Task-oriented Communication for Edge AI

Jun Zhang

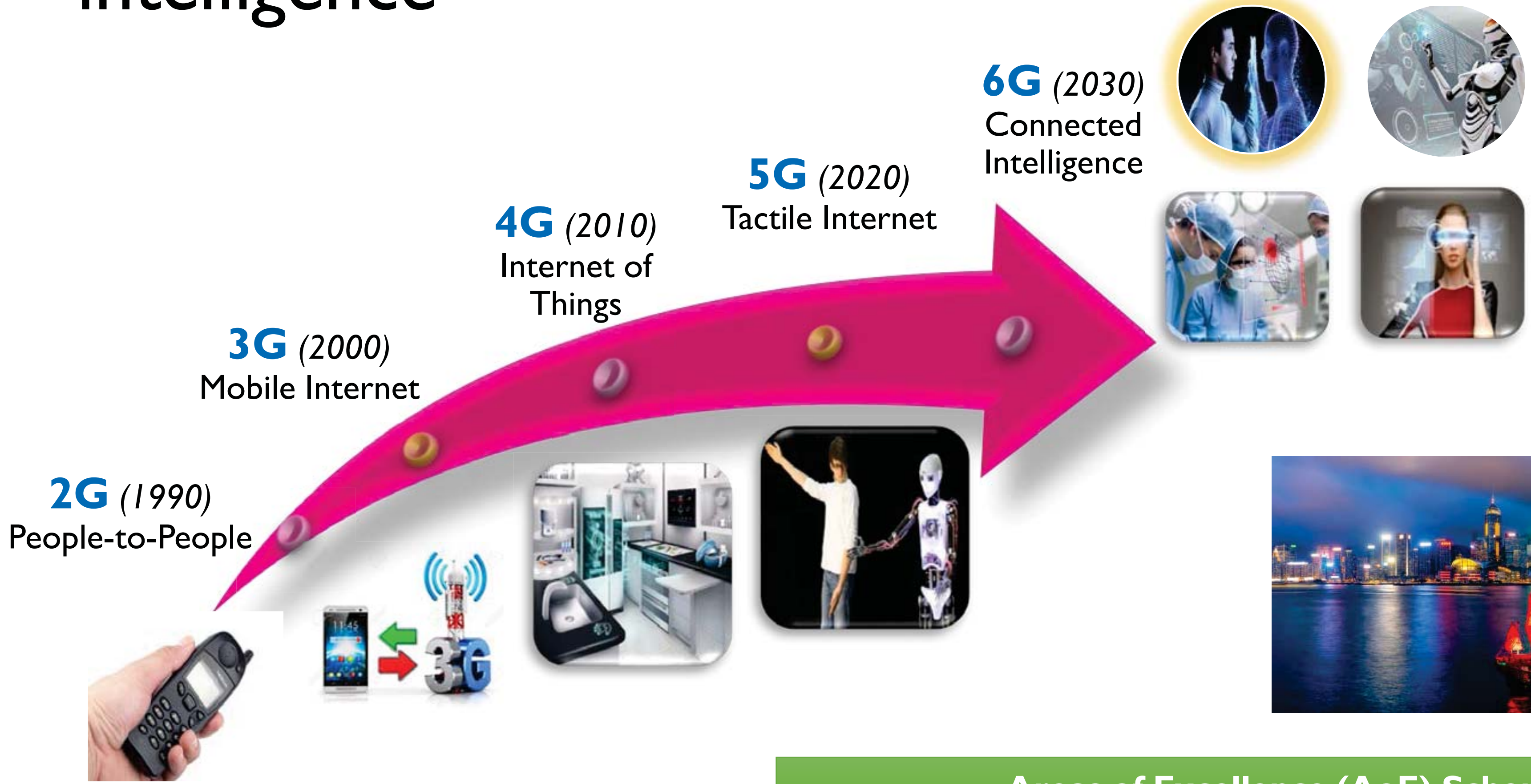
香港科技大學
THE HONG KONG
UNIVERSITY OF SCIENCE
AND TECHNOLOGY

Outline

- Introduction
- Design principles
  - Single-device inference via information bottleneck (IB)
  - Cooperative inference via distributed information bottleneck (DIB)
- Case studies
  - Edge video analytics
  - Edge-assisted localization for autonomous driving
  - EdgeGPT for autonomous edge AI
- Conclusions

Introduction

Wireless evolution: From "connected things" to "connected intelligence"



IEEE Hong Kong 6G Wireless Summit

Wireless@HKUST

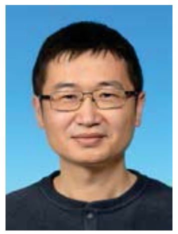
Areas of Excellence (AoE) Scheme

"6G: Wireless Access and Connectivity for an Intelligent and Sustainable World,"
HK$ **87,304,000**, 2023 – 2027.

