边缘智能-
边缘计算时代的人工智能

报告人：谢 磊

南京大学

2020-3-15

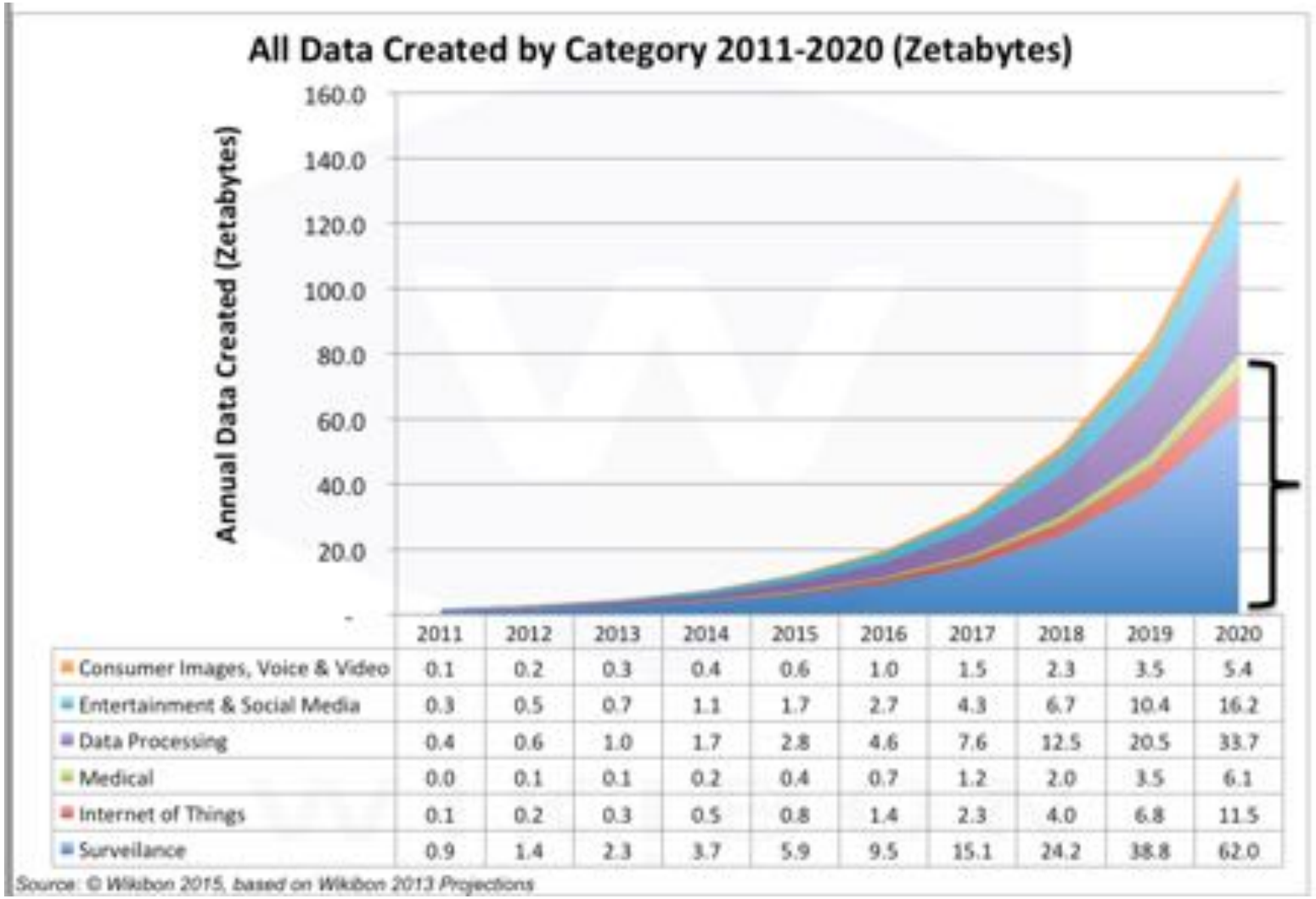
提纲

- 边缘智能简介
- 基于边缘智能的视频大数据分析
- 案例：基于视频分析的抬头检测系统

提纲

- 边缘智能简介
- 基于边缘智能的视频大数据分析
- 案例：基于视频分析的抬头检测系统

“大数据”的产生：云 → 边缘



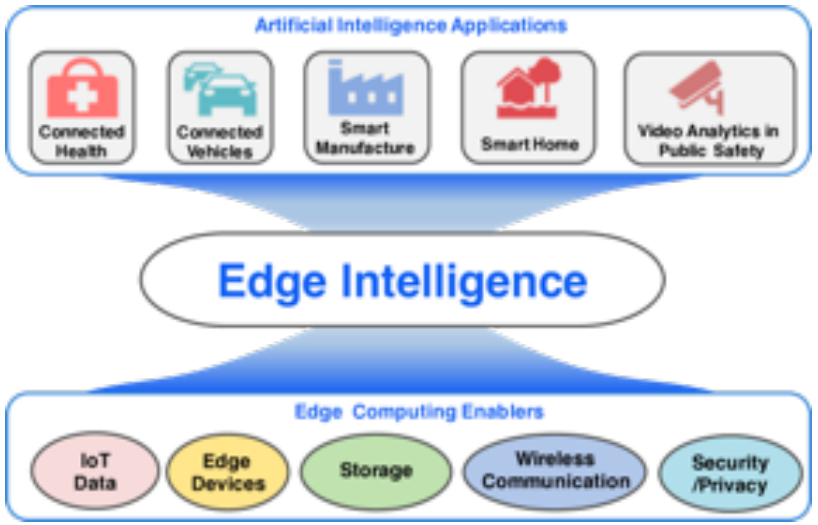
	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Consumer Images, Voice & Video	0.1	0.2	0.3	0.4	0.6	1.0	1.5	2.3	3.5	5.4
Entertainment & Social Media	0.3	0.5	0.7	1.1	1.7	2.7	4.3	6.7	10.4	16.2
Data Processing	0.4	0.6	1.0	1.7	2.8	4.6	7.6	12.5	20.5	33.7
Medical	0.0	0.1	0.1	0.2	0.4	0.7	1.2	2.0	3.5	6.1
Internet of Things	0.1	0.2	0.3	0.5	0.8	1.4	2.3	4.0	6.8	11.5
Surveilance	0.9	1.4	2.3	3.7	5.9	9.5	15.1	24.2	38.8	62.0

预计在2020年，边缘数据将达到80%。

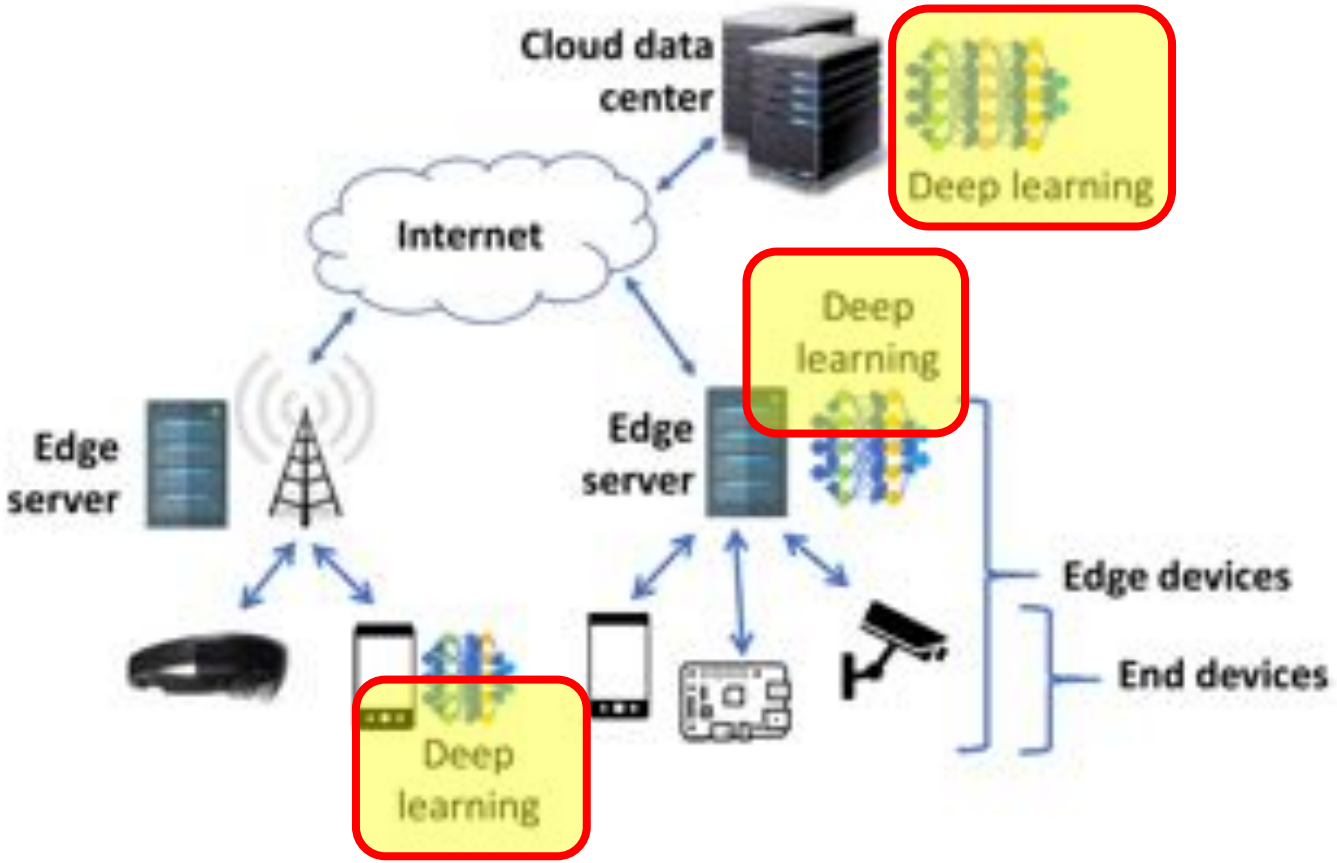
为什么需要边缘智能(Edge Intelligence)?

