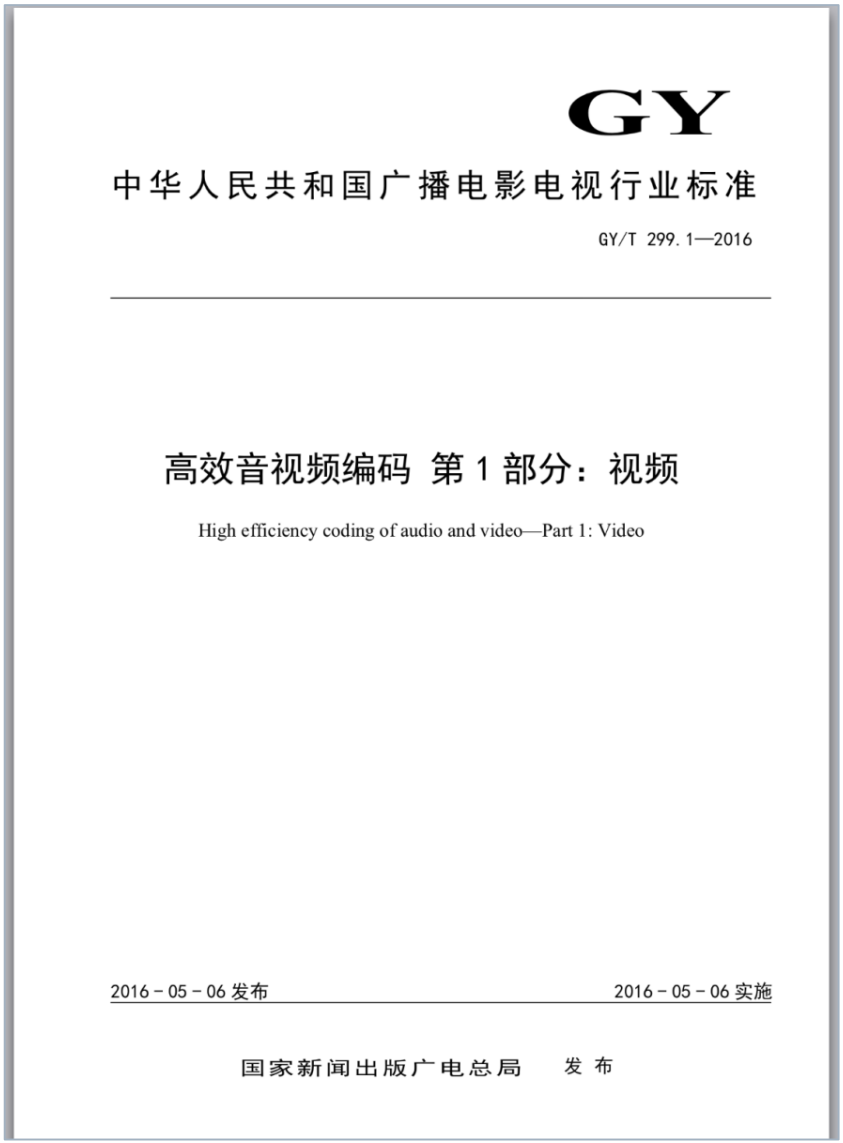
Surveillance Videos	BEO vs. HM 12.0			
	BD Rate (Y,U,V)			Time Saving
Crossroad-cif	-18.39%	-46.41%	-43.20%	32.28%
Overbridge-cif	-30.60%	-79.59%	-51.80%	26.03%
Snowgate-cif	-55.88%	-77.13%	-74.02%	44.22%
Snowroad-cif	-53.18%	-66.21%	-66.40%	60.06%
Bank-sd	-48.88%	-72.46%	-73.78%	60.79%
Crossroad-sd	-29.24%	-71.06%	-67.37%	37.73%
Office-sd	-16.17%	-54.70%	-50.88%	27.28%
Overbridge-sd	-46.91%	-71.84%	-70.48%	56.05%
Intersection-hd	-21.45%	-33.74%	-31.28%	26.28%
Mainroad-hd	-70.15%	-83.13%	-75.49%	65.59%
Average	**-39.09%**	**-65.63%**	**-60.47%**	**43.63%**

Conference Videos	BEO vs. HM			
	BD Rate (Y,U,V)			Time Saving
FourPeople-720p	-8.02%	-15.86%	-14.41%	37.31%
Johnny-720p	1.82%	-15.91%	-14.53%	48.33%
Kristen&Sara-720p	-9.06%	-19.28%	-18.70%	41.18%
Vidyo1-720p	-5.99%	-11.15%	-13.02%	38.14%
Vidyo3-720p	-10.10%	-16.53%	-33.67%	56.90%
Vidyo4-720p	-0.26%	-13.37%	-15.18%	40.19%
Average	**-5.27%**	**-15.35%**	**-18.25%**	**43.68%**

Results: BHO can achieve ~40% bit saving and 43.63% complexity reduction on surveillance videos, while those are ~6% and 43.68% on conference videos.

Xianguo Zhang, Yonghong Tian, Tiejun Huang, Siwei Dong, Wen Gao, Optimizing the Hierarchical Prediction and Coding in HEVC for Surveillance and Conference Videos with Background Modeling, ***IEEE Transactions on Image Processing***, 23(10), Oct. 2014. 4511-4526.

AVS2 with B model, published in 2016



GY

中华人民共和国广播电影电视行业标准

GY/T 299.1—2016

高效音视频编码 第1部分：视频

High efficiency coding of audio and video—Part 1: Video

2016-05-06 发布　　2016-05-06 实施

国家新闻出版广电总局　发布

ICS 35.040
L 71

GB

中华人民共和国国家标准

GB/T 33475.2—2016

信息技术 高效多媒体编码
第2部分：视频

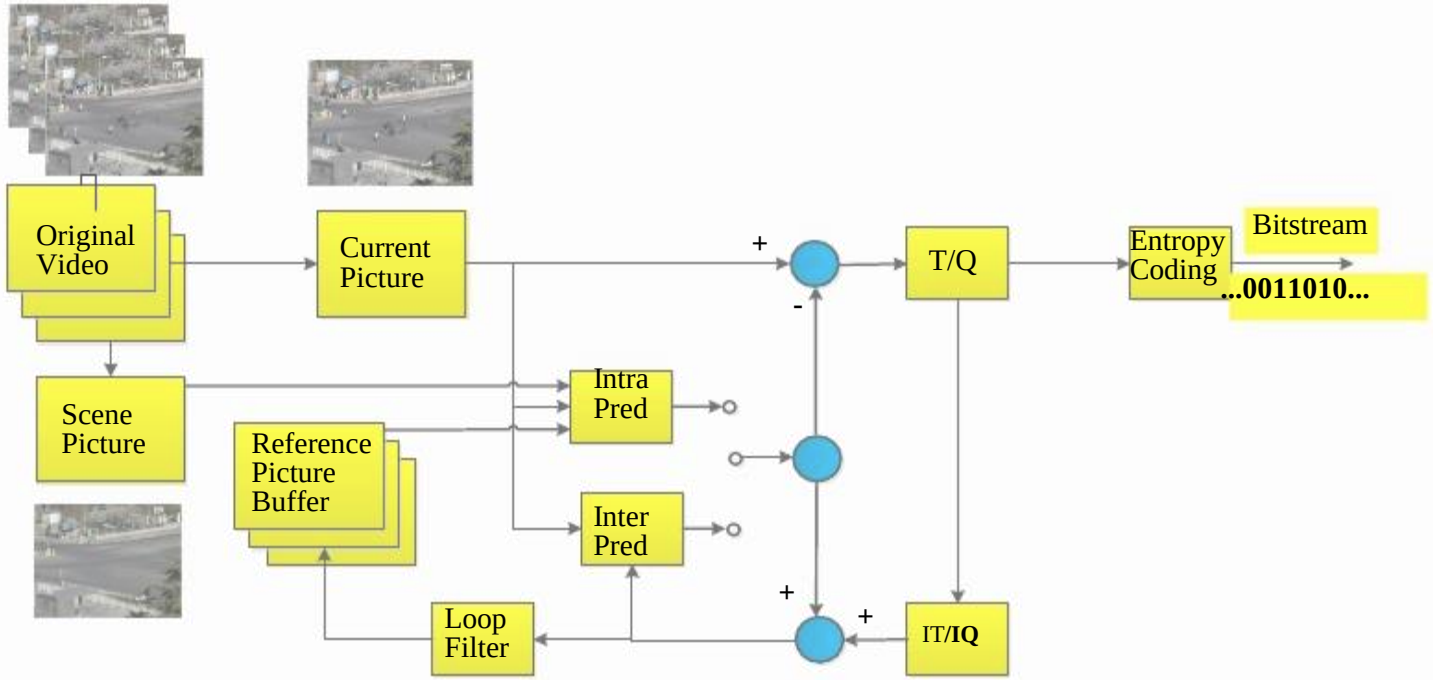
Information technology - High efficiency media coding - Part 2: Video

2016-12-30 发布　　2017-07-01 实施

中华人民共和国国家质量监督检验检疫总局
中国国家标准化管理委员会　发布

IEEE 1857.4 published in 2019

Encoder

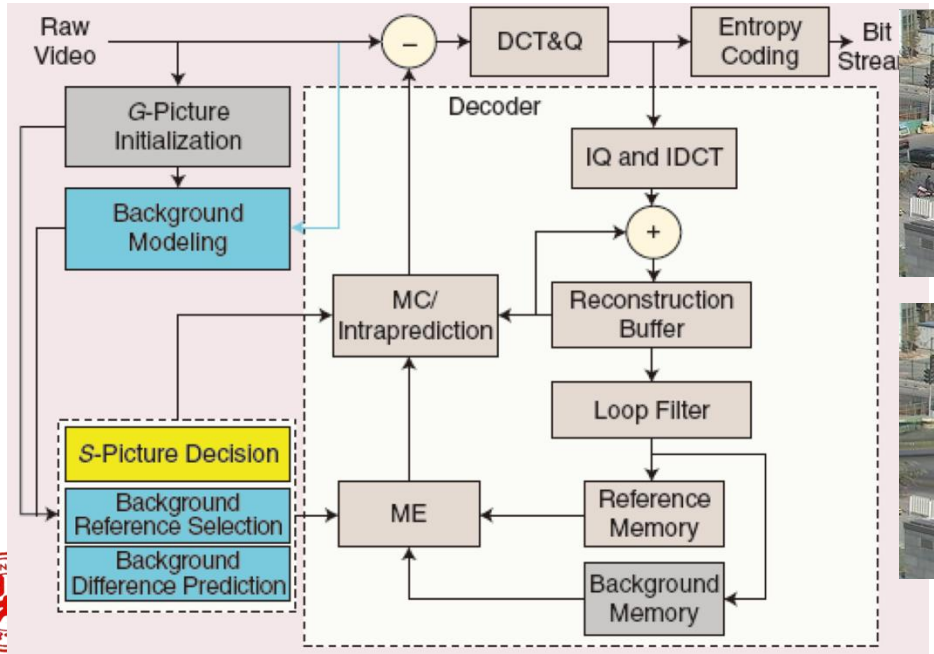


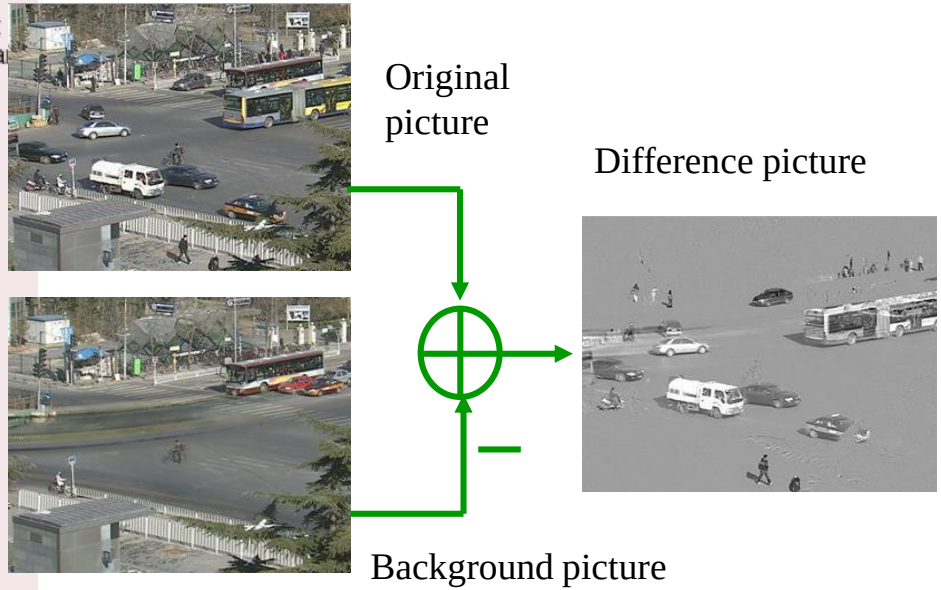
Key Features

- Quad Tree Based Partition: CU/PU/TU
- Directional Intra Prediction and Flexible Inter Prediction
- Transform: Secondary Transform
- Entropy Coding: CBAC
- In-loop Filter: Deblocking Filter + SAO + ALF
- Scene Video Coding：Background picture prediction coding

Enhanced Scene Video Coding in IEEE 1857.4

- Define G/GB reference picture for scene video coding
  - G picture: the I picture in the original stream used as an background picture
  - GB picture: the background picture constructed with background modeling
- Achieve significant bitrate saving for scene video