Khaled Letaief
(IEEE Fellow)

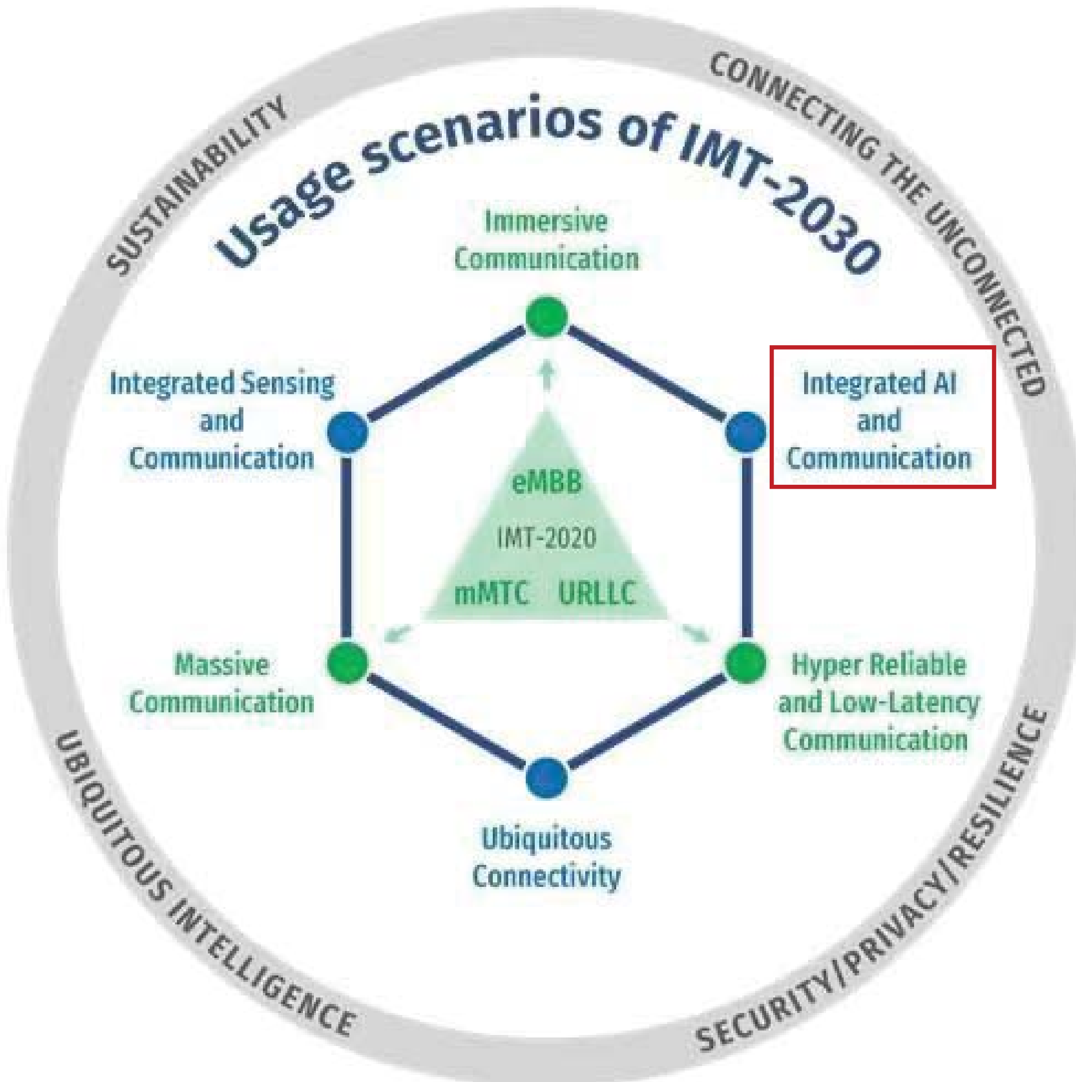
Ross Murch
(IEEE Fellow)

Vincent Lau
(IEEE Fellow)



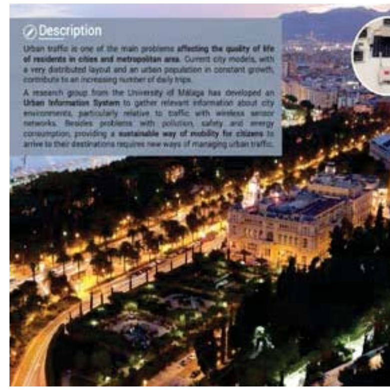
Jun Zhang
(IEEE Fellow)

Usage scenarios of IMT-2030 (6G visions)