- 如在云端处理大数据，会产生带宽、时延问题
  - 基于大数据的智能应用(如自动驾驶、AR等)需要实时处理
  - 利用云计算的算力集中式处理，会带来带宽、时延的巨大开销

- 需要把AI能力推送边缘，构成边缘智能(EI)
  - 近端感知数据中的变化，以响应快速变化的环境，提高操作效率
  - 边缘节点本地处理，避免原始数据传输带来的网络负担



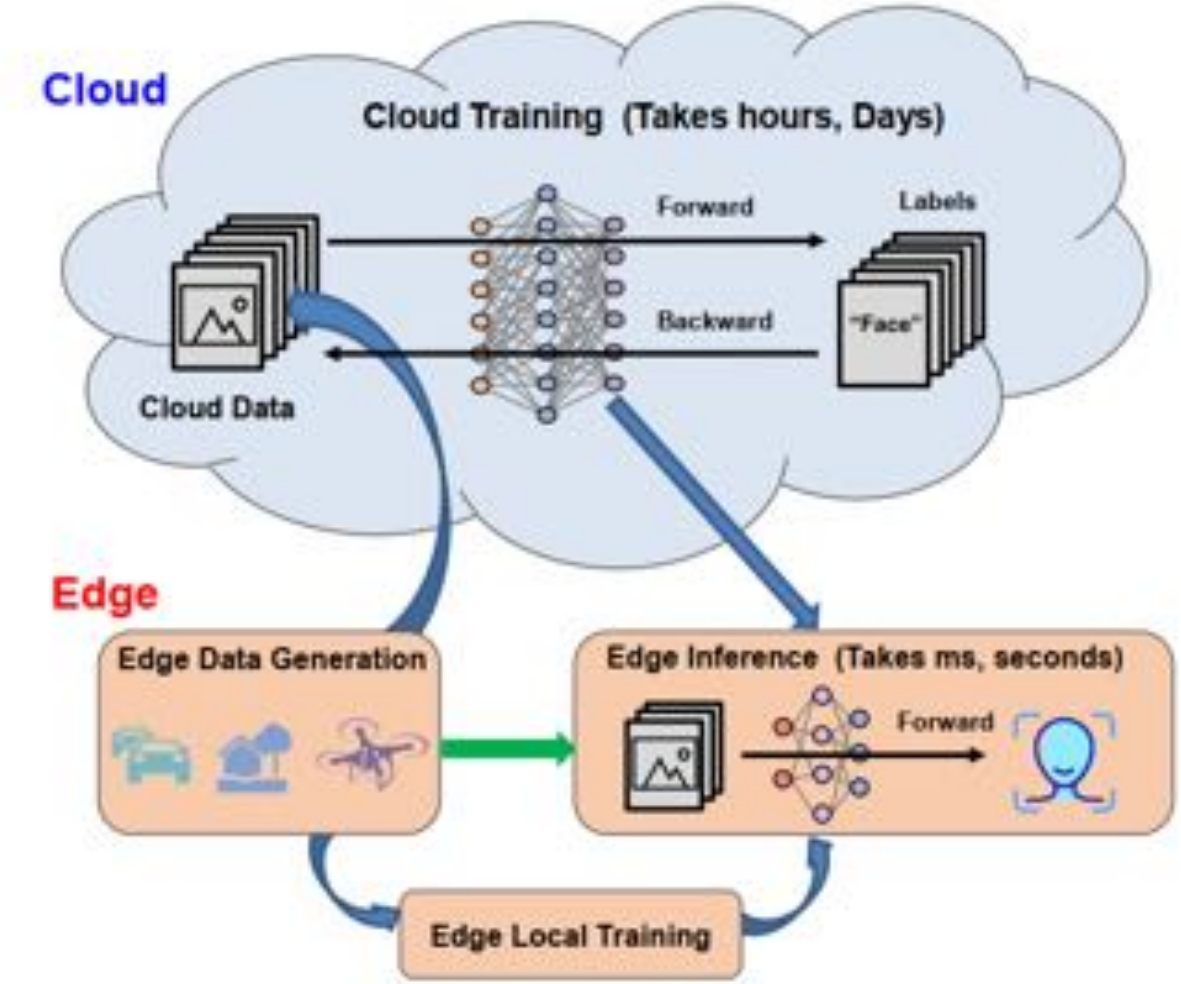
边缘智能(Edge Intelligence)



- 云-边-端协同处理框架
- 将人工智能算法（当前主要是深度学习）推向边缘
- 加速AI的发展
  - 驱动力有四个方面：算法、硬件、数据和应用场景
  - 边缘端丰富的应用场景和多样化的数据促进了AI进一步发展

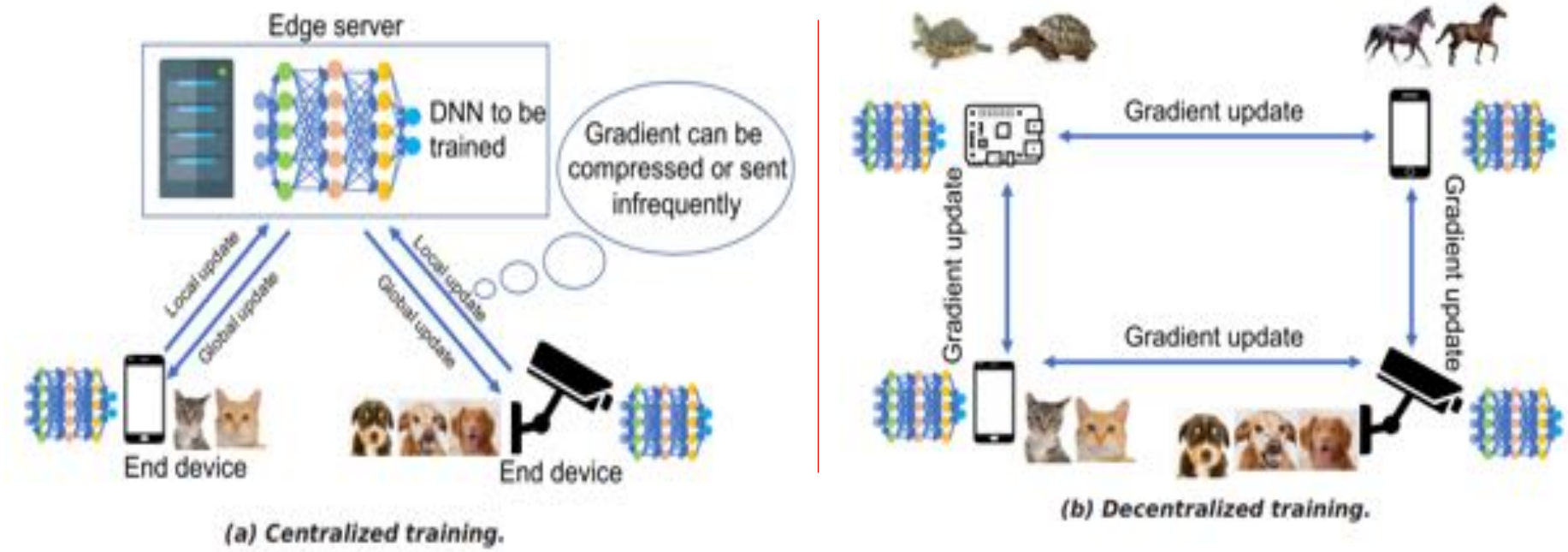
[1] JIASI CHEN AND XUKAN RAN, Deep Learning With Edge Computing: A Review, Proceedings of the IEEE, May 2019.

深度学习的训练和推理



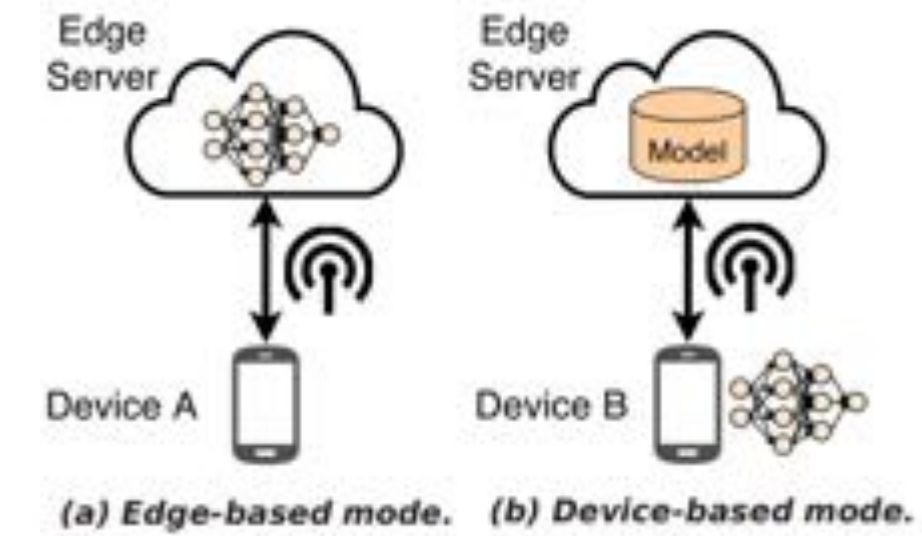
- 经典模式：云端训练+边缘推理
- 选配模式：边缘训练，基于本地的数据特性来进行定制化的训练
  - 通过Pruning、Dropout等方法降低训练参数量
  - 通过迁移学习快速训练
  - 通过强化学习迭代训练

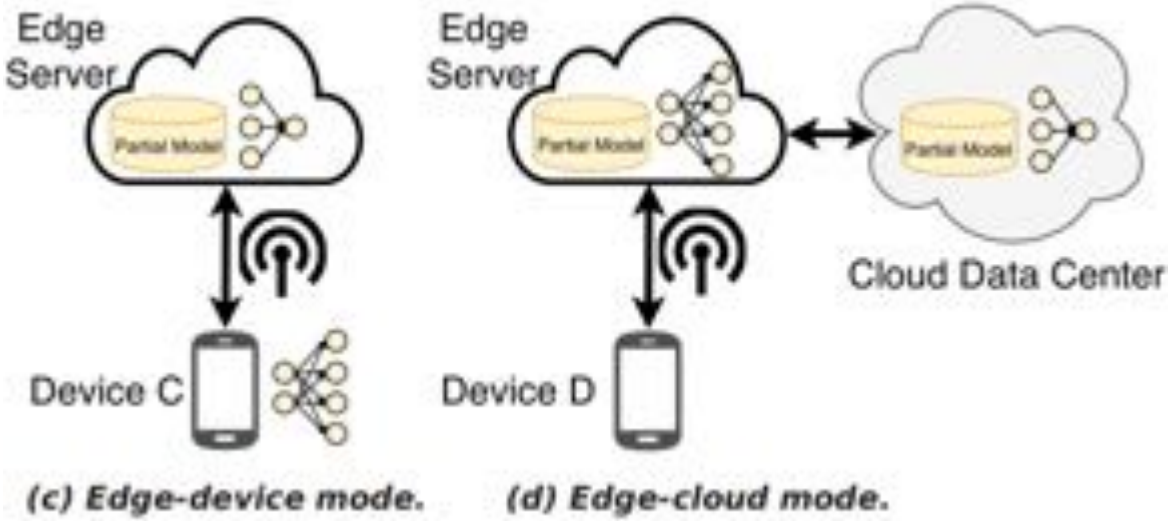
边缘训练模型



- **集中式训练:** Each worker computes local data set, which are then collected by a central parameter server, and the updates sent back to the workers. 代表性模型：联邦学习
- **分布式训练**: Each device computes its own gradient updates based on its training data and then communicates its updates to some of the other devices.
- **两种模型比较**：DNN模型一致性、带宽限制等。