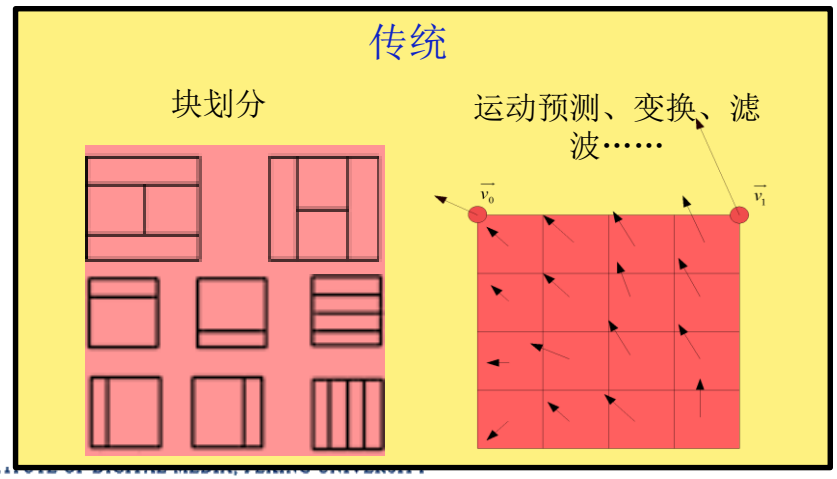


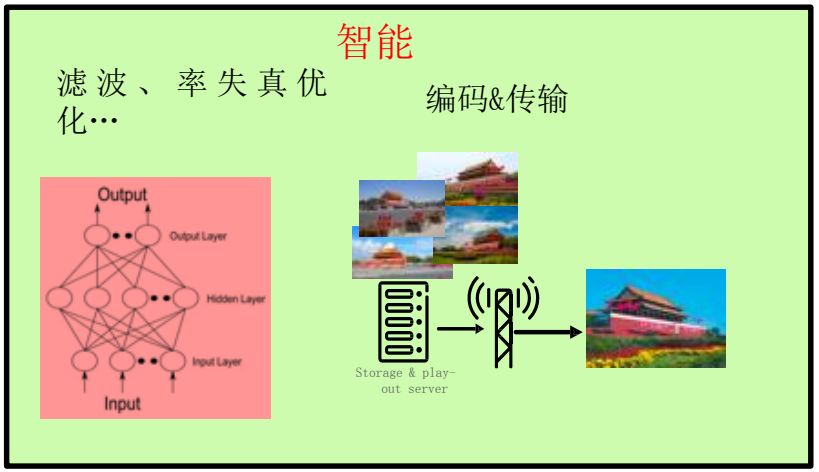
AVS3, one year ahead to VVC/H.266

- AVS3标准，主要面向8K视频需求，早于H.266完成，第一次实现领跑
  - 第六十四次会议，深圳，2018年03月28日-03月31日
  - 第六十五次会议，合肥，2018年06月21日-06月23日
  - 第六十六次会议，长春，2018年08月29日-09月01日
  - 第六十七次会议，厦门，2018年12月05日-12月08日
  - 第六十八次会议，青岛，2019年03月06日-03月09日，AVS Baseline Profile finished
  - 第六十九次会议，成都，2019年06月12日-06月15日
  - 第七十次会议，　海口，2019年08月28日-08月31日

2019年3月发布第一版，AVS3 Baseline

面向8K、VR等视频应用，编码效率比AVS2再提升一倍

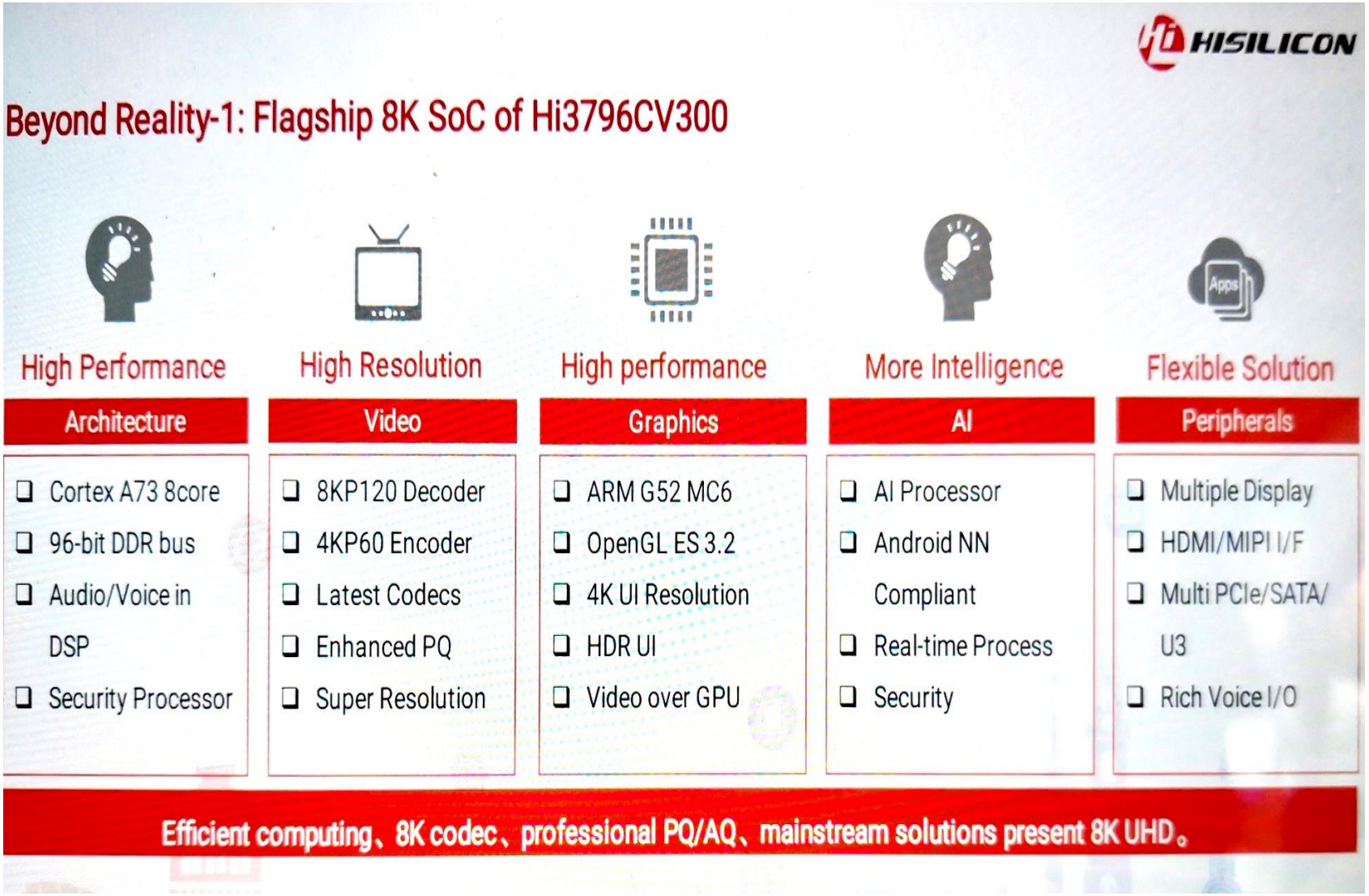




World first AVS3 Chip

- Announced and Demo on Sept. 13th, 2019, at IBC2019, Amsterdam
- By HiSilicon
- AVS3/8K/120P

Hi3796CV300



Enable Tech 2: feature coding

- Our key contribution:
  - CDVS, MPEG-7 part 13
  - CDVA, MPEG-7 part 15

Why compress visual features?

Raw features are still too large

The size of JPEG image > **1MB**

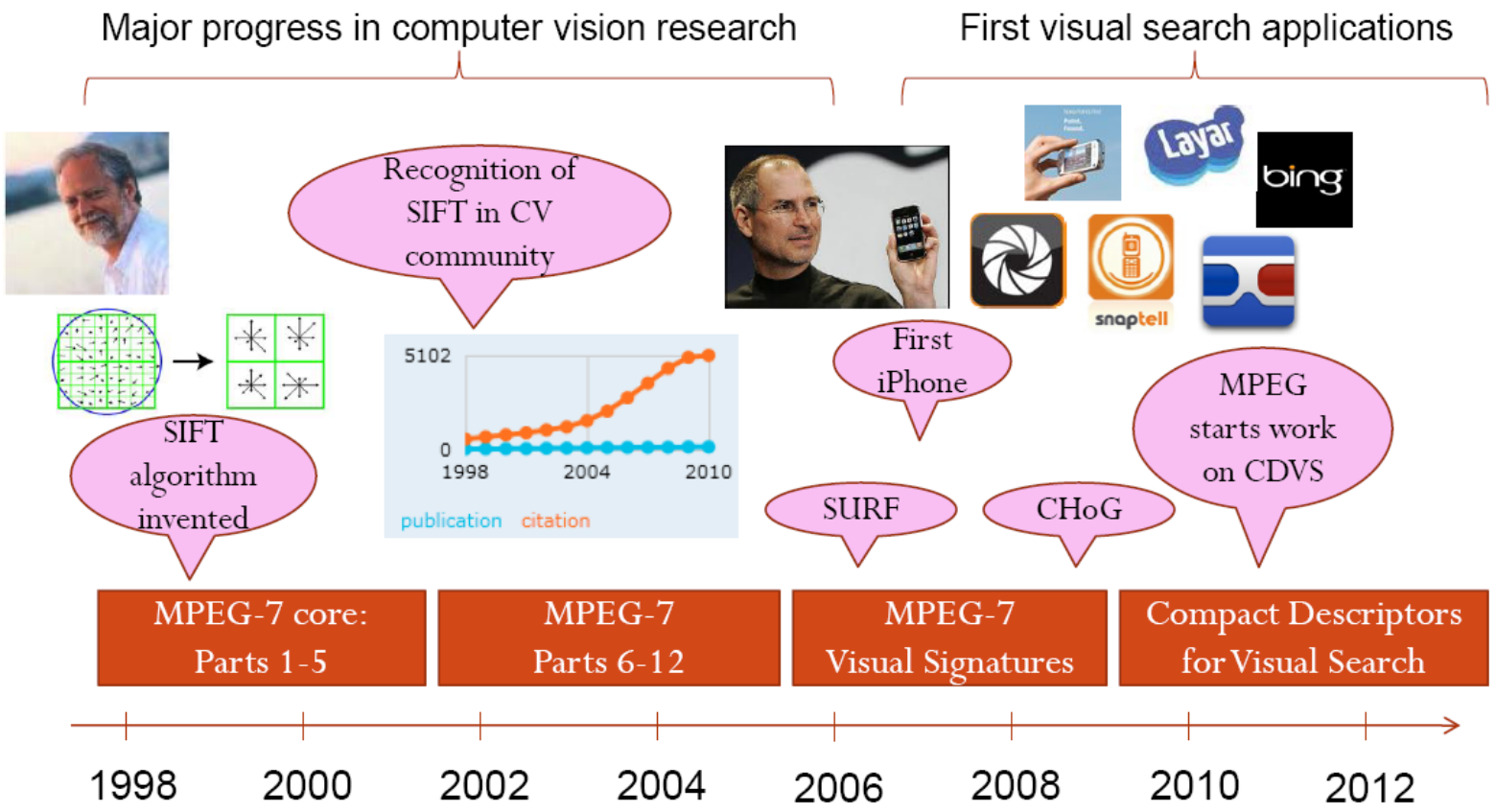
The size of raw SIFT features ~ **520KB**

(Float, 128-dim, ~1000 interest points)

Compact image descriptor **512B, 1KB, 2KB, 4KB, 8KB, 16KB**

Feature Compression – multimedia description

- Moving Picture Experts Group (MPEG), the formal title “ISO/IEC JTC1 SC29 WG11”, initiated the **Compact Descriptors for Visual Search (CDVS)** standard activity at the 91st MPEG meeting (Kyoto, Jan. 2010).



Standardization of MPEG7 compact descriptor

Query

Image

Video

I2V

I2I

V2V

V2I

MPEG7 CDVS

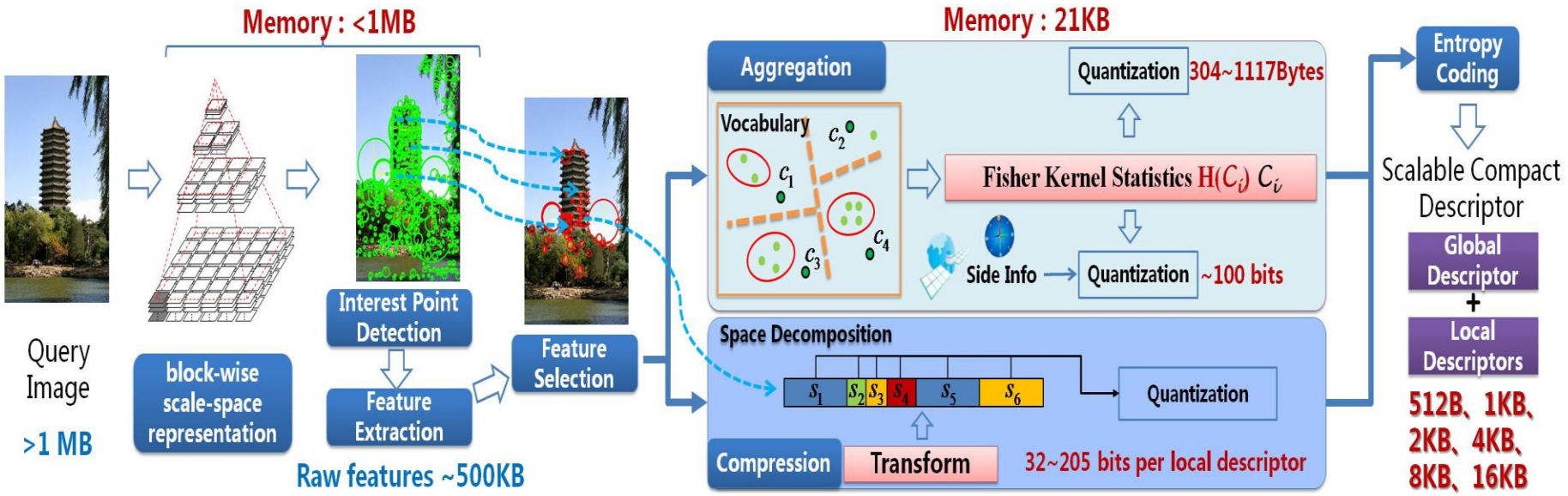
MPEG7 CDVA

Video

Image

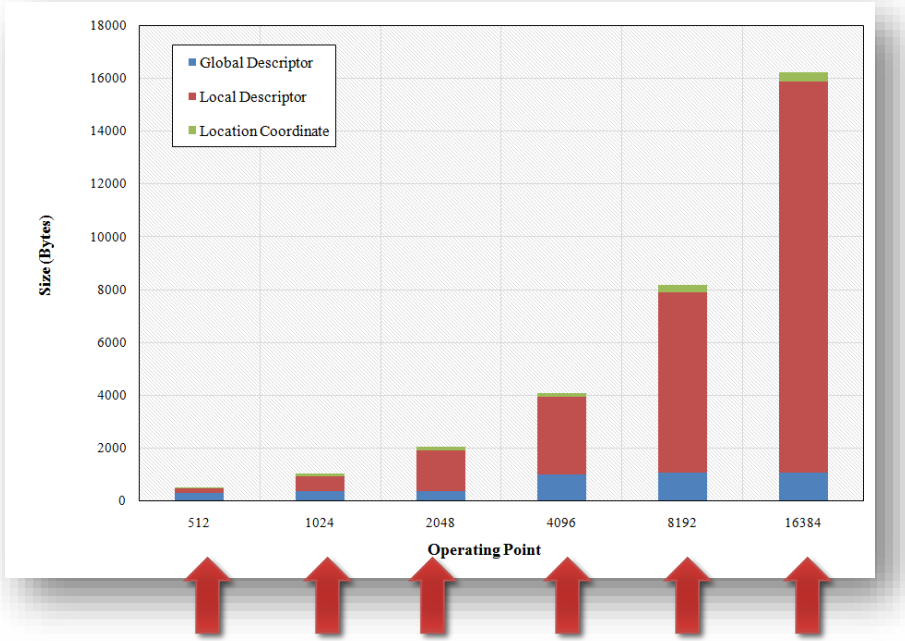
Reference

MPEG7 P13 - CDVS (Handcrafted)