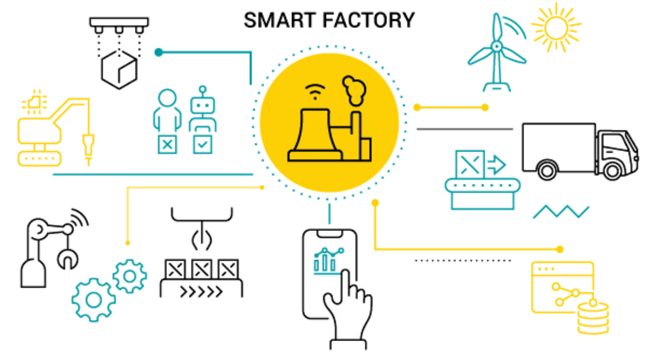


The rise of AI at the network edge

Smart home



Smart city



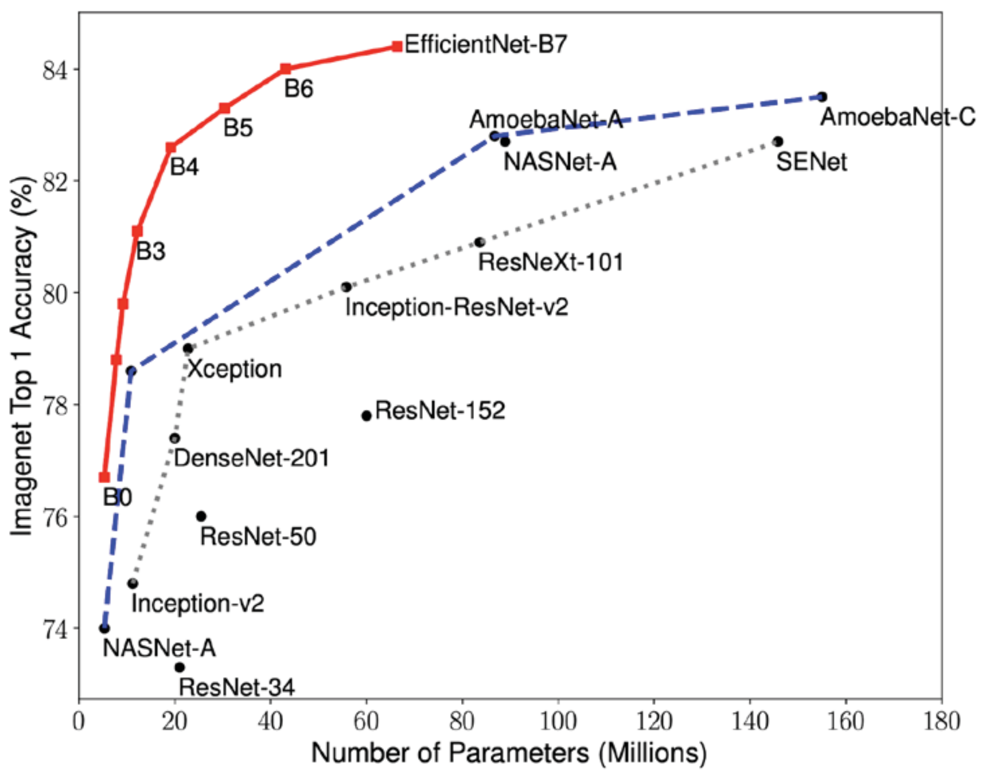
Smart factory

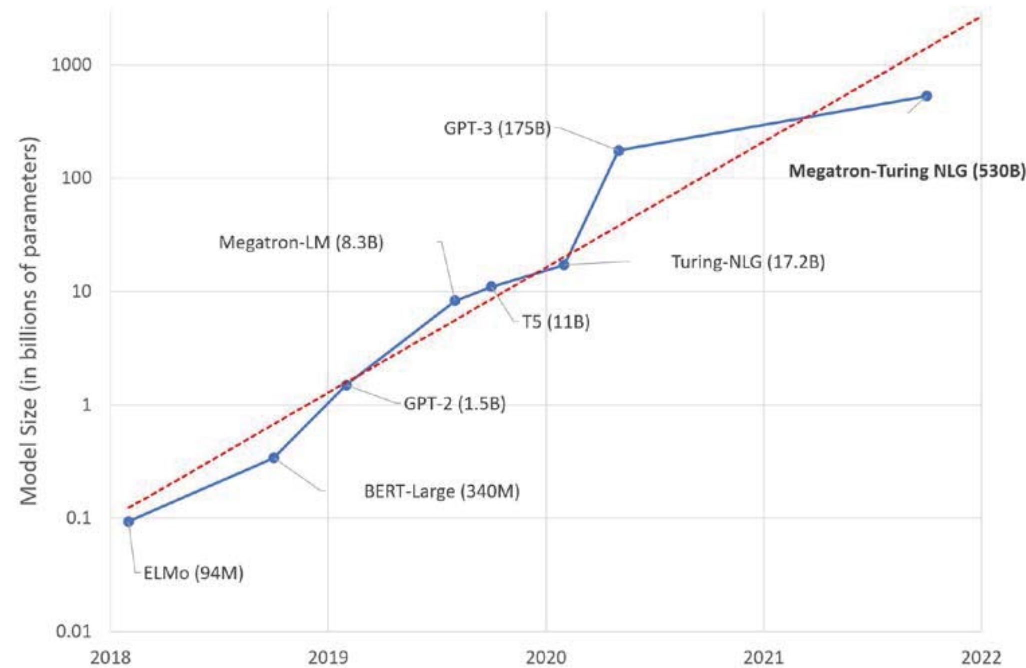
Autonomous vehicles

Drones

AR/VR