边缘推理模型





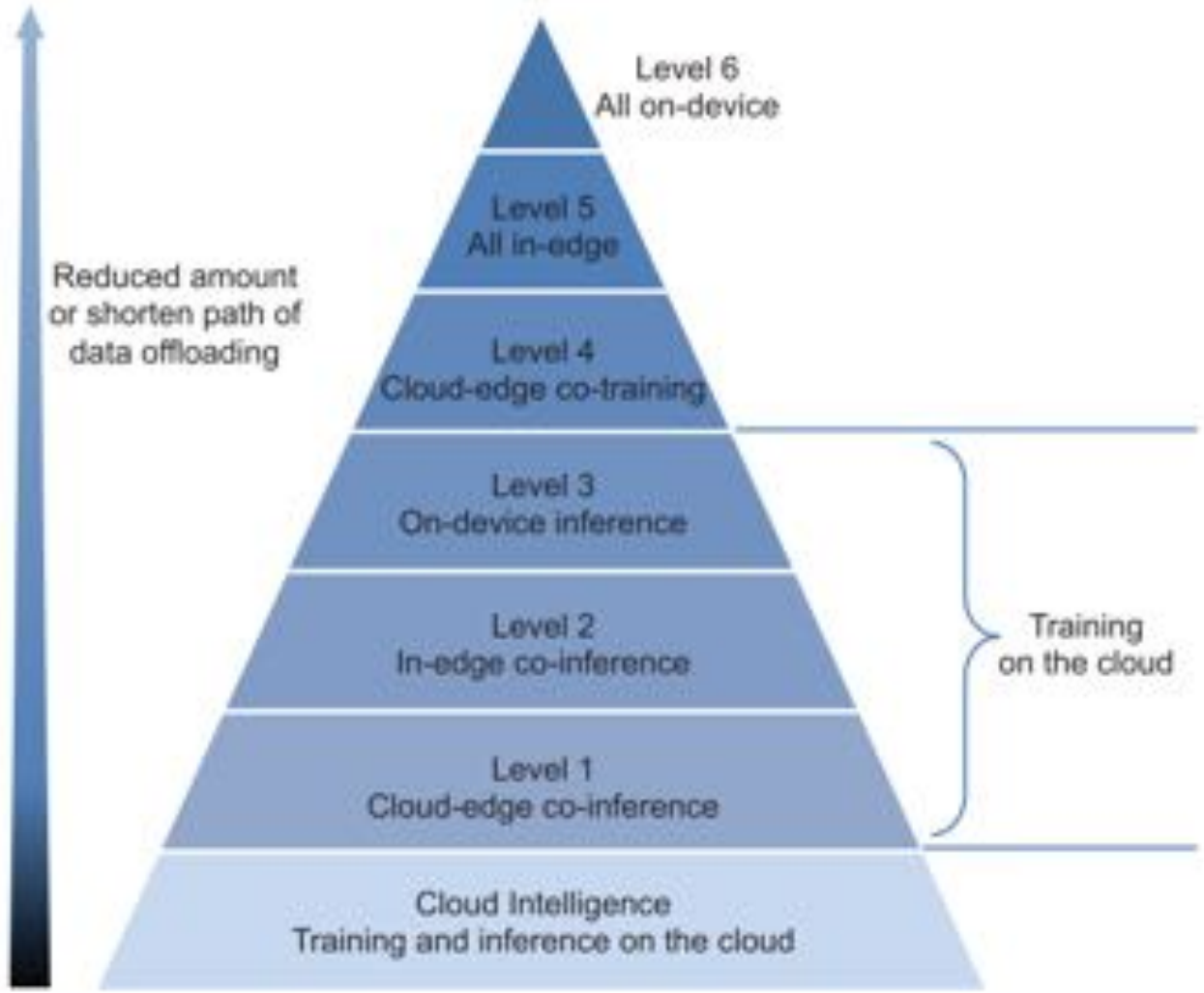
（a）Edge-based mode

（b）Device-based mode

（c）Edge-device mode

（d）Edge-cloud mode

边缘智能的6个等级



- 根据数据卸载的数量和路径长度，我们将边缘智能分成6个等级
- **层级越高，数据卸载的数量和路径长度越少**，从而时延、带宽、安全有优势，但会增加计算延迟和能耗代价

[2] Zhi Zhou, Xu Chen, En Li, Liekang Zeng, Ke Luo, and Junshan Zhang, Edge Intelligence: Paving the Last Mile of Artificial Intelligence with Edge Computing, Proceedings of the IEEE, May 2019.

边缘智能的关键性能指标

（1）**Accuracy（精度）**：从推理中获得的正确预测输入样本数量和总输入样本数量的比值；

（2）**Latency（时延）**：整个推理过程中所占时间，推理时延、传输时延；

（3）**Memory Footprint（内存占用）**：其主要被原始DNN模型大小和加载大量DNN参数的方法所影响。

（4）**Energy consumption（能耗）**：能量效率被DNN模型大小和边缘设备资源所影响；

（5）**Communication overhead（通信开销）**：其依赖于DNN推理方式和可用带宽；

（6）**Privacy（隐私）**：其依赖于处理原始数据的方式；

提纲

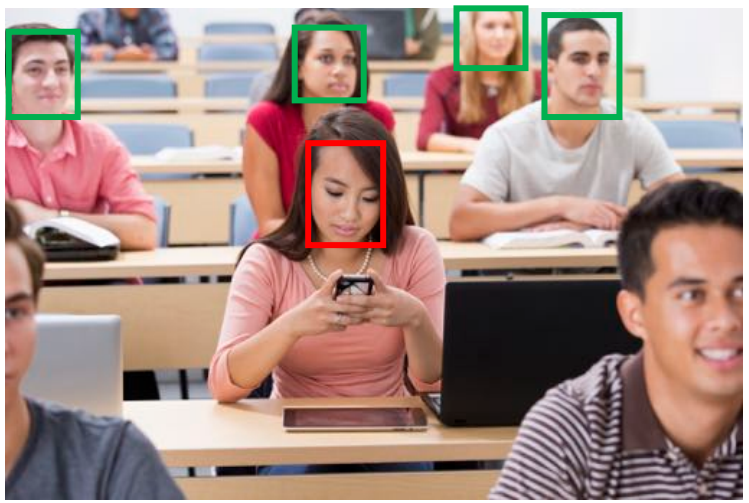
- 边缘智能简介
- 基于边缘智能的视频大数据分析
- 案例：基于视频分析的抬头检测系统

当前视频大数据分析应用特征

- 高分辨率高帧率视频
- 大量/海量监控目标
- 高精度实时响应需求

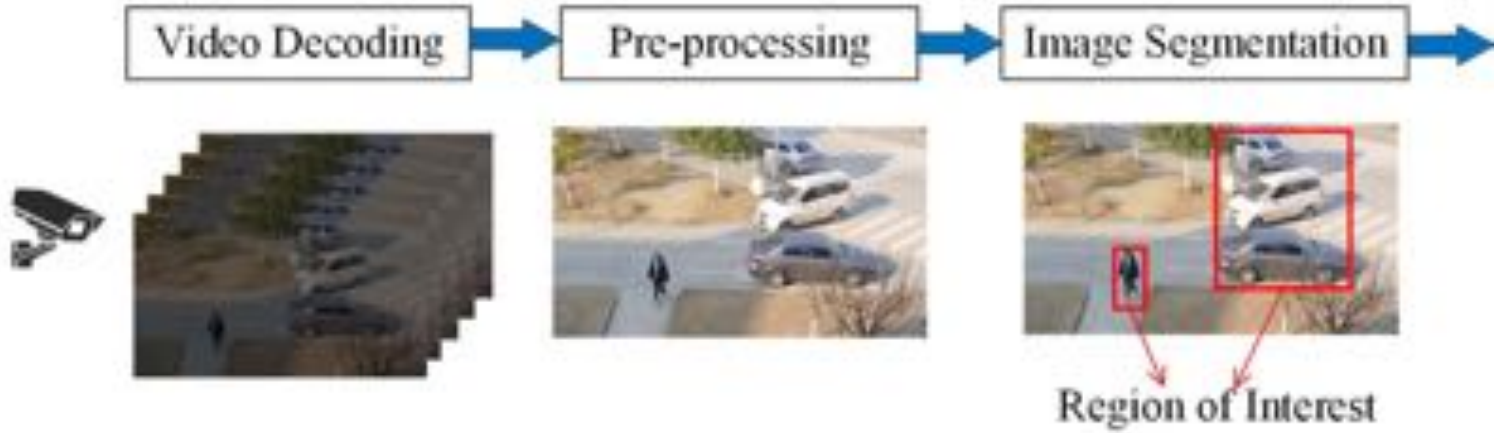
交通车辆监控

人流流量管控



课堂教学评估

视频分析的典型处理流程



代表性成果：Neurosurgeon（ASPLOS'17）

*"**Neurosurgeon:** Collaborative intelligence between the cloud and mobile edge"*

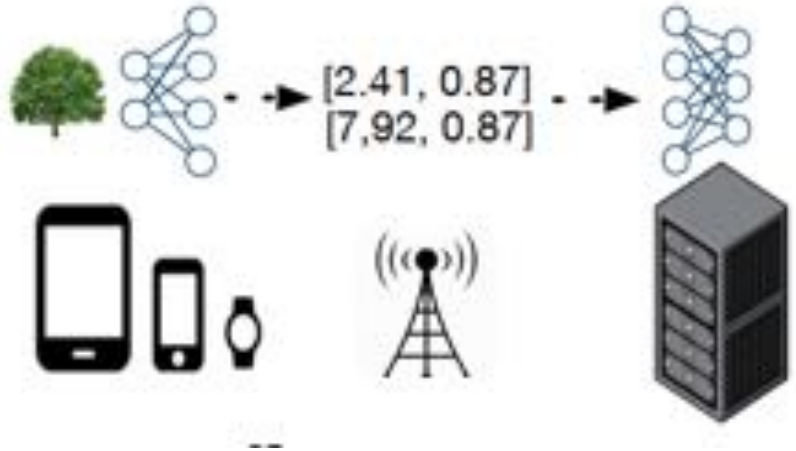
- Tradition approach: **cloud-only or device-only** performs all computation
- Neurosurgeon approach: **partitions computation** between the cloud and mobile device.
  （low Latency + low Energy）