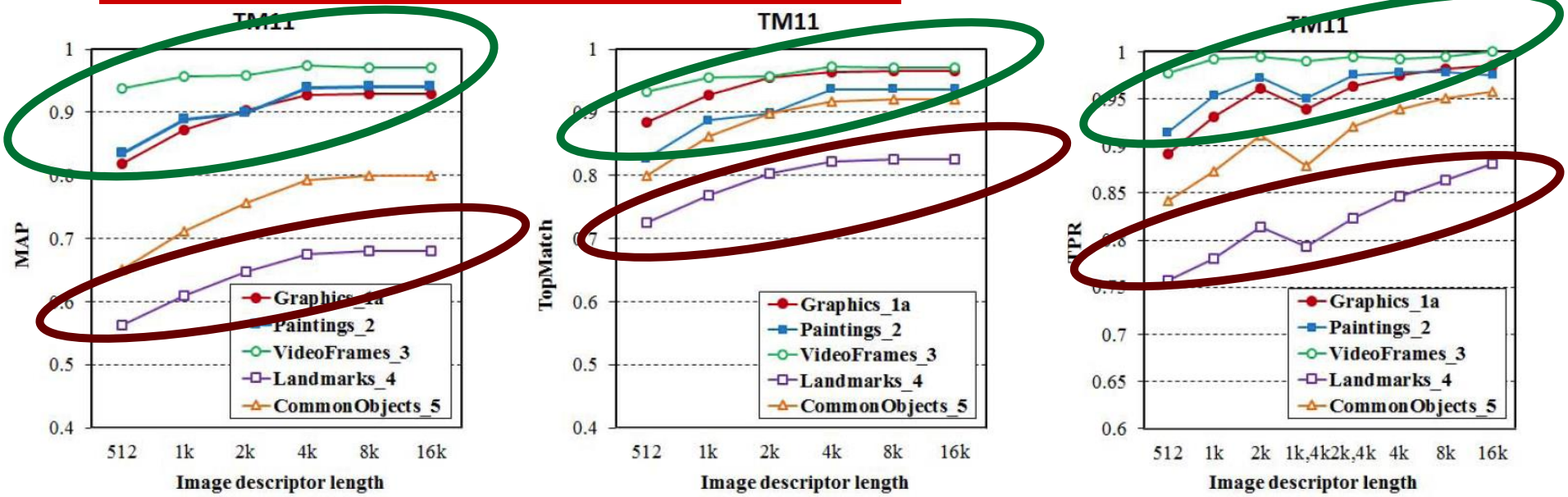


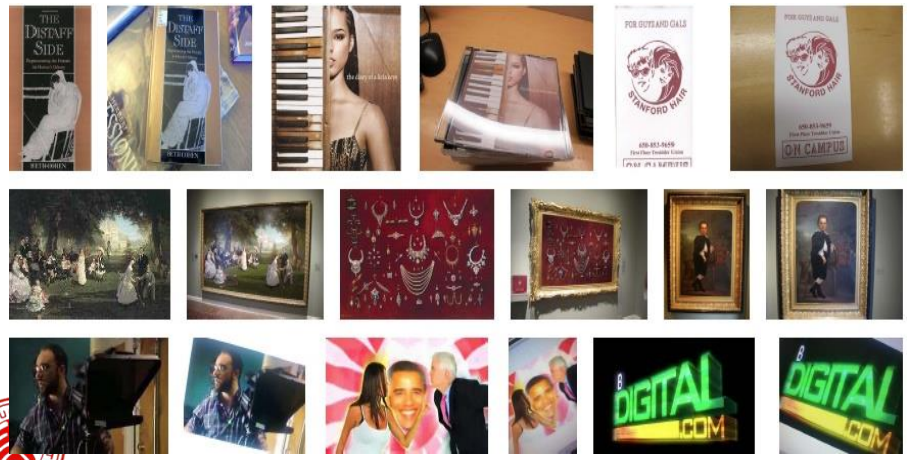
- Interest point detection
- Local feature selection
- Local feature descriptor compression
- Local feature descriptor aggregation
- Local feature location compression

MPEG-CDVS (Handcrafted)



MPEG-CDVS (Handcrafted)





MPEG-CDVS Standardization

- Behind the syntax is high performance techniques

TM	Date	Place	Adopted techniques
TM1.0	Feb. 2012	San Jose, USA	Bag-of-words, feature selection, DISTRAT (m22672) [28].
TM2.0	May 2012	Geneva, Switzerland	Local feature aggregation (REVV, a global descriptor m23578 [57]).
TM3.0	Jul. 2012	Stockholm, Sweden	Scalar quantizer to compress local feature descriptors (m25929 [19]), Location coordinate coding (m25883 [40]).
TM4.0	Oct. 2012	Shanghai, China	Scalable local feature aggregation (SCFV, a scalable global descriptor, m26726 [52]), Multi-Stage Vector Quantization (MSVQ) for local feature descriptor compression (m26727 [34]), weighted matching (m25795 [41]).
TM5.0	Jan. 2013	Geneva, Switzerland	Enhanced SCFV with the addition of accumulated gradient vector with respect to the variance of the Gaussian functions for higher bit rates, i.e., 4 KB, 8 KB, and 16 KB (m28061 [51]).
TM6.0	Apr. 2013	Incheon, Korea	A block-wise frequency domain LoG filter (BFLoG, m28891 [25]), two-way key point feature matching schemes with slightly improved performance (m29359 [68]), MBIT (a fast indexing structure, m28893 [67]).
TM7.0	Jul. 2013	Vienna, Austria	Software maintenance, no technology was adopted
TM8.0	Nov. 2013	Geneva, Switzerland	A Low-degree Polynomial extrema detector (ALP, m31369 [23]).
TM9.0	Jan. 2014	San Jose, USA	Improved SCFV by increasing the number of Gaussian functions from 128 to 256 and incorporating the bit selection mechanism for the lowest descriptor length of 512 bytes (m32261 [53]).
TM10.0	Apr. 2014	Valencia, Spain	Combination of BFLoG and ALP (The block-wise processing has incorporated LoG filtering, extrema detection, and orientation assignment of keypoints, m33159 [24]), further improved SCFV by increasing the number of Gaussian functions from 256 to 512, and introducing the standard deviation-based selection method of Gaussian functions (m33189 [54]).
TM11.0	Jul. 2014	Sapporo, Japan	Software maintenance, no technology was adopted
TM12.0	Oct. 2014	Strasbourg, France	Software maintenance, no technology was adopted
TM13.0	Feb. 2015	Geneva, Switzerland	Software maintenance, no technology was adopted
TM14.0	Jun. 2015	Warsaw, Poland	Software maintenance, no technology was adopted

Standardization of MPEG7 compact descriptor

Query

Image

Video

I2V

I2I

MPEG7 CDVS

V2V

V2I

MPEG7 CDVA

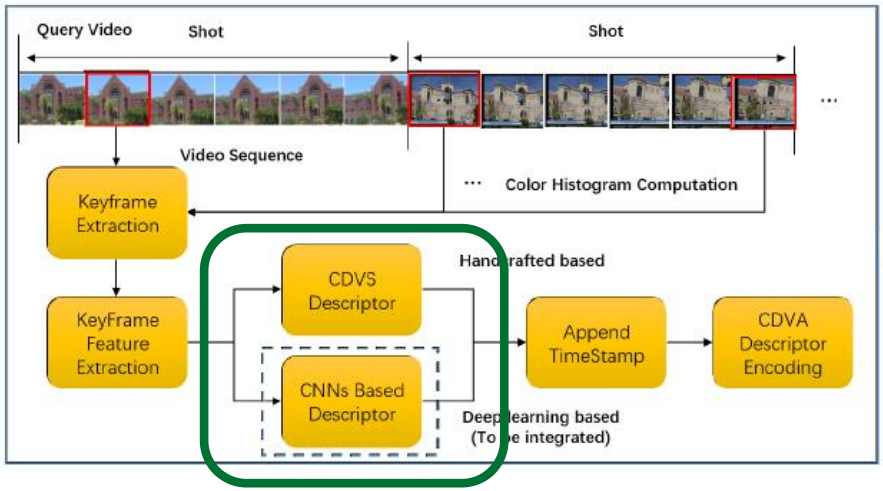
Video

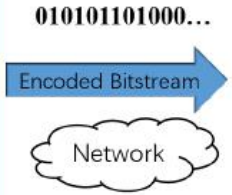
Image

Reference

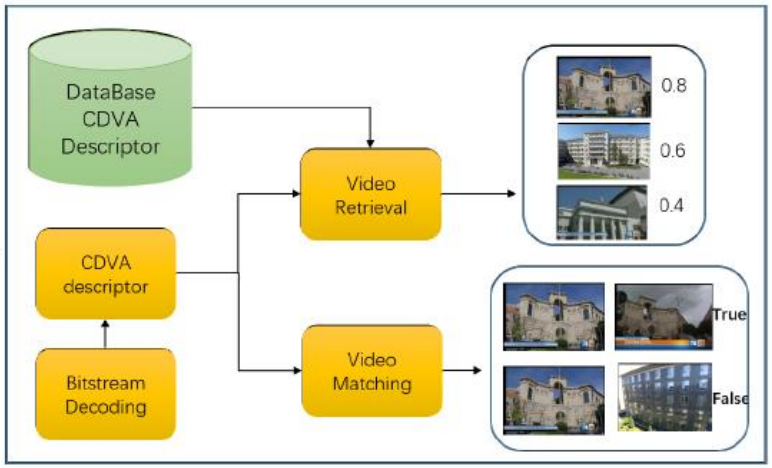
MPEG7 P15 - CDVA (Deep Learning)

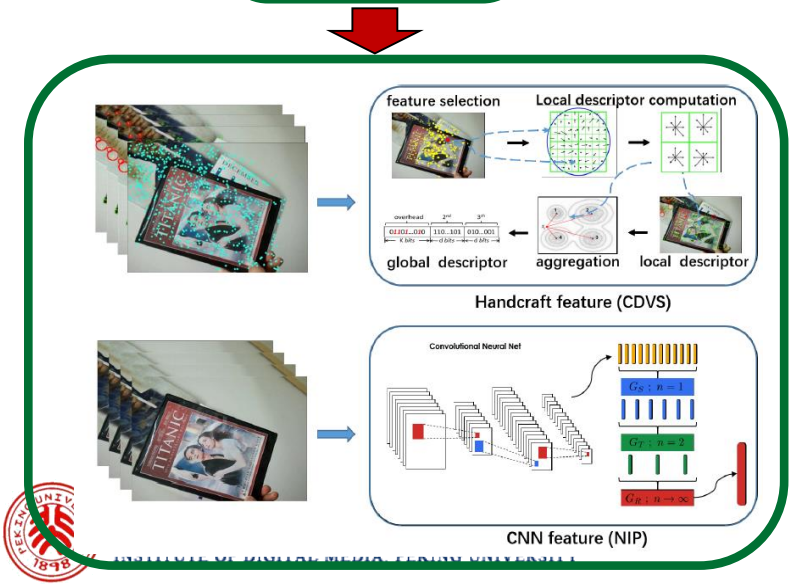
Generating Compact Descriptors





Video Analysis





Leveraging both handcrafted features (MPEG-CDVS) and deep learning features (CNN-NIP)

	mAP	Precisian@R	TPR@FPR=0.01	Localization Accuracy
CXM0.1	0.66	0.655	0.779	0.365
CXM0.2	0.721	0.712	0.836	0.544
CXM1.0	0.721	0.712	0.836	0.662
NIP	0.768	0.736	0.879	0.725
NIP+SCFV	**0.826**	**0.803**	**0.886**	**0.723**
NIP (compressed model)	0.763	0.773	0.87	0.722
NIP (compressed model) + SCFV	0.822	0.798	0.878	0.722
Binarized NIP	0.71	0.673	0.86	0.713
Binarized NIP+ SCFV	0.799	0.775	0.872	0.681

MPEG-CDVA (Deep Learning)

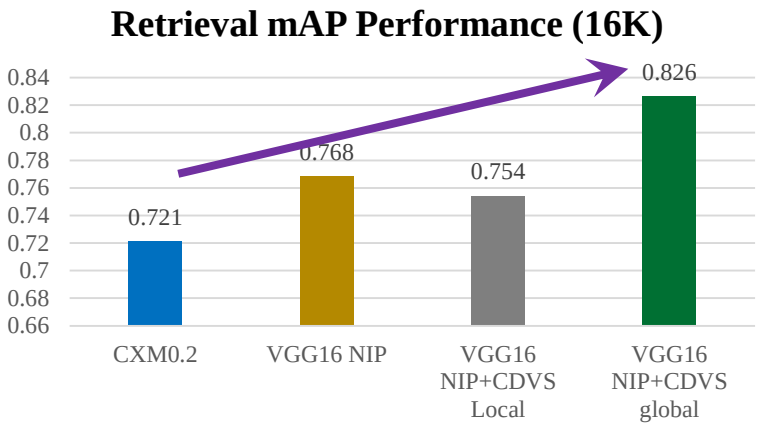
STATISTICS ON THE MPEG CDVA BENCHMARK DATASETS. IOI: ITEMS OF INTEREST. Q.V. : QUERY VIDEOS. R.V.: REFERENCE VIDEOS.