Table 3: Benchmark Specifications

App	Abbr.	Network	Input	Layers
Image classification	IMC	AlexNet [21]	Image	24
	VGG	VGG [26]	Image	46
Facial recognition	FACE	DeepFace [27]	Image	10
Digit recognition	DIG	MNIST [28]	Image	9
Speech recognition	ASR	Kaldi [29]	Speech features	13
Part-of-speech tagging	POS	SENNA [30]	Word vectors	3
Named entity recognition	NER	SENNA [30]	Word vectors	3
Word chunking	CHK	SENNA [30]	Word vectors	3

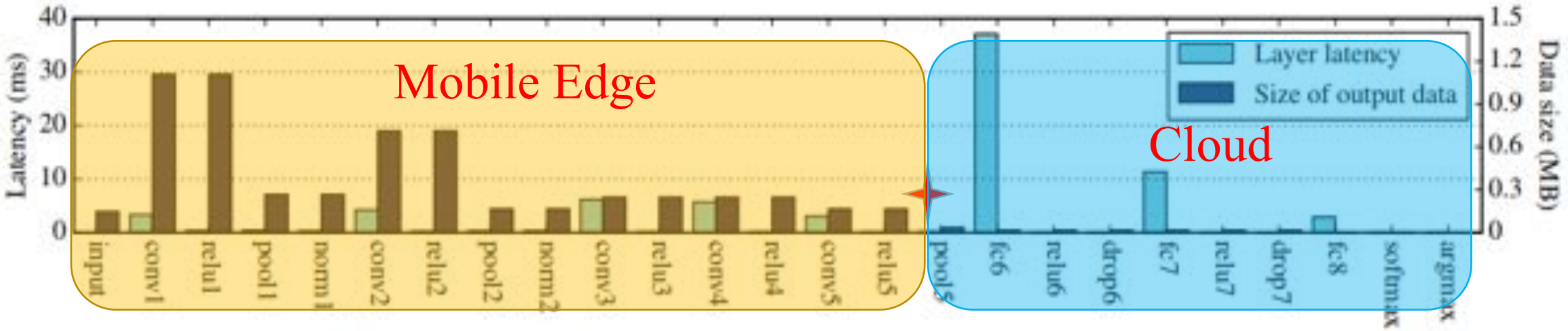
[3] Y. Kang et al., "Neurosurgeon: Collaborative intelligence between the cloud and mobile edge," ACM SIGPLAN Notices, vol. 52, no. 4,pp. 615–629, 2017.

Neurosurgeon（AlexNet as an example）



Jetson TK1 mobile platform

Key Observation:
(1)**Full-Connected Layers(FC)** incur high latency;
(2) **Convolution Layers(CONV)** increasing size of output data;



Neurosurgeon（Generalizing to more DNNs）

Table 4: Neurosurgeon's partition point selections for best end-to-end latency.

Mobile	Wireless network	Benchmarks							
		IMC	VGG	FACE	DIG	ASR	POS	NER	CHK
CPU	Wi-Fi	input	input	input	input	input	fc3		
	LTE	input	input	input	argmax	input	fc3		
	3G	argmax	input	input	argmax	input	fc3		
GPU	Wi-Fi	pool5	input	input	argmax	input	fc3		
	LTE	argmax	argmax	input	argmax	input	fc3		
	3G	argmax	argmax	argmax	argmax	input	fc3		

Table 5: Neurosurgeon partition point selections for best mobile energy consumption.

Mobile	Wireless network	Benchmarks							
		IMC	VGG	FACE	DIG	ASR	POS	NER	CHK
CPU	Wi-Fi	input	input	input	input	input	fc3		
	LTE	input	input	input	input	input	fc3		
	3G	input	input	input	argmax	input	fc3		
GPU	Wi-Fi	input	input	input	argmax	input	fc3		
	LTE	pool5	input	input	argmax	input	fc3		
	3G	argmax	argmax	input	argmax	input	fc3		

Neurosurgeon **adapts to various** DNN architectures, hardware platforms, wireless connections, and server load levels, and **chooses the partition point** for **best latency and best mobile energy consumption**.

代表性成果： EdgeAssisted（MobiCom'19）

"Edge Assisted Real-time Object Detection for Mobile Augmented Reality"

- Detect and classify complex objects in the real world
- AR/MR应用处理需要60FPS，每帧端到端处理时延 < 16.7ms
- Offload object detection to the edge or cloud server

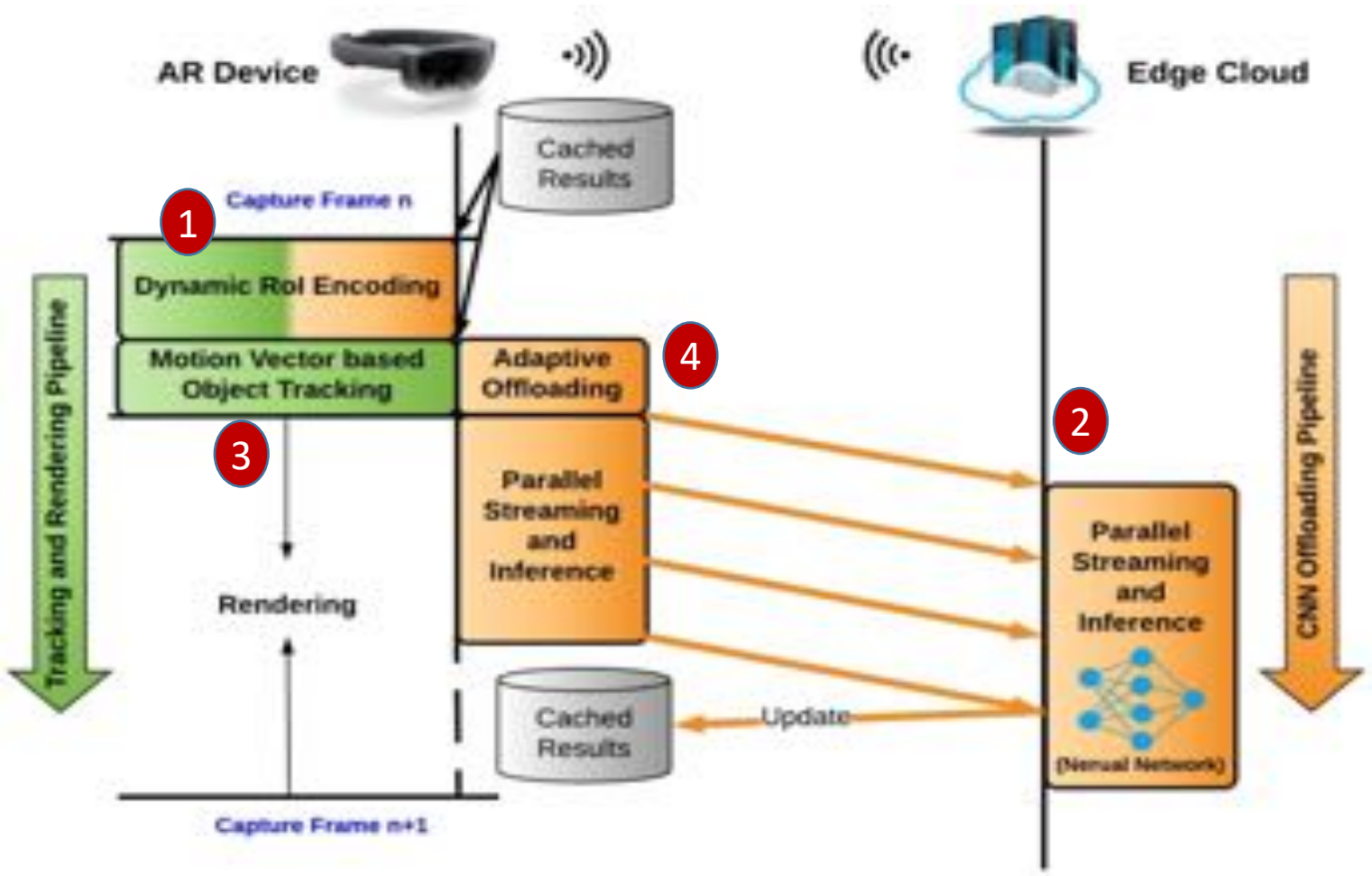
  Detection Accuracy + E2E Latency

(a) Dangerous traffic warning.

(b) Pickachu sits on her shoulder.

[5] Luyang Liu, Hongyu Li, Marco Gruteser, "Edge Assisted Real-time Object Detection for Mobile Augmented Reality", *ACM MobiCom*, 2019.

EdgeAssisted System Framework



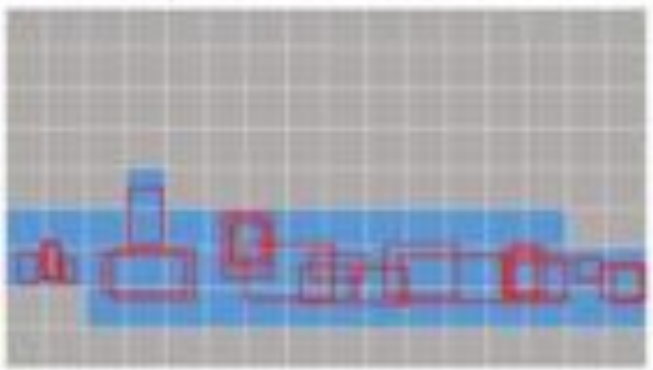
- A local tracking and rendering system on a mobile device and a pipelined objected detection system on the edge cloud.
- Decouple the rendering process and CNN offloadding process into two separate pipelines.

Dynamic Roi Encoding

ROI：Region of Interest

- Adjust the encoding quality on each frame to reduce the transmission latency based on the Regions of Interest (Rols)detected in the last offloaded frame
- It provides higher quality encodings in areas where objects are likely to be detected and uses stronger compression in other area.
- The key innovation lies in identifying the regions with potential objects of interest from candidate regions on prior frames.

(a) Detect Rols on the last offloaded frame.



(b) Mark macroblocks that overlap with Rols.

(c) Change encoding quality on the current frame.

Motion Vector based Fast Object Tracking

- Estimate the object detection result of the current frame
- Indicate the offset of pixels among frames to achieve a higher compression rate.

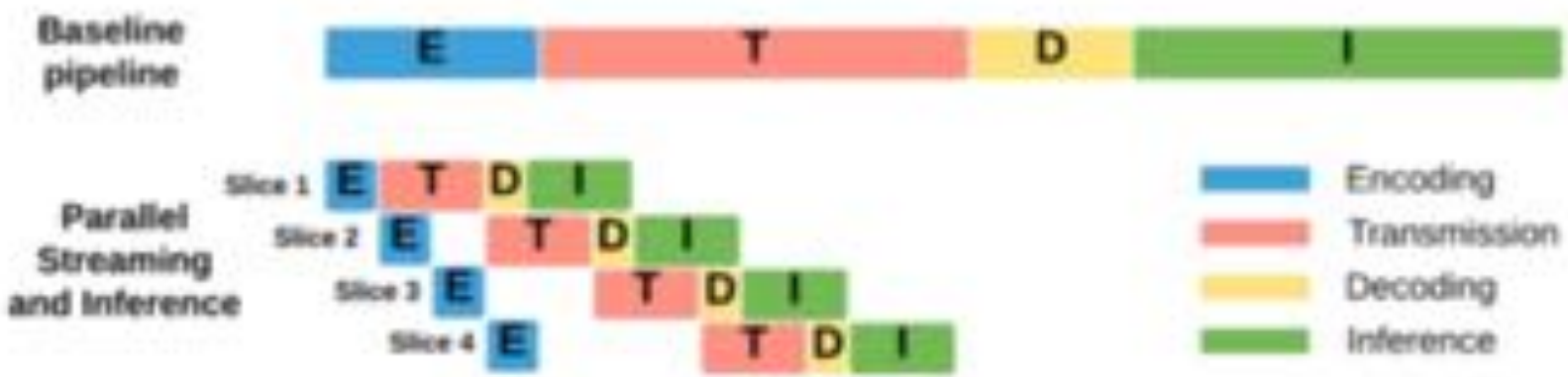
(a) Cached detection result of the last offloaded frame.

(b) Motion vectors extracted from the current encoded frame.

(c) Shift the bounding box based on the motion vectors.

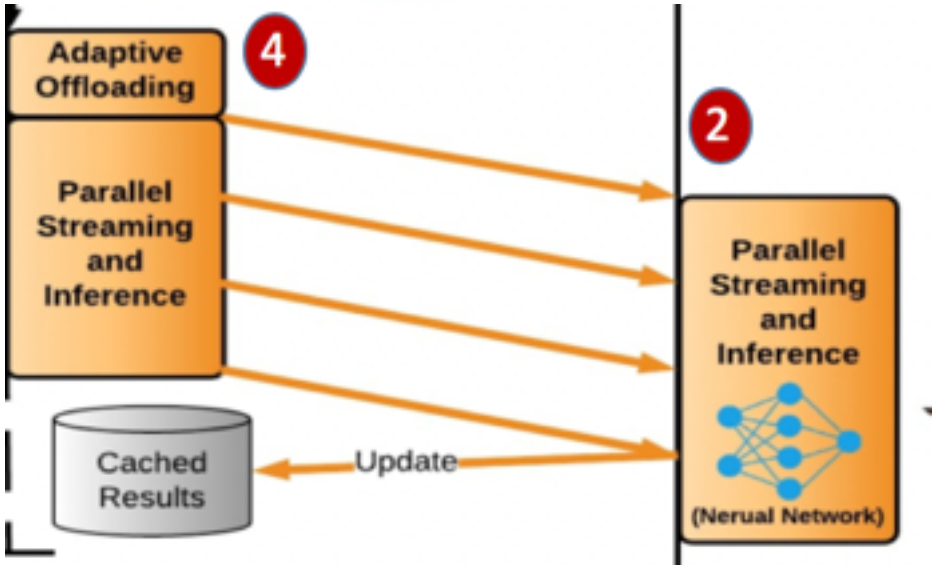
Parallel Streaming and Inference

- Enable inferences on slices of a frame
- the streaming and inference can be effectively pipelined and executed in parallel



Adaptive Offloading

- a frame will only be eligible to be offoaded if the previous offoaded frame has been completely received by the edge cloud
- a frame will be considered for offloading if it differs significantly from the last offloaded frame



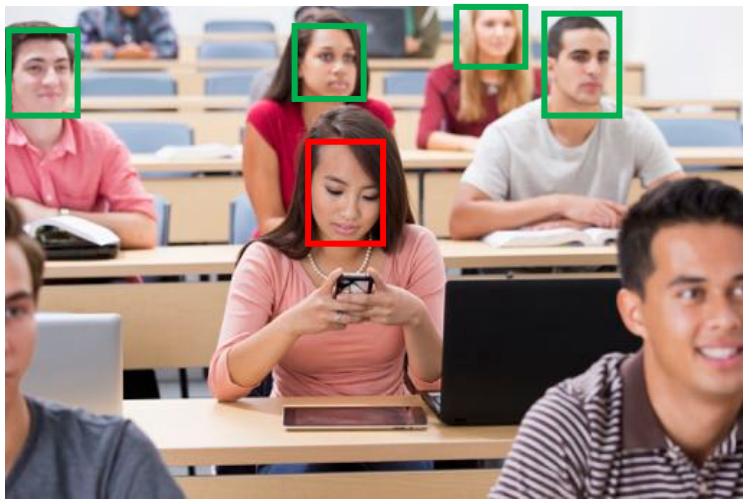
The result shows that the system can improve the detection accuracy by 20.2%-34.8% for the object detection and human keypoint detection tasks, and only requires 2.24ms latency for object tracking on the AR device.