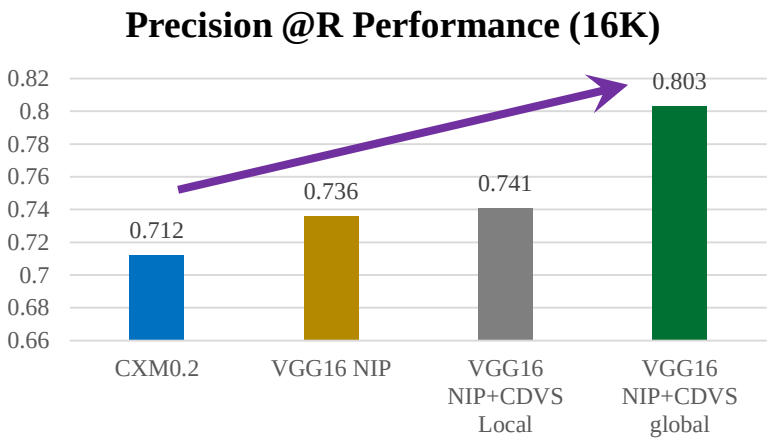
dataset	IoI	# q. v.	hours	frames	# r. v.	hours	frames
All	796	9974	87	8.6M	5127	42	4.5M
Landmarks	489	5224	42	4.2M			
Scenes	71	915	35	3.4M			
Objects	236	3835	10	1.0M			

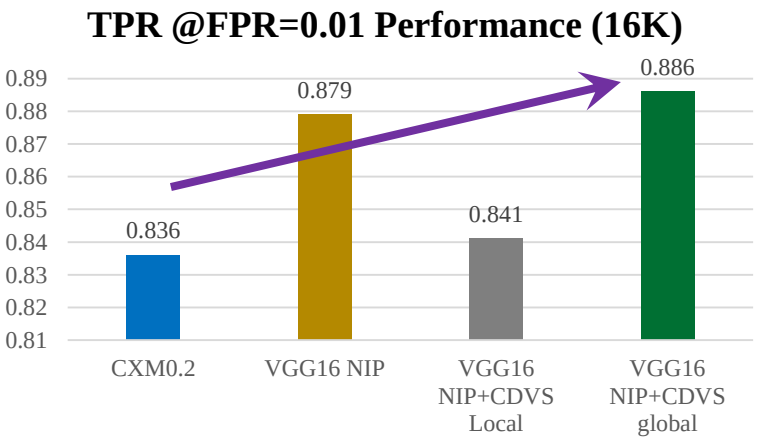
	# videos	hours	frames
Distractors	14537	1029	88M

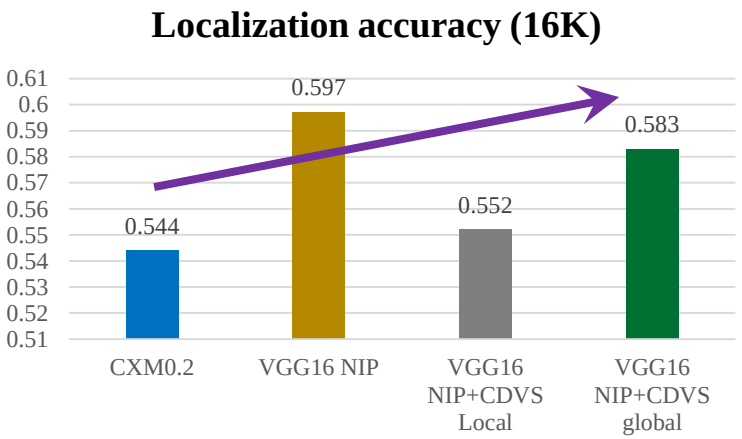
	# video pairs
Matching pairs	4693
Non-matching pairs	46911

MPEG-CDVA (Deep Learning)









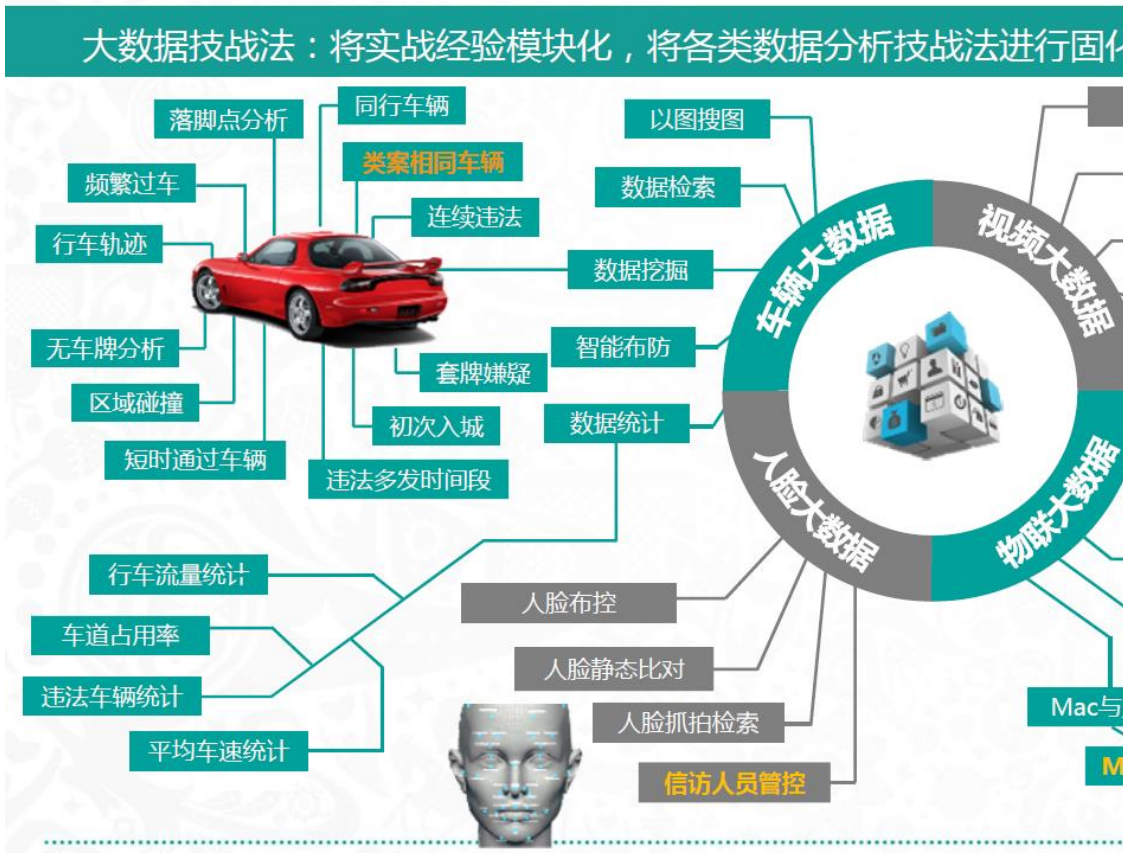
Elegantly leveraging the complementary effects of Handcrafted features (MPEG-CDVS) and deep learning features (NIP)

Applications on CDVS

- Core techniques adopted by Baidu, Tencent, etc.

Applications on CDVA

- Core techniques adopted by Hisense.

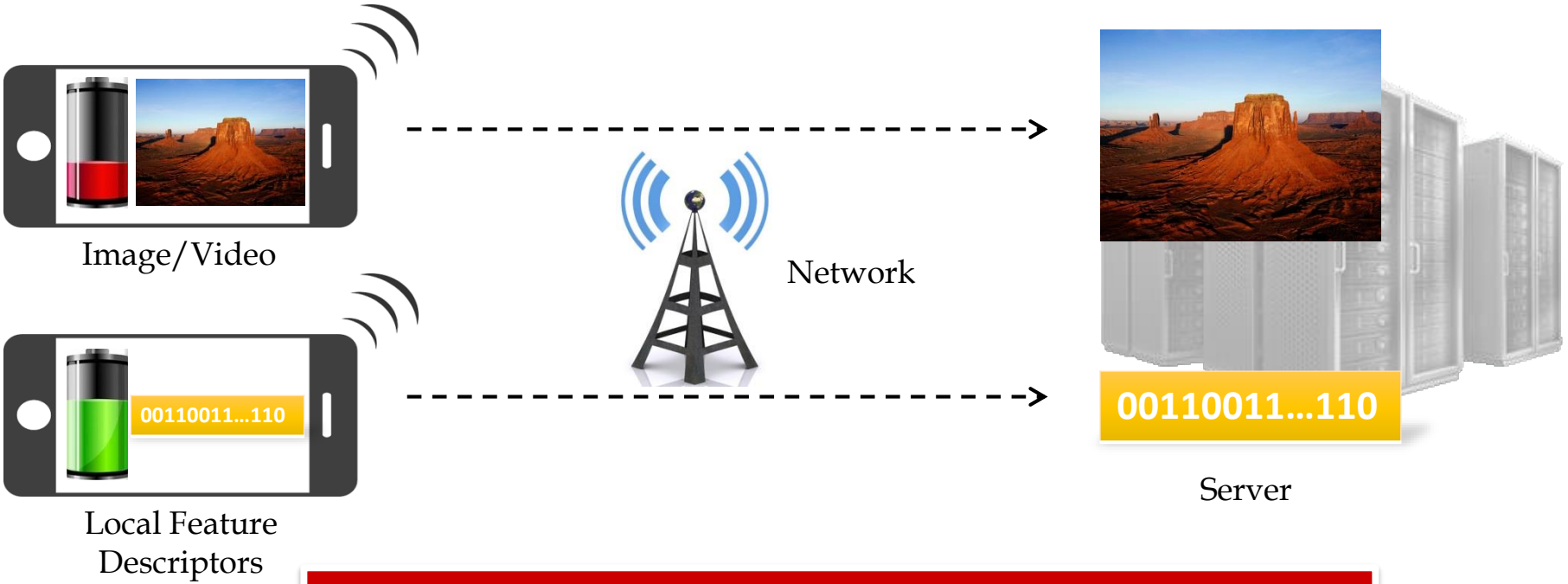


Enable Tech 3: Joint optimization

- Joint optimization between video coding and feature coding, and maybe model coding
- Our key contribution:
  - Joint R-D and R-A Optimization

Joint R-D and R-A Optimization

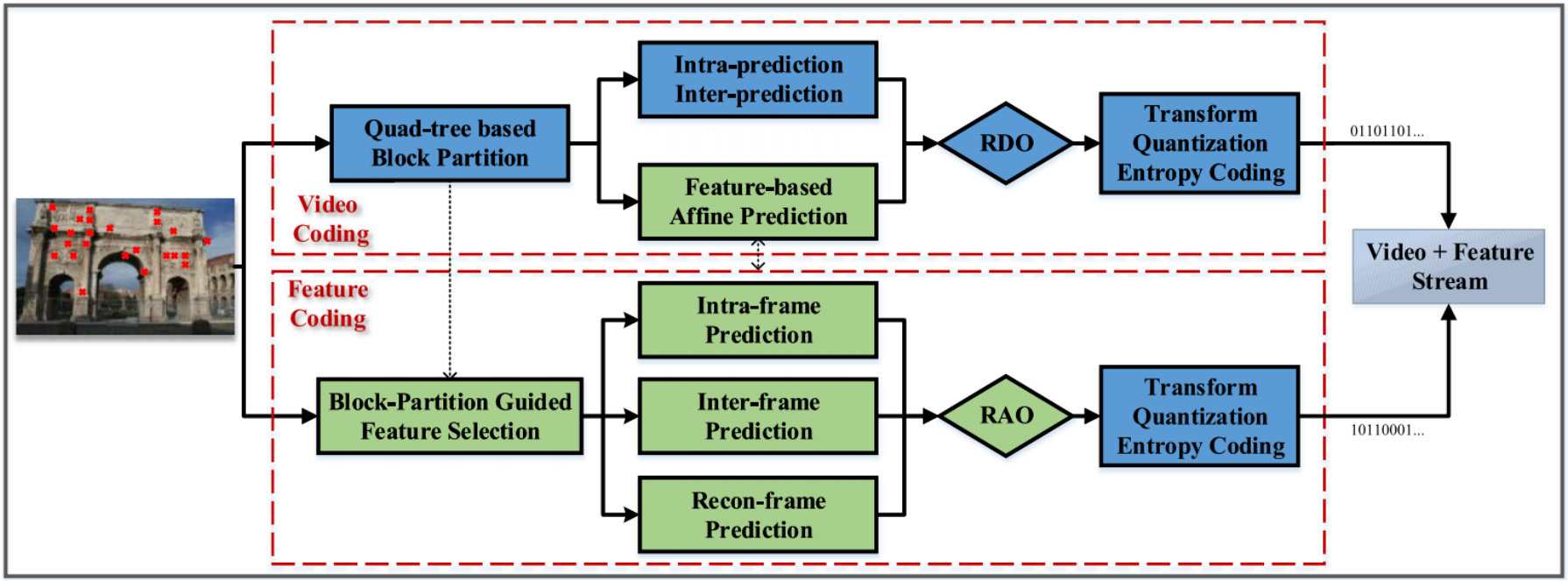
- Video Coding and Local Feature Descriptors
  - Video coding: towards high-efficiency transmission and restoration
  - Local Feature Descriptor: losing image texture info.