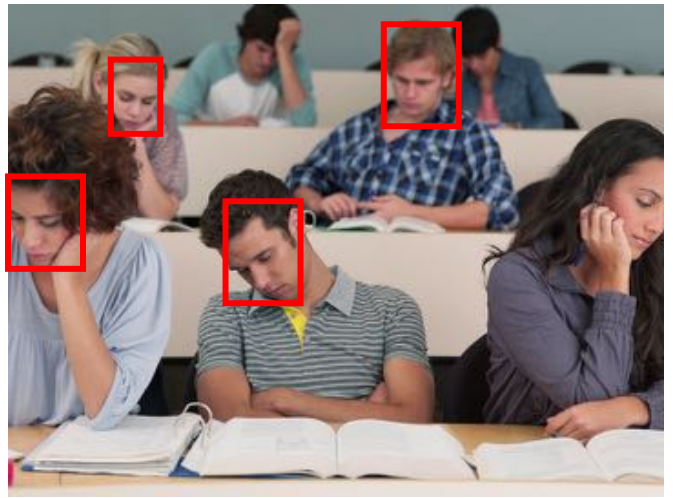
提纲

- 边缘智能简介
- 基于边缘智能的视频大数据分析
- 案例：基于视频分析的抬头检测系统

当人工智能遇上现代教学

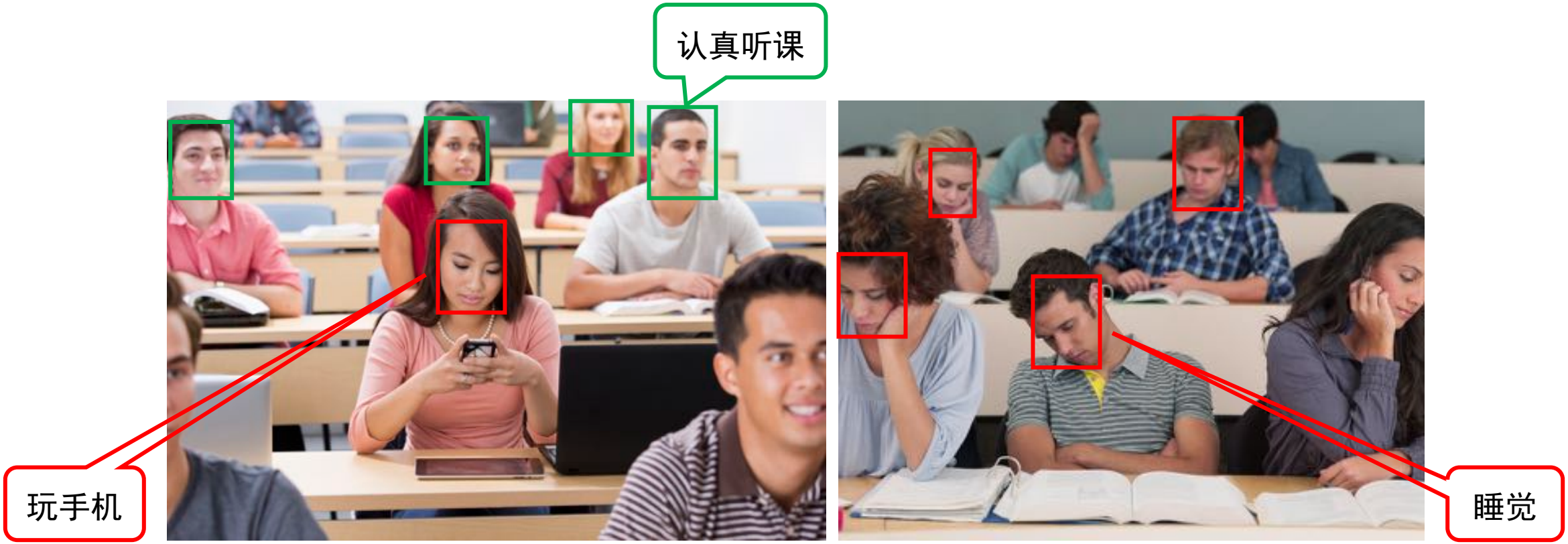
基于多目标视频分析的抬头检测系统





研究动机

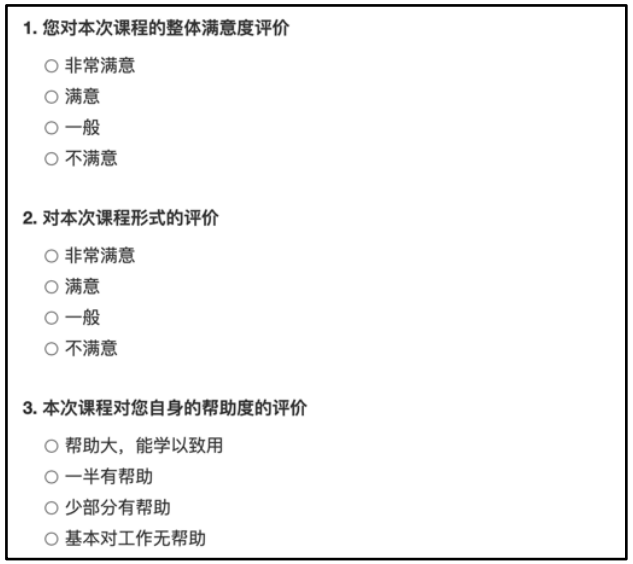
如何利用更为客观的指标来评估课堂教学质量？



研究动机

- 传统解决方案
  - 专家评估：现场评估，回看课堂视频
  - 学生评分：教务网打分
- 传统解决方案的不足之处：耗费时间，存在主观性

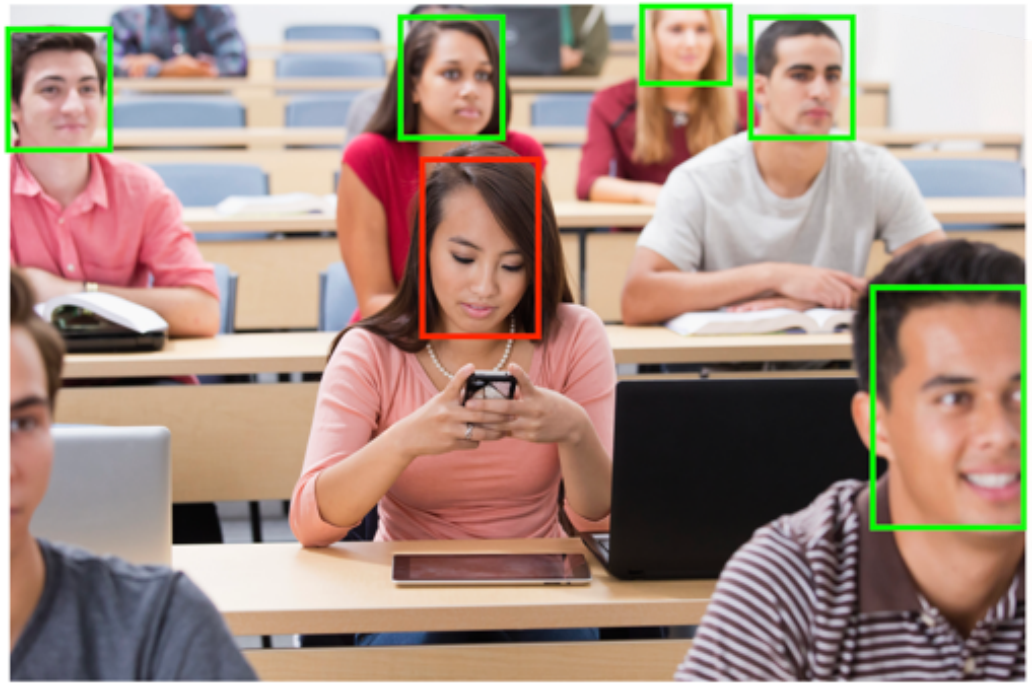
现场评分



问卷评分

研究动机

统计听课人数，评估课堂教学质量：$\frac{\text{听课人数}}{\text{总人数}}$

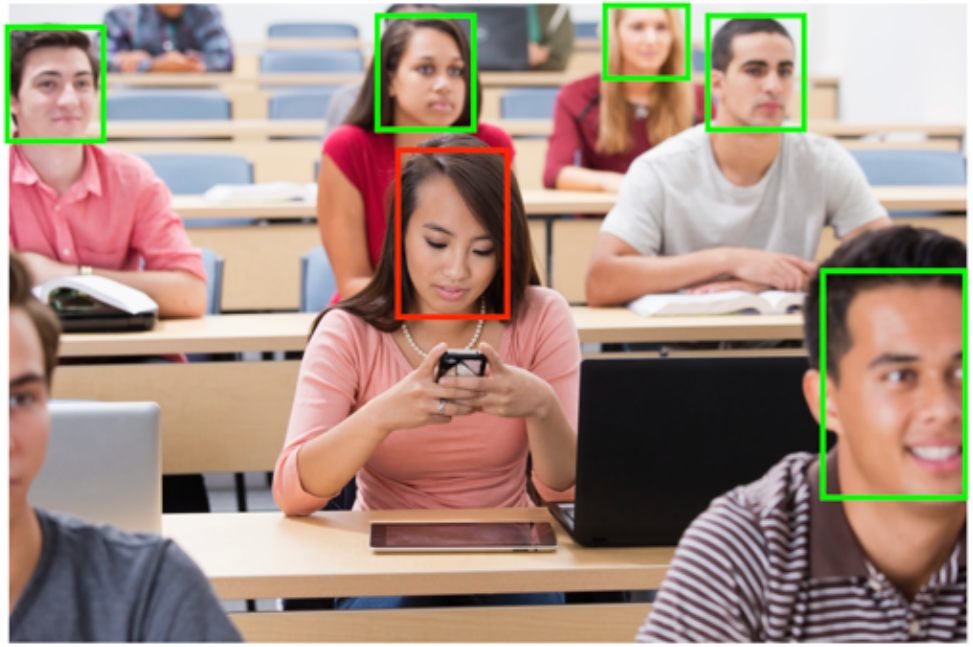


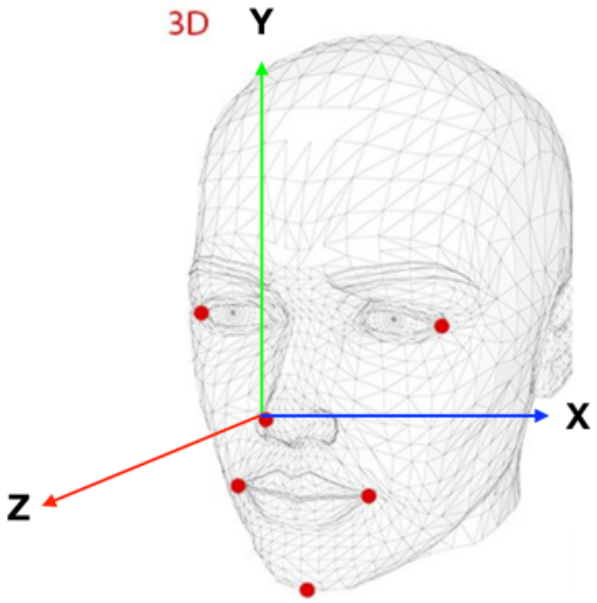
可否将AI技术用于教学？

可否处理图像/视频数据，检测**听课状态**与**非听课状态**？

研究思路

基于摄像头获取的数据，利用计算机视觉等方法检测课堂听课人数比例。其中，**是否抬头**作为是否听课的主要判断依据。

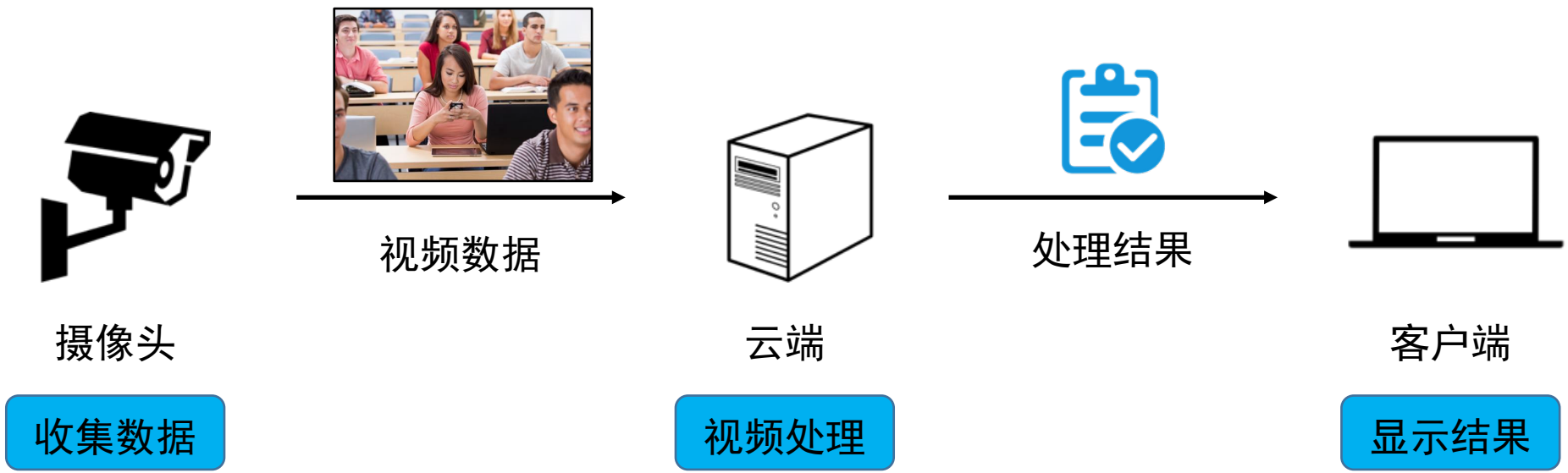




是否抬头可以根据**脸部姿态**判断，即图中脸部绕X轴旋转的角度，例如，该角度小于一定阈值，可以判定为抬头状态。

现有方案及存在的问题

- 目前，大部分视频处理任务需要交付云端集中式处理，结合实际应用场景，已有解决方案如下：



- 虽然该方案可以结合云计算的优势，但是仍然存在一些问题

现有方案及存在的问题

- 上述解决方案存在的问题：
  - 带宽问题：不同应用对视频质量的需求不同，视频质量需求越高，所需带宽越大
  - 时延问题：时延主要包括视频数据传输时延和云端计算时延，当网络状况差时，传输时延会显著增加
  - 隐私问题：在特定应用场景中，视频数据可能包含用户位置和肖像等个人隐私，从而不便上传到云端

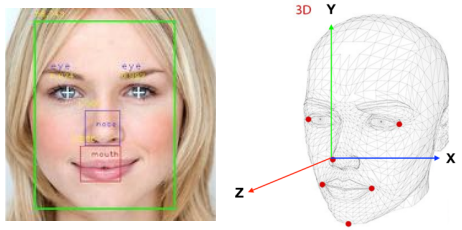
带宽

时延

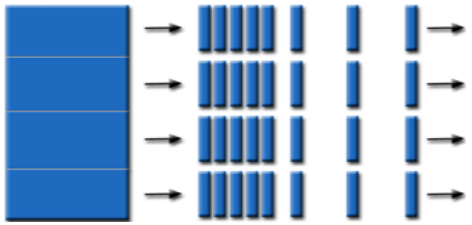
隐私

研究方法

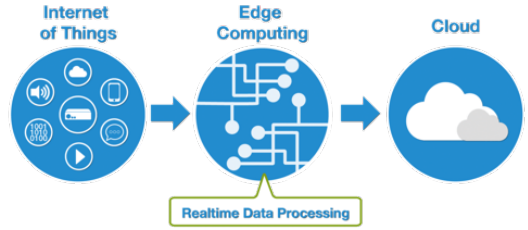
1. **人工智能和模式识别**：对视频数据进行处理，包括人脸检测、脸部特征点检测以及脸部姿态估计等。计算机视觉等人工智能方法可以提供有效帮助。



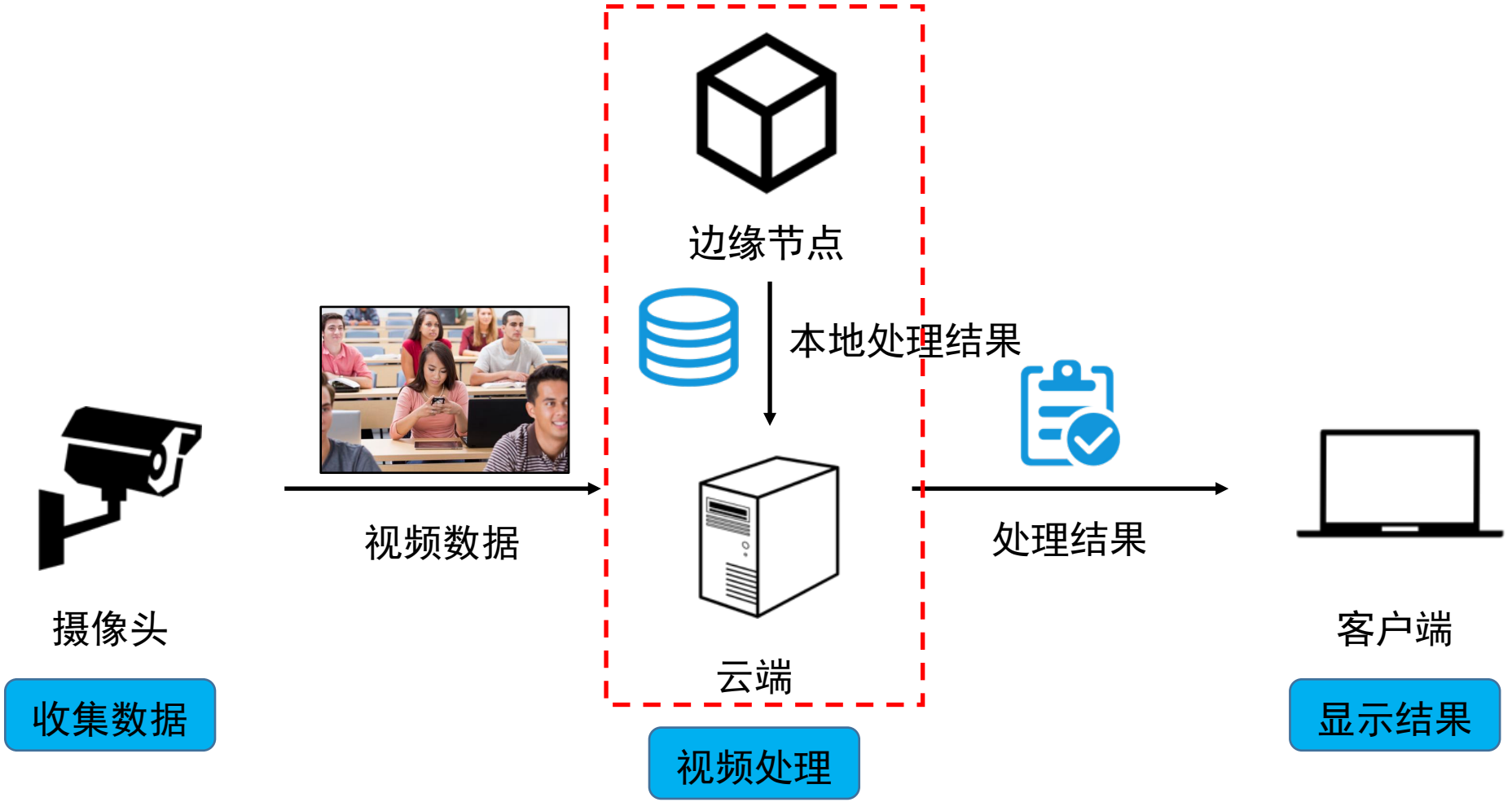
2. **并行处理**：计算机视觉等人工智能和模式识别方法，往往会引入较大的处理时延，采用并行处理相关技术，降低系统处理时延。



3. **边缘计算**：对于视频数据预处理等计算开销小的任务实行近端处理，交付给距离感知源较近的边缘节点完成，而对于计算开销较大的任务，进一步由边缘节点传输给云端完成。