1. How to use feature info. to guide video coding?
2. How to use video info. to guide feature coding?

Joint R-D and R-A Optimization

- Framework
  - Feature coding
  - Video coding



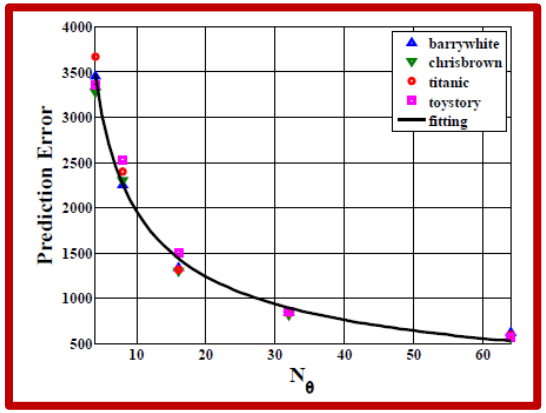
Rate-Distortion Optimization

- Multiple prediction modes
  - Intra-frame prediction
    - Search most-similar feature in the current frame
  - Inter-frame prediction
    - Search most-similar feature in the previous frame
    - Using motion vector to speed up the searching process
  - Reconstructed-frame prediction
    - Fast feature extraction: coding scale, orientation parameters
    - How to achieve optimal orientation quantization?

$$\widetilde{N_\theta} = \underset{N_\theta}{\operatorname{argmin}}(D(N_\theta) + \lambda \cdot R(N_\theta))$$

$$R(N_\theta) = \log_2(N_\theta) \qquad D(N_\theta) = aN_\theta{}^b + c$$

$$\widetilde{N_\theta} = \left(\frac{-\lambda}{ab\ln 2}\right)^{1/b}$$

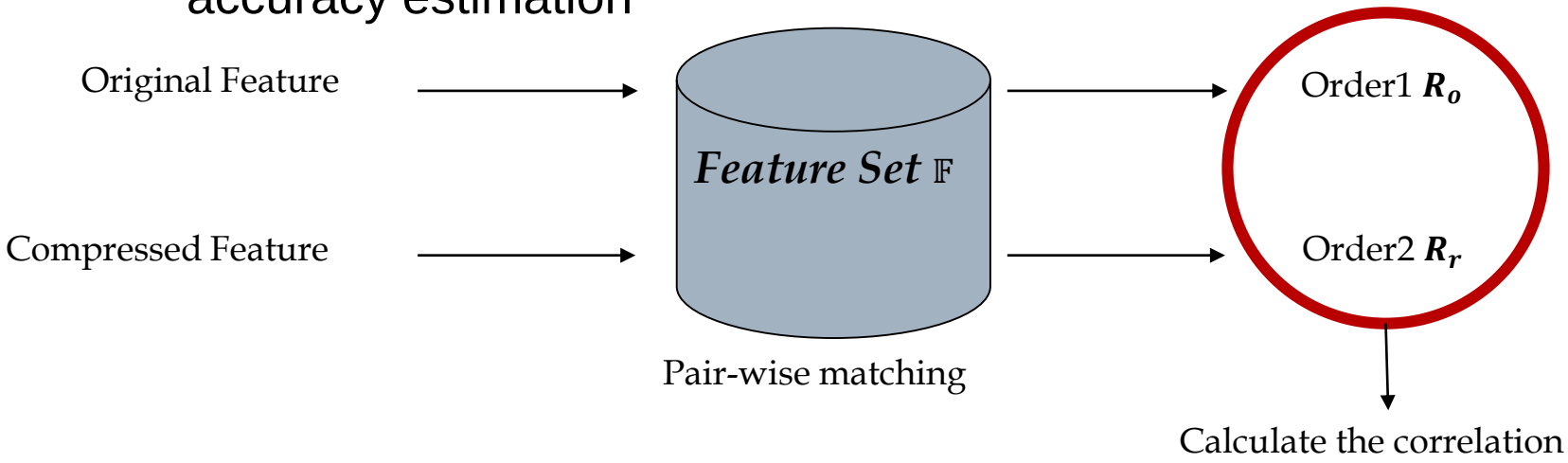


Rate-Accuracy Optimization

- Rate-Accuracy Optimization Model

$$\min(J_A),\ where\ J_A = D_A + \lambda_A R$$

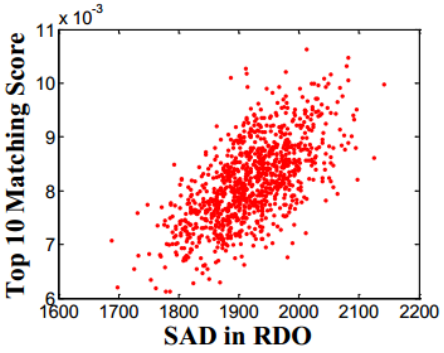
- Local Feature Descriptors are used for matching and retrieval
  - Using the matching performance degradation as the accuracy estimation



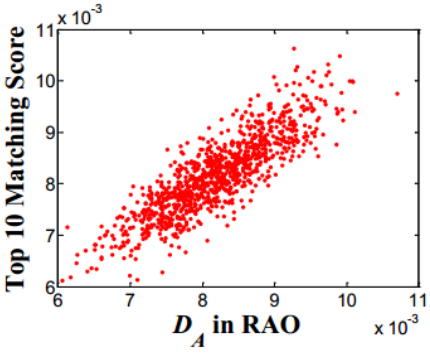
$$D_A \triangleq 1 - SROCC(\boldsymbol{R_o}, \boldsymbol{R_r}) = \frac{6\sum\left(r_o^i - r_r^i\right)^2}{K(K^2-1)}$$

Rate-Accuracy Optimized Compression of Local Feature Descriptors

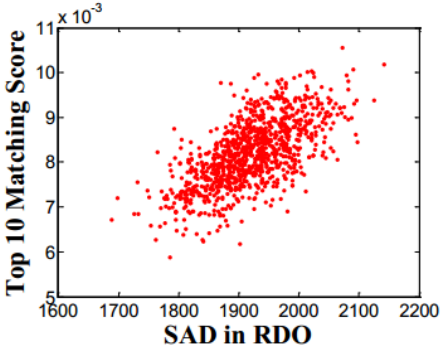
Experimental Results



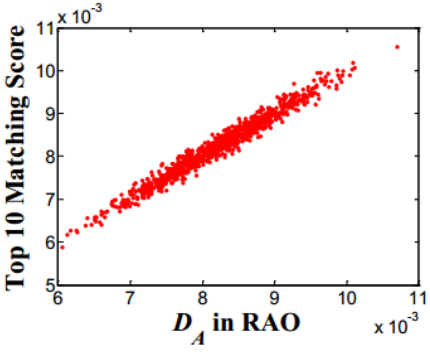
(a) RDO (random feature set)



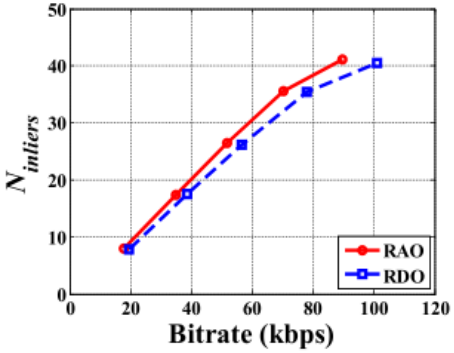
(b) RAO (random feature set)



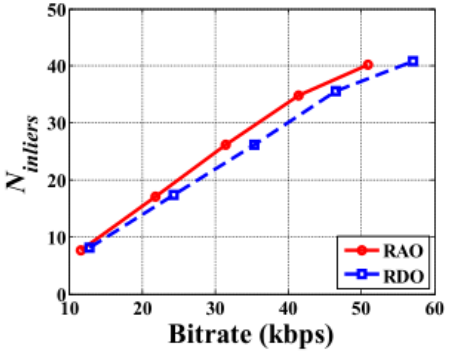
(c) RDO (reconstructed feature set)



(d) RAO (reconstructed feature set)



(a) $QP_F = 35$



(b) $QP_F = 45$

Xiang Zhang; Siwei Ma; Shiqi Wang; Xinfeng Zhang; Huifang Sun; Wen Gao, A Joint Compression Scheme of Video Feature Descriptors and Visual Content, IEEE Transactions on Image Processing2017, 26(2), 633 - 647,

Joint R-D and R-A optimization

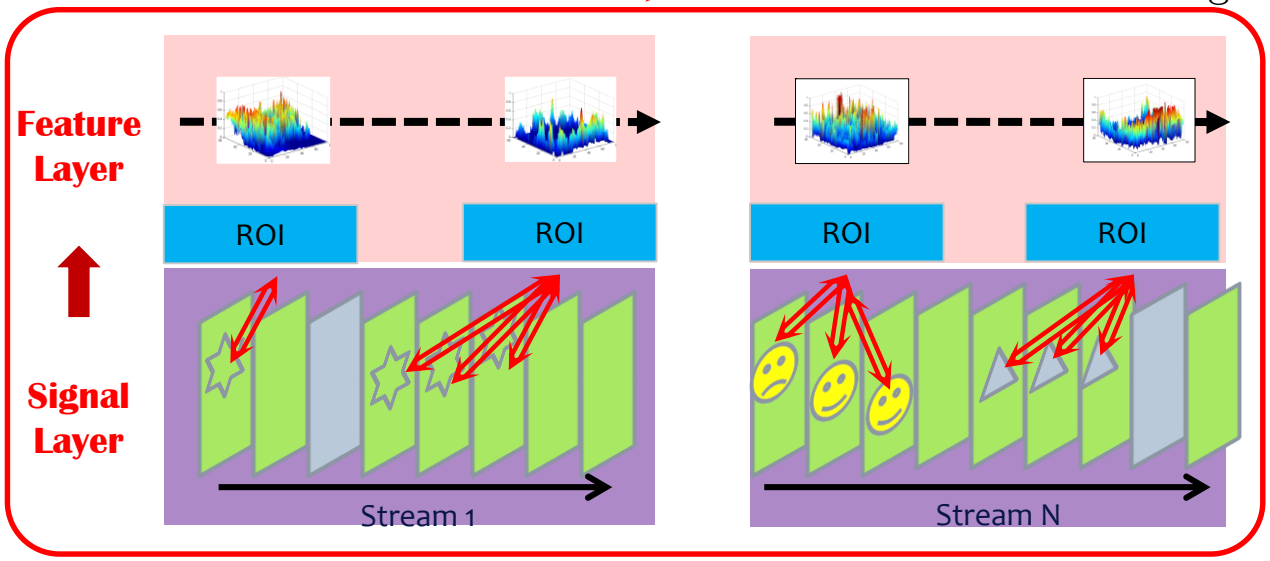
Input: Data $\boldsymbol{V}$ **Output**: Optimal representation $\boldsymbol{S}^*$

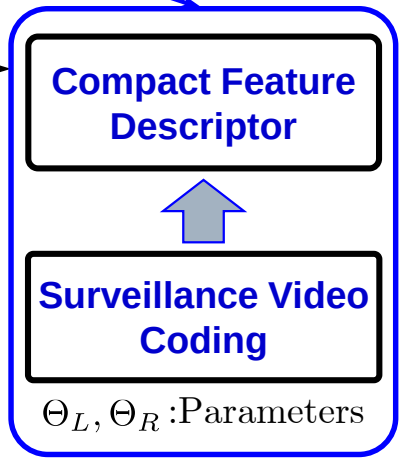
$$S^* = \arg\min_S \mathcal{L}(S|V;T;\Theta_L) + \lambda\mathcal{R}(S;\Theta_R)$$

Representation fidelity

Processing task

Cost





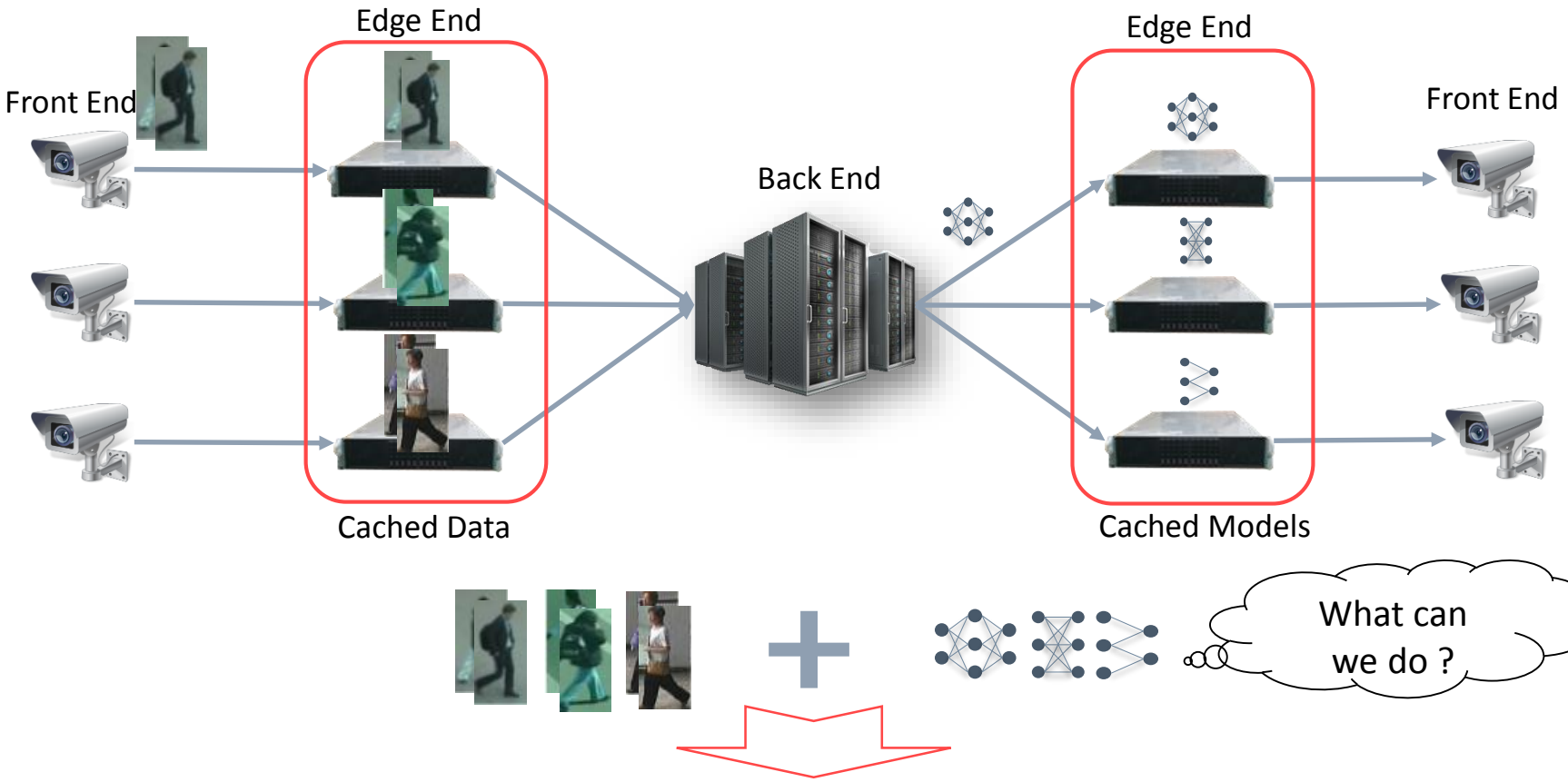
Solution? On going

Enable Tech 4: model coding

- Joint optimization between video coding and feature coding, and maybe model coding
- Our key contribution:
  - Joint R-D and R-A Optimization

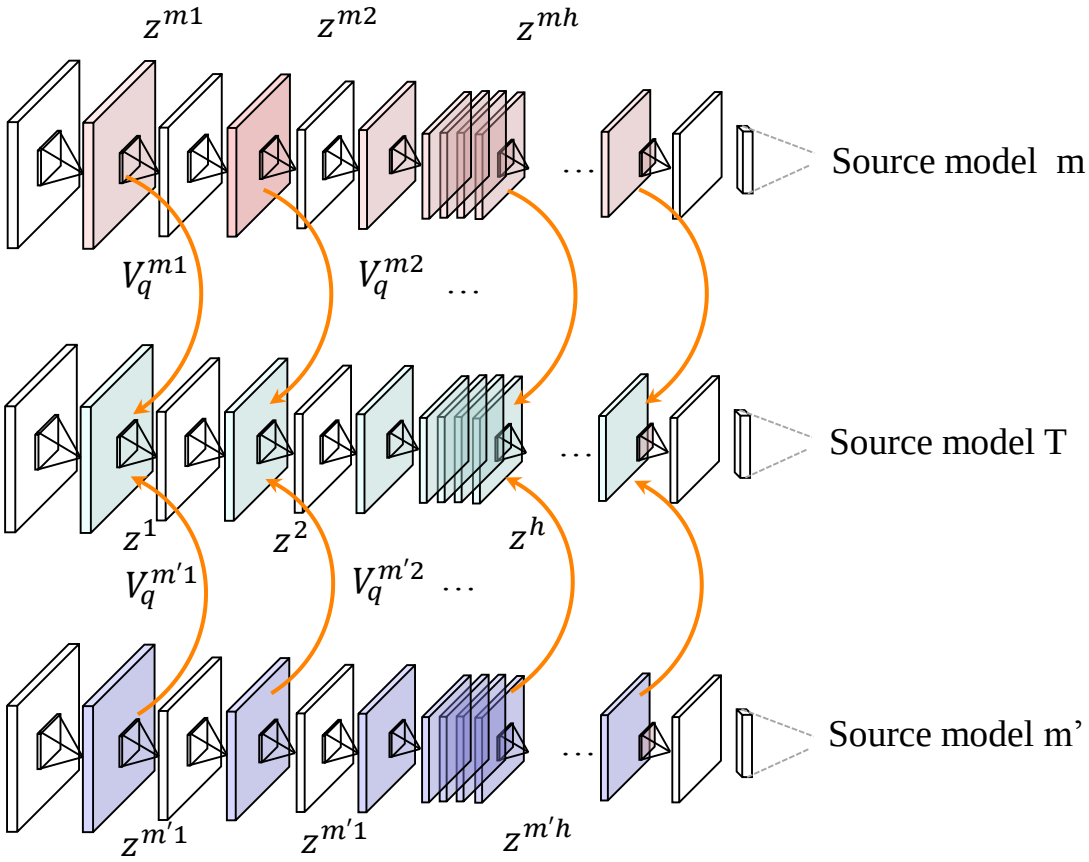
Data and Model Interaction in Digital Retina

- The models are generated at back-end, utilized at edge end and transmitted to the front end.



Multiple Model Reuse

- **Model Reuse aims to reuse existing models to promote target model training.**



Reuse
More Domain
Knowledge!

Multiple Model Reuse

- Input :
  - Data x | Label y
  - Fixed source Model m
- Learning Parameter:
  - Target Model W^h (h layers)
  - Transformation V_Q^{mh} in the h layers
- Learning Objective:

$$\epsilon(\Theta_T;\ \{\mathbf{x_n}, y_n\}) = \frac{1}{N_l}\sum_{n=1}^{N_l}\left(\underbrace{L(f_T(\Theta_T;\ \mathbf{x}_n), y_n)}_{\textit{Target Task}} + \gamma \underbrace{\sum_{n=1}^{N_l+N_u}\sum_{h=1}^{H'<H} R\left(\mathcal{Z}_{\boldsymbol{n}}^h;\ \{\mathcal{Z}_n^{mh}\}\right)}_{\textit{Reuse term}}\right)$$

$$R\left(\mathbf{z}_{\boldsymbol{n}}^h;\ \{\mathbf{z}_n^{mh}\}\right) = \sum_{m=1}^{M}\alpha_m\left\|\mathcal{Z}_n^{mh} - \mathcal{Z}_n^h \times_1 V_1^{\mathrm{mh}} \ldots \times_Q V_Q^{mh}\right\|_F^2$$

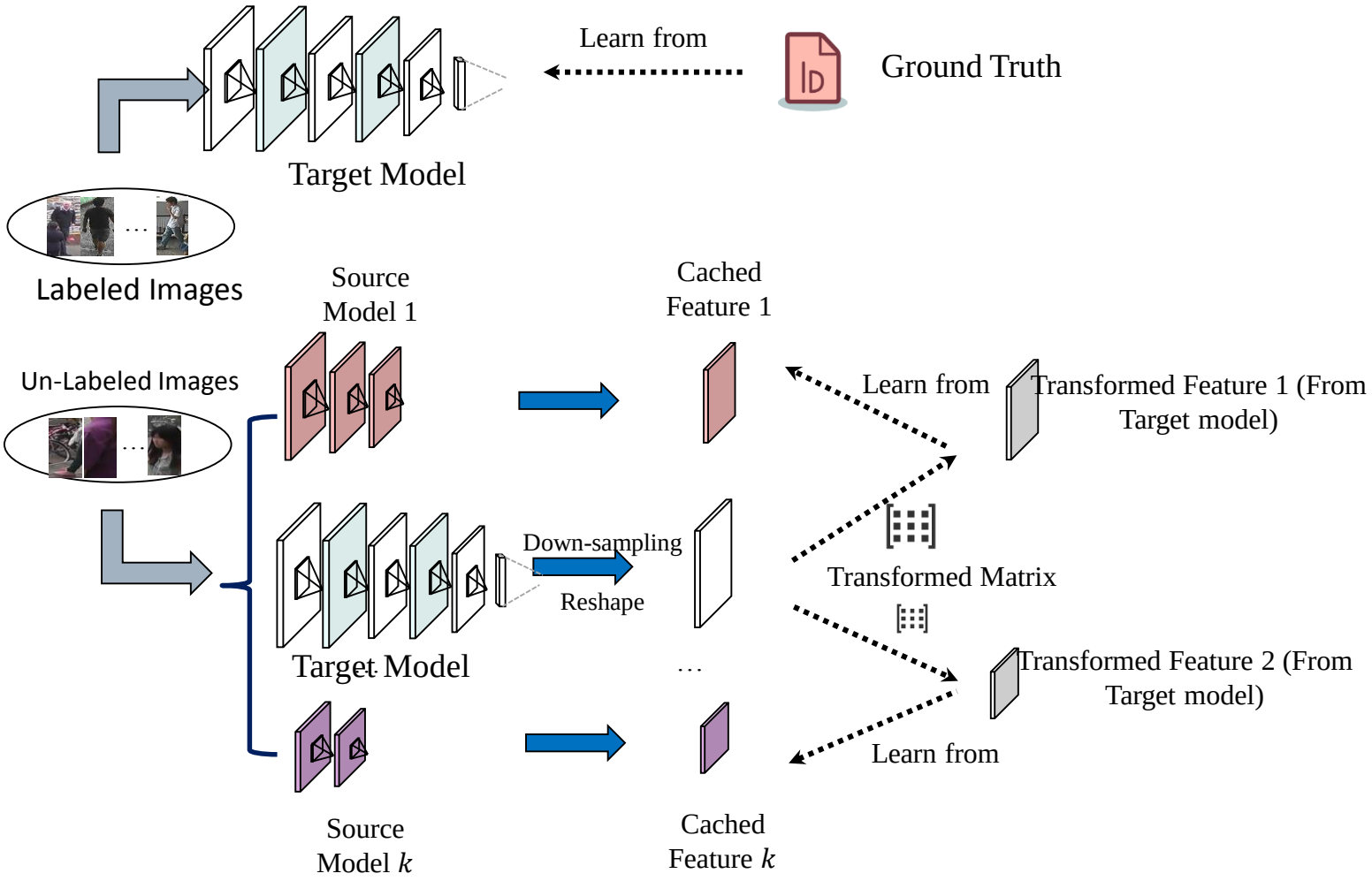
Reuse from Multiple Source Models

- Theoretical Result:

$$\mathcal{R}\left(f_{T,N_l}\right) \le 2M\sum_{m=1}^{M}\alpha_m^2\mathcal{R}(V^m f_m)(r^2/\gamma + 1)$$

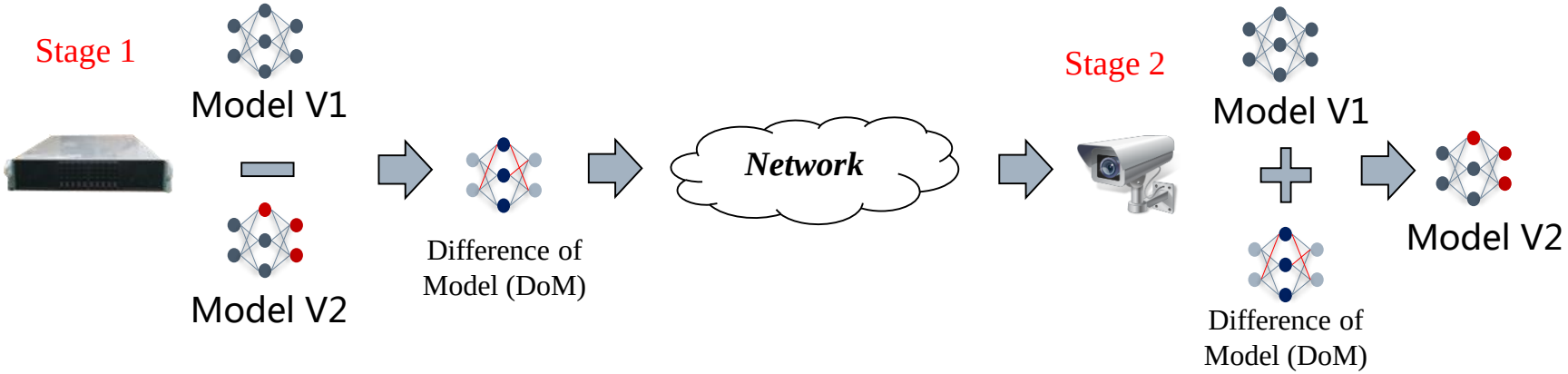
The expected risk of the target model is guaranteed to be low when the source models are well-trained

Learning Details in Multiple Model Reuse



Difference of Model (DoM) Compression

- Problem: Model redundancy in frequent model updating takes large transmission cost
- Solution: Compress difference of Model for model prediction.
  - Stage 1: Compute and encode difference of model.
  - Stage 2: Predict model with reconstructed differences



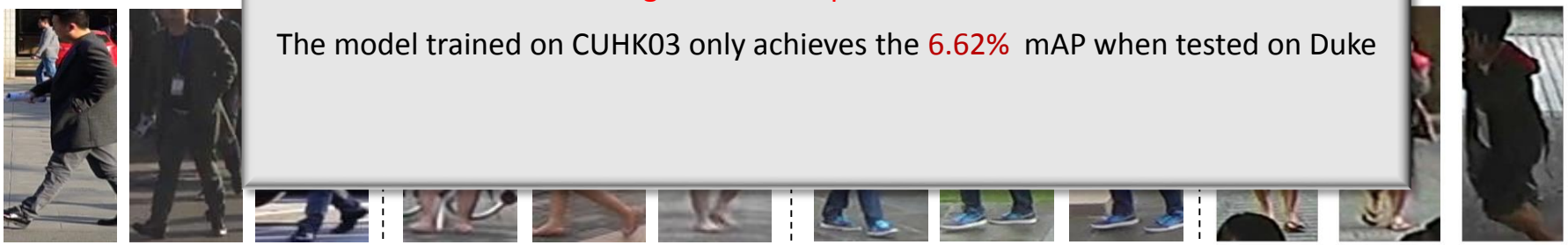
Experiments Setup

- We conduct experiments on person ReID in smart city application
  - V… sets.

Why?

Huge Domain Gap in Person ReID

The model trained on CUHK03 only achieves the 6.62% mAP when tested on Duke

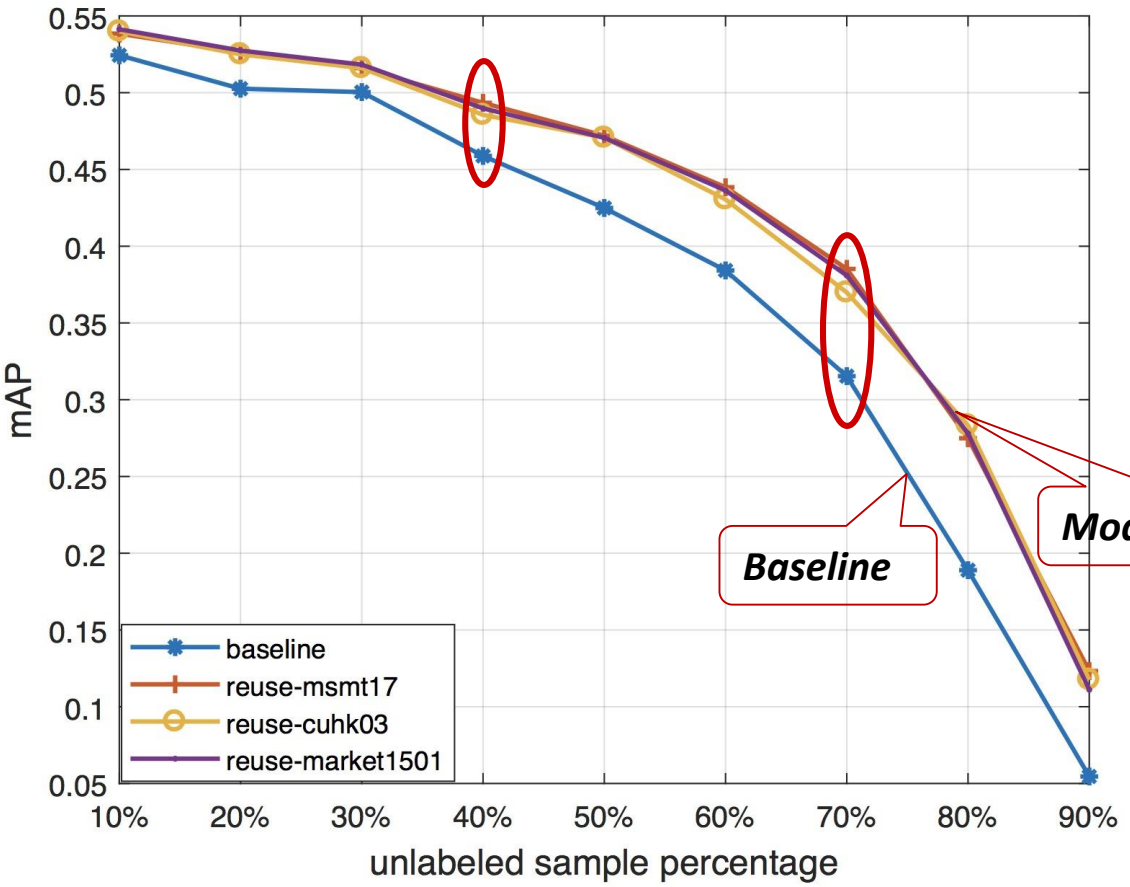


MSMT17 Market1501 Duke CUHK03

Dataset	Images/IDs	Train	Test
Duke	36,411/1,812	16,522/702	19,919/1,110
Market1501	32,688/1,501	12,936/751	19,752/750
MSMT17	126,441/4,101	32,621/1,041	93,820/3,060
CUHK03	28,192/1,467	26,264/1,367	1,928/100