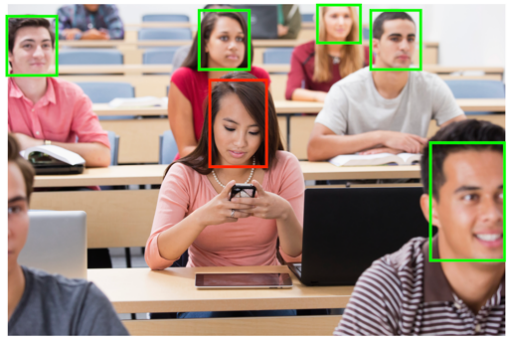
基于“边缘智能”的解决方案

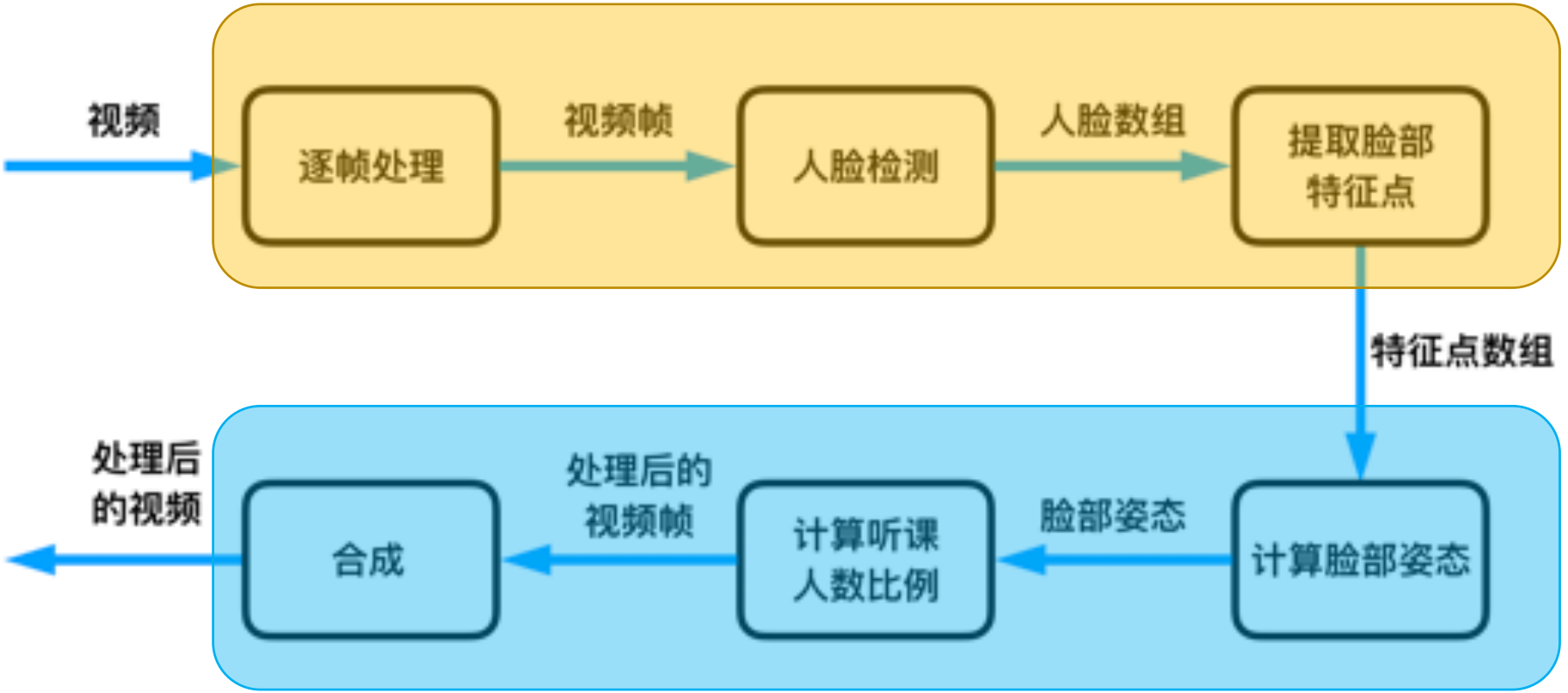


系统处理流程

边缘端处理

对于视频数据预处理等计算开销小的任务实行近端处理，交付给距离感知源较近的边缘节点完成

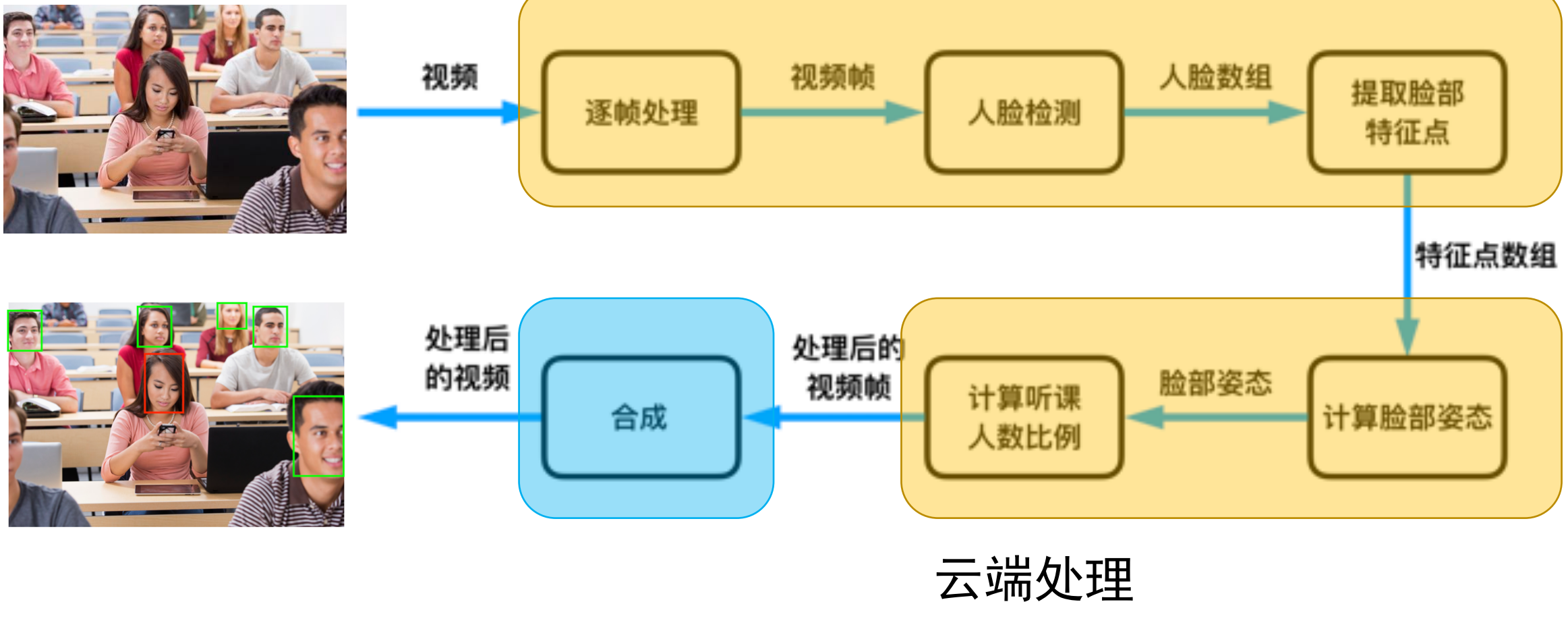




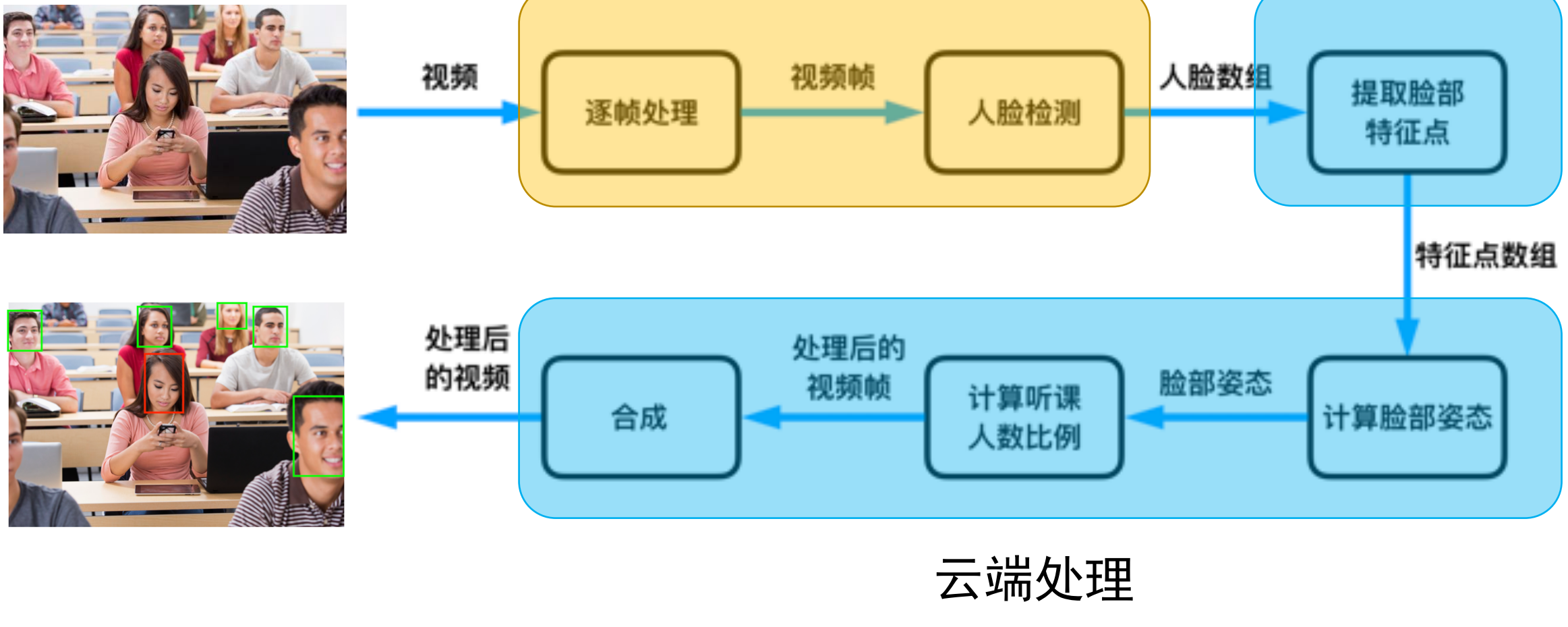
云端处理

对于计算开销较大的任务，进一步由边缘节点传输给云端完成

“端-边-云”协同的弹性处理框架



“端-边-云”协同的弹性处理框架



系统部署

系统实现


Total : 12 Detected : 9
up -2.23
up 1.19
up 1.01
up 3.41
up 2.64
up -11.57
other -35.31

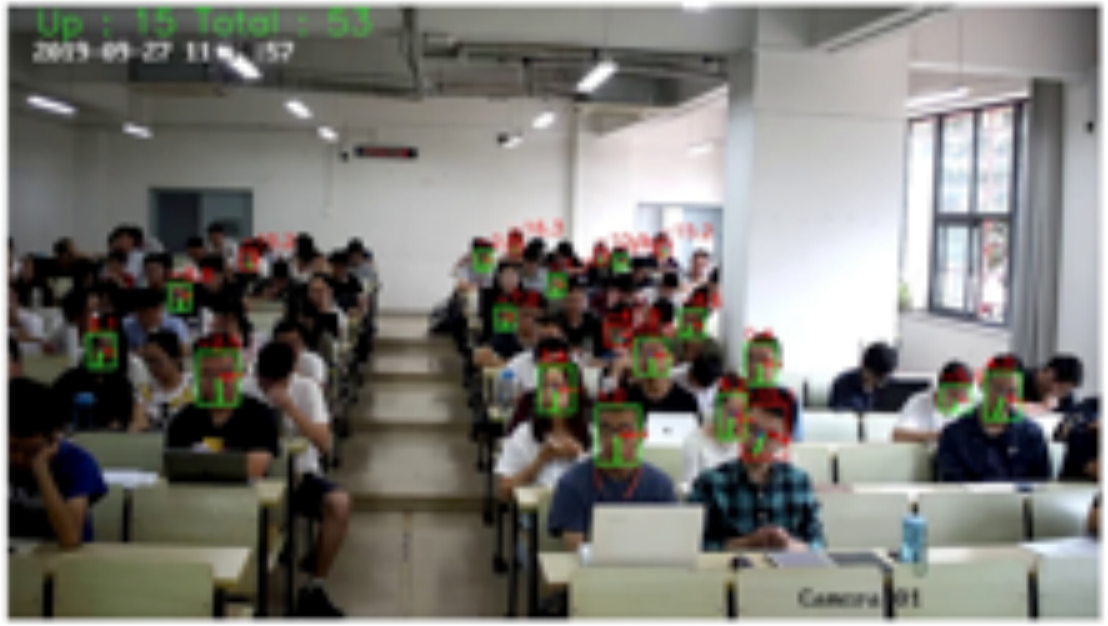
系统实现


视频管理
课堂情况
视频实时显示区域
角度:off
听课人数比例
听课:6 其他:6
是否显示抬头角度等辅助信息
听课人数比例

系统实现

课堂情况

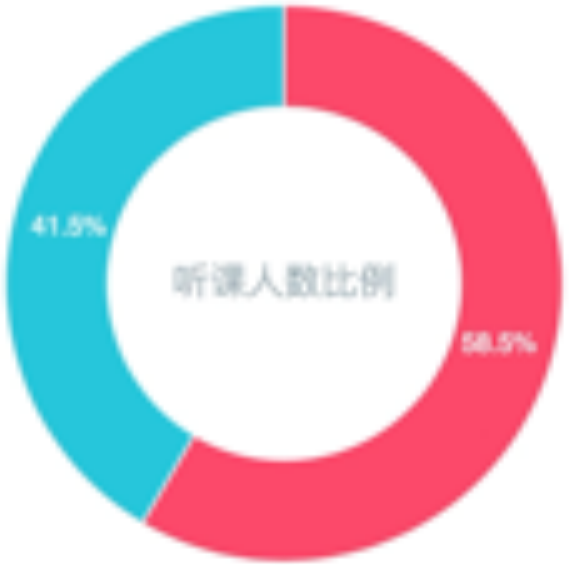
视频画面



角度 : on

听课人数比例

课堂听课情况统计



听课 : 22　其他 : 31

“边缘智能”系统原型

边缘节点

广角
摄像头

GPU
处理器

边缘智能-

边缘计算时代的人工智能

报告人：谢 磊

敬请各位专家批评指正！

References

[1] JIASI CHEN AND XUKAN RAN, Deep Learning With Edge Computing: A Review, Proceedings of the IEEE, May 2019.

[2] Zhi Zhou, Xu Chen, En Li, Liekang Zeng, Ke Luo, and Junshan Zhang,“Edge Intelligence: Paving the Last Mile of Artificial Intelligence with Edge Computing,” Proceedings of the IEEE, May 2019.

[3] Y. Kang et al., “Neurosurgeon: Collaborative intelligence between the cloud and mobile edge,” ACM SIGPLAN Notices, vol. 52, no. 4,pp. 615–629, 2017.

[4] Ben Zhang, Xin Jin, Sylvia Ratnasamy et al., “AWStream: Adaptive Wide-Area Streaming Analytics,” ACM SIGCOMM, 2018.

[5] Luyang Liu, Hongyu Li, Marco Gruteser, “Edge Assisted Real-time Object Detection for Mobile Augmented Reality”, ACM MobiCom, 2019.

[6] H. Zhang, G. Ananthanarayanan, P. Bodik, M. Philipose, P. Bahl, and M. J. Freedman, “Live video analytics at scale with approximation and delay-tolerance,” in Proc. USENIX NSDI, 2017, pp. 377–392.

[7] C.-C. Hung et al., “VideoEdge: Processing camera streams using hierarchical clusters,” in Proc. IEEE/ACM Symp. Edge Comput. (SEC), Oct. 2018, pp. 115–131