Evaluation: mean Average Precision (mAP)

Results of Model Reuse



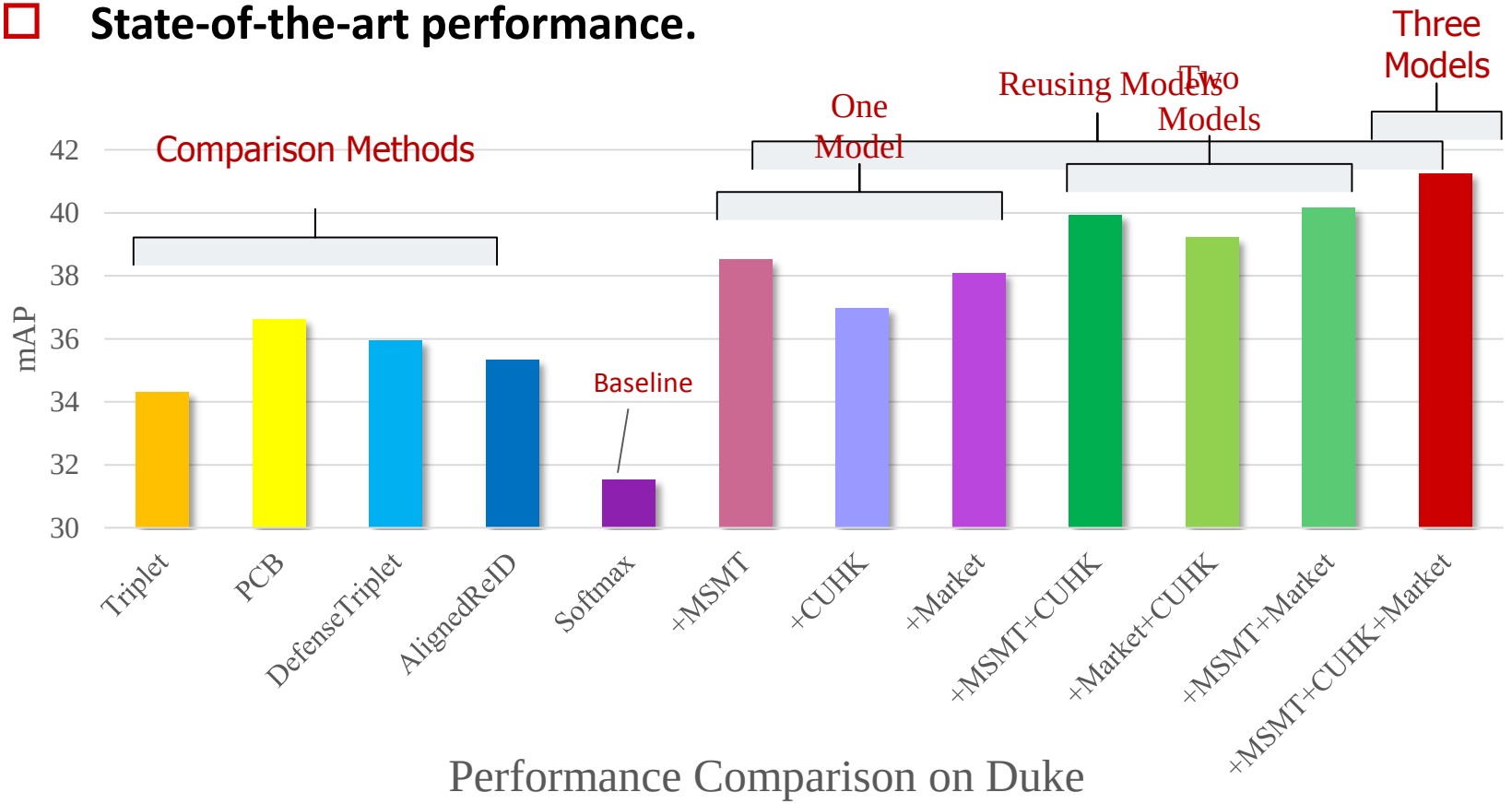
More training data, more performance gain

Model reuse significantly boosts baseline

The performance of reusing different **single models** on **Duke** train set by varying the **percentage** of unlabeled data

Results of Model Reuse

- Reusing additional models achieves better performance.
- More reused models, better performance.
- State-of-the-art performance.



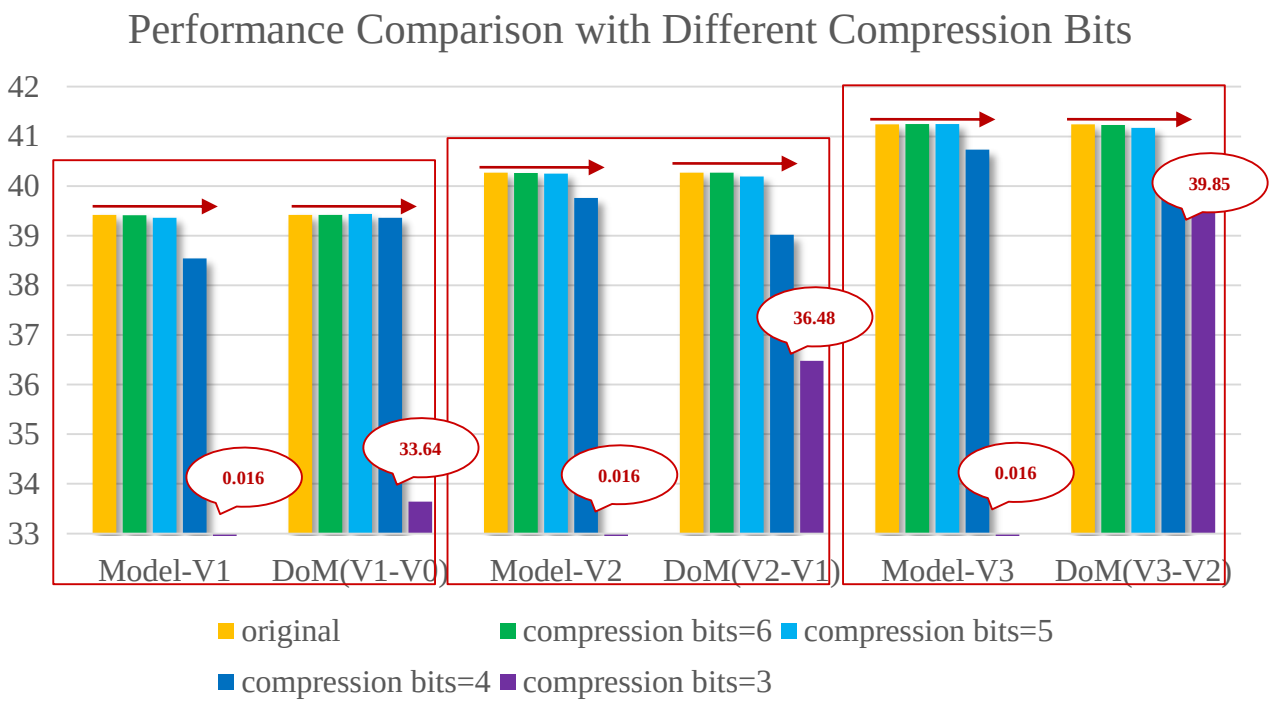
Performance Comparison on Duke

Results of DoM Compression

- Lower compression bits, lower performance.
- Using model sharing information improves performance, especially for low bits compression.

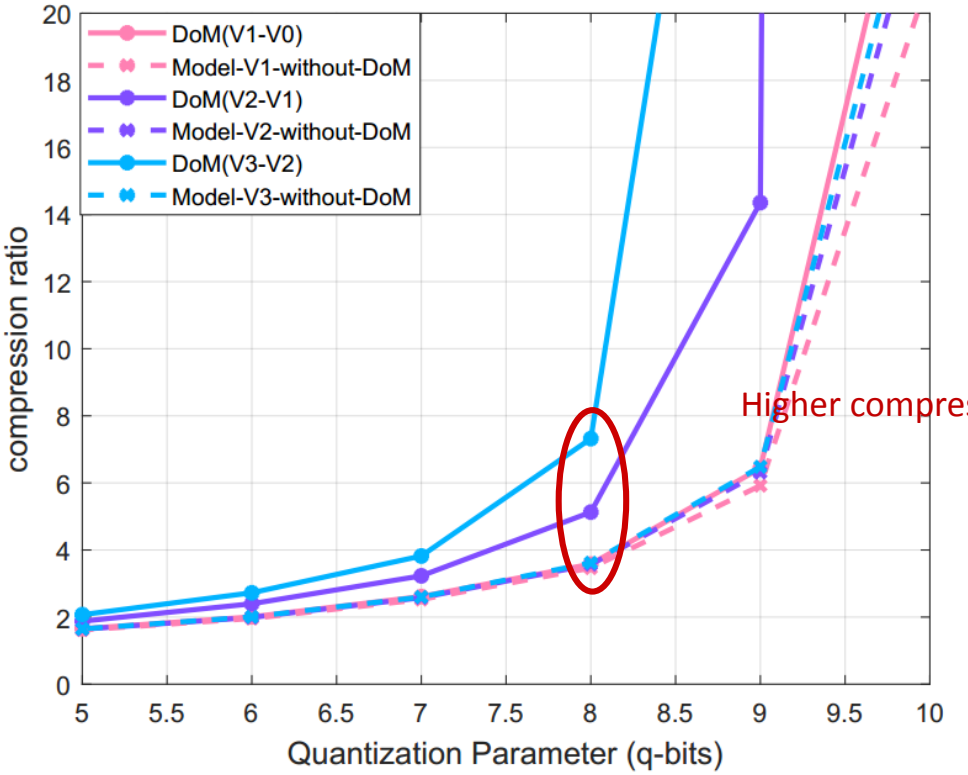
Model Setting

- Model v1/v2/v3: models with different performance
- DoM(v2-v1): Difference between model v2 and v1



Results of DoM Compression

- The DoM strategy significantly outperforms the simple single model compression.

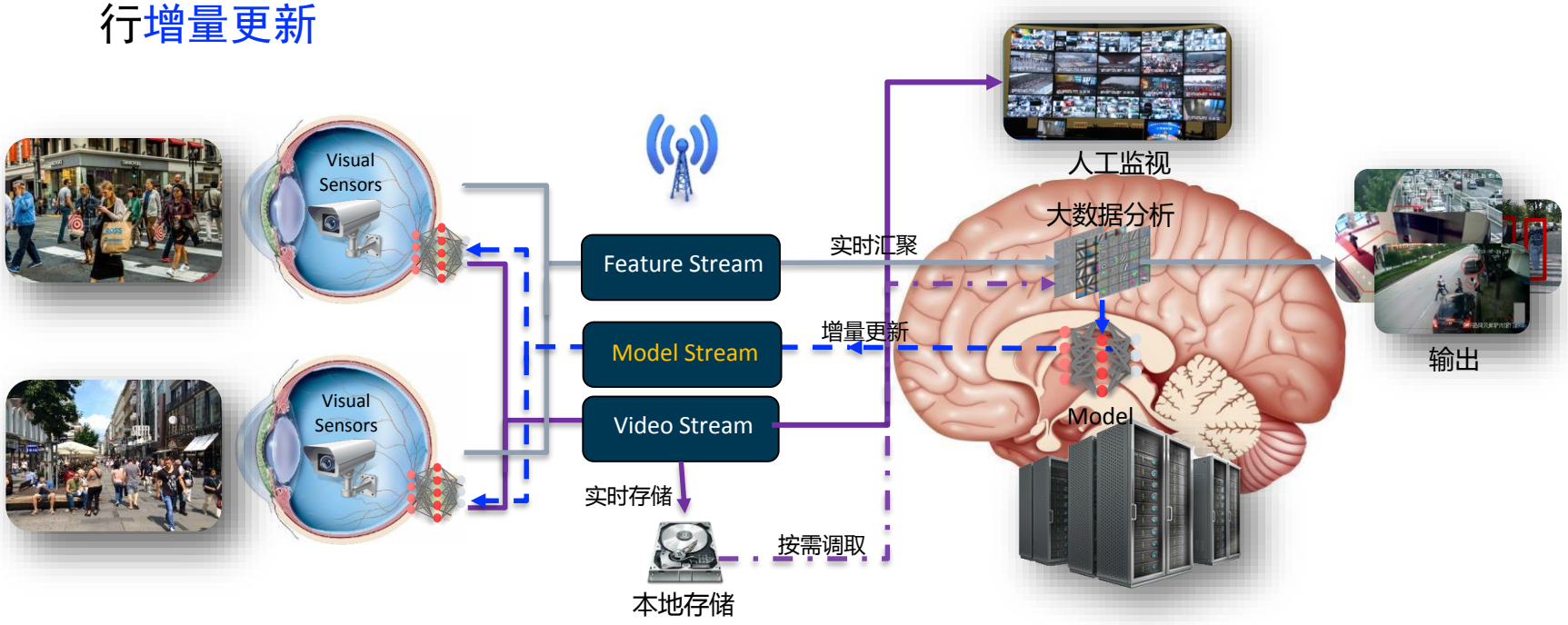


Higher compression ratio

The compression ratios between with/without DoM

Digital Retina： three coding streams

- Video coding： 用于存档和人工监视，存放本地服务器，按需调取到云端大脑
- Feature coding： 实时汇聚到云端大脑，用于大数据分析与检索
- Model coding： 在云端大脑训练，针对不同端/边设备迁移与压缩，进行增量更新



First digital retina chip: GV9631

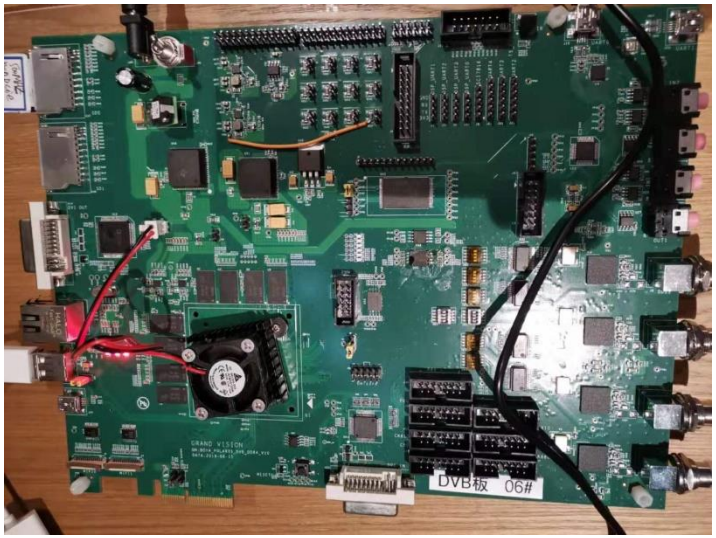
鸿图™ GV9531

数字视网膜技术的完整诠释

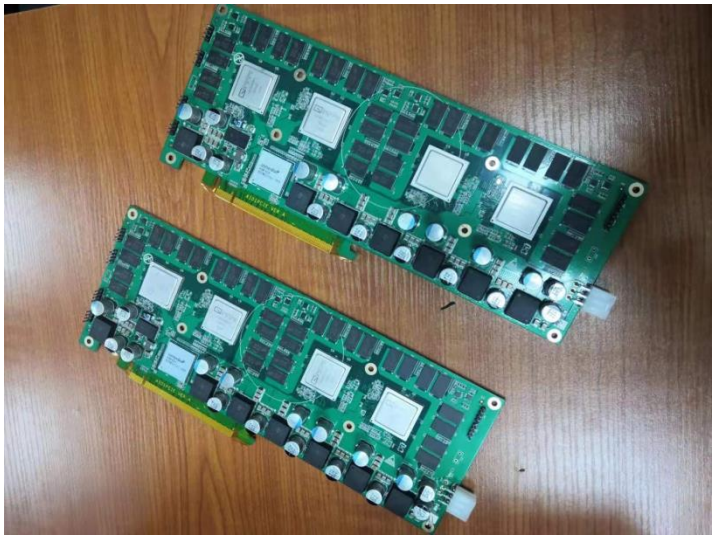
- 高效的视频编码，AVS2/H.265
- 高效的特征编码，CDVS/CNN
- 视频与特征联合编码
- 功能/性能（模型和参数）的软件定义
- 大幅降低视觉智能计算成本
- 符合自主可控要求

编解码性能	支持AVS2，AVS+，H.265，H.264 支持4K@30分辨率的视频编解码双工处理 支持多标准音频编解码 支持网络适配及信道容错编解码功能， 支持CBR、VBR码率控制，支持低延迟编码
视频处理	支持数字降噪、图像增强、去雾等预处理 视频缩放、叠加、镜像等后处理
智能处理	支持底层视觉特征描述及压缩（CDVS） 5Tops CNN算力 支持背景帧提取和ROI提取 支持6路视频运动目标检测和车辆、行人、人脸识别及结构化分析
处理器内核	国产多核高性能CPU
外设接口	DDR4/3/3L 接口 视音频输入输出接口 SPI Flash，Nand Flash接口 PCI-E接口、SDIO 2.0 接口、USB 2.0 接口、千兆网口
生产工艺	28nm工艺国内流片，封装大小为25mmx25mm

GV9531 reference design card



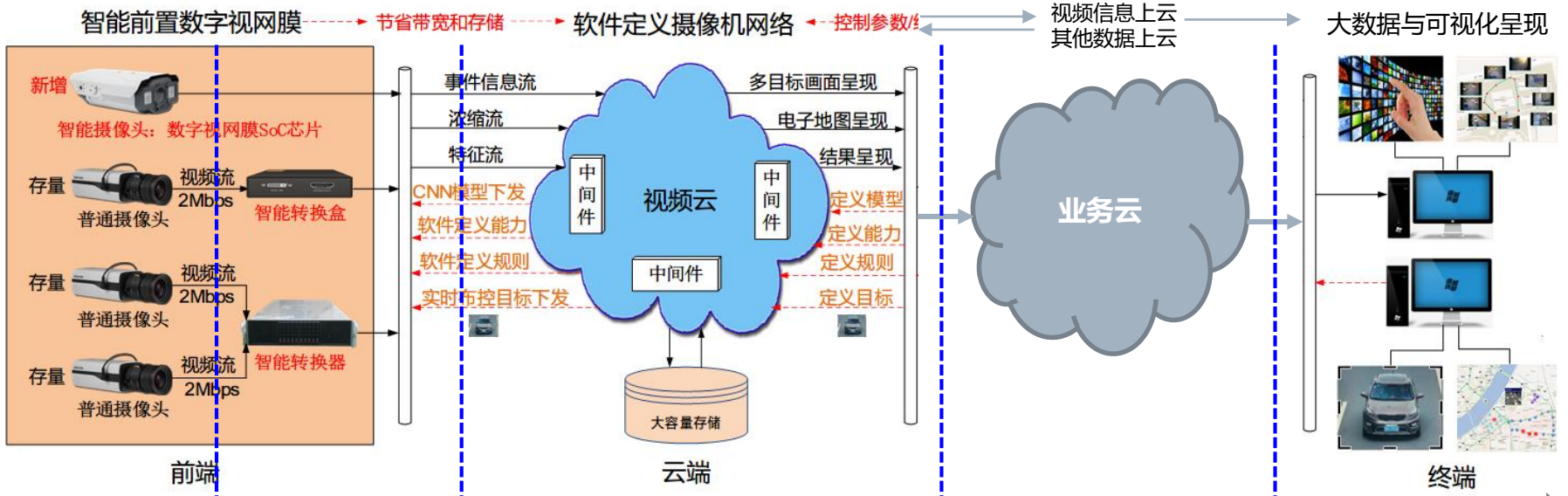
提供全功能的参考设计给应用开发用户



4颗芯片的PCIe标准板卡产品
可支持超高清视频编转码及AI服务器的产品开发

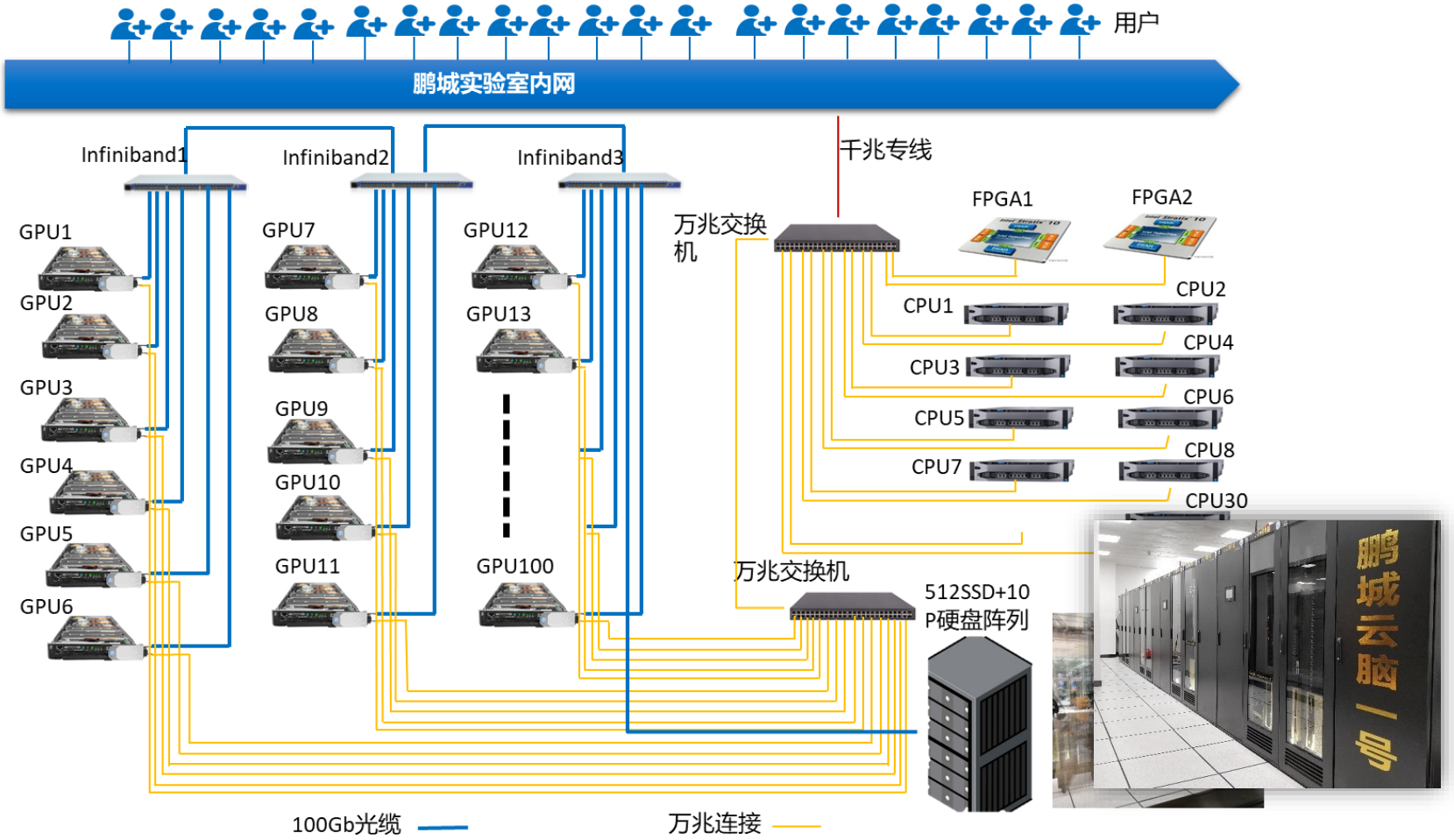
Eco system for digital retina: system

仿人类视网膜机理，通过算法和计算对视频逐级提取和浓缩，为大数据云计算提供高质量的视频信息数据



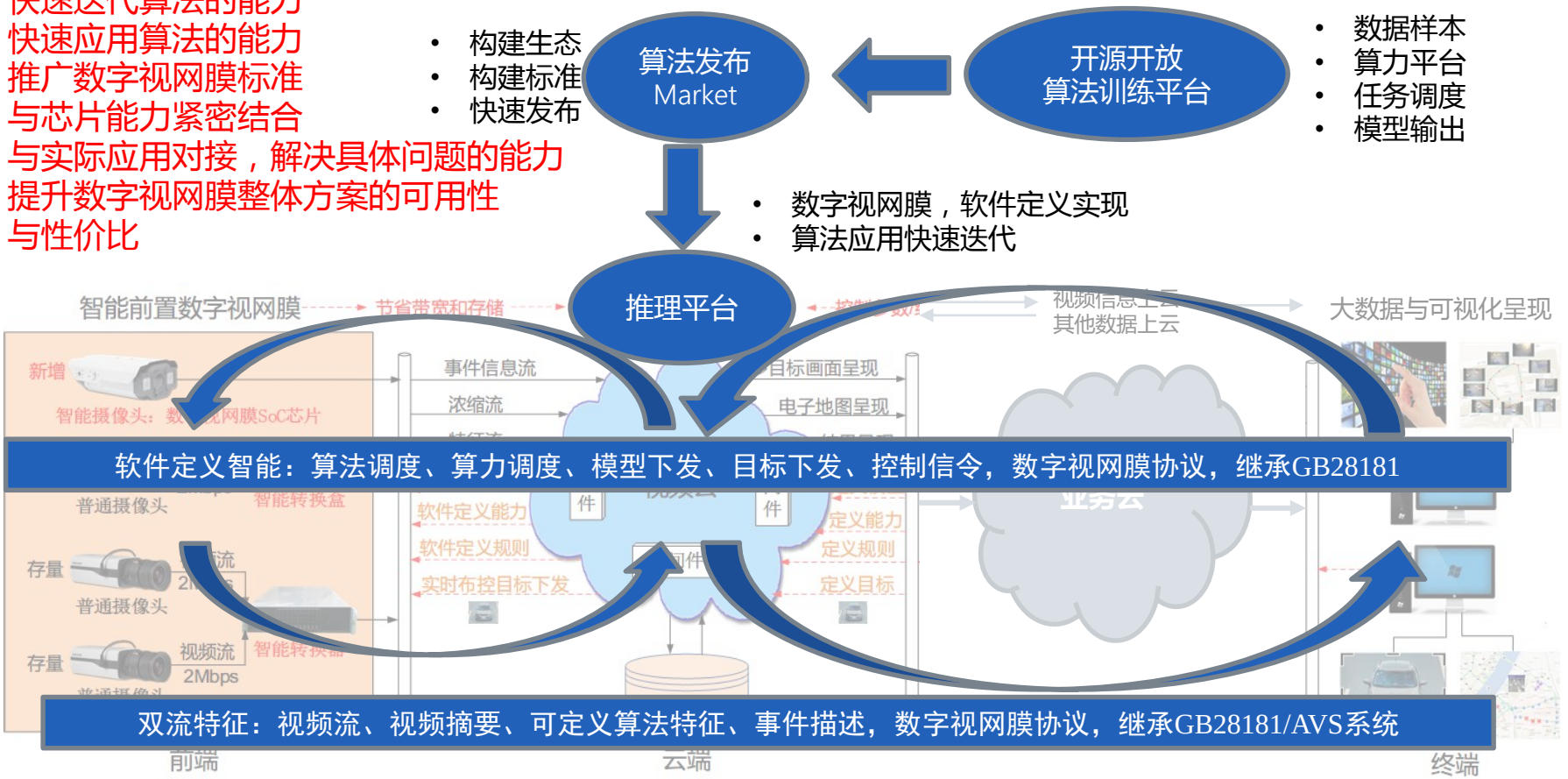
Eco system for digital retina: cloud

- PC AI-Cloud1：100P GPU、10PB storage



Eco system for digital retina: open source

- 快速迭代算法的能力
- 快速应用算法的能力
- 推广数字视网膜标准
- 与芯片能力紧密结合
- 与实际应用对接，解决具体问题的能力
- 提升数字视网膜整体方案的可用性与性价比



Summary

- Current CVS is not efficient in many senses
- Digital retina 1.0, 8 items
  1. Unified time stamp
  2. The exact geographical location
  3. High efficient video coding
  4. High efficient feature coding
  5. Joint optimization between video coding and feature coding
  6. High efficient model updating
  7. Top-down attention
  8. Software defined function X
- Standards for Digital retina 1.1, mostly ready
- Future work of Digital Retina?
  - Implementation of Digital retina 1.1, off-line, GPU, and ASIC
  - Further effort on Digital retina 1.1
    - Making feedback functions between camera and perceptual system
  - Digital retina 2.0?
    - Spike coding based VCC system

Thanks!
Q&A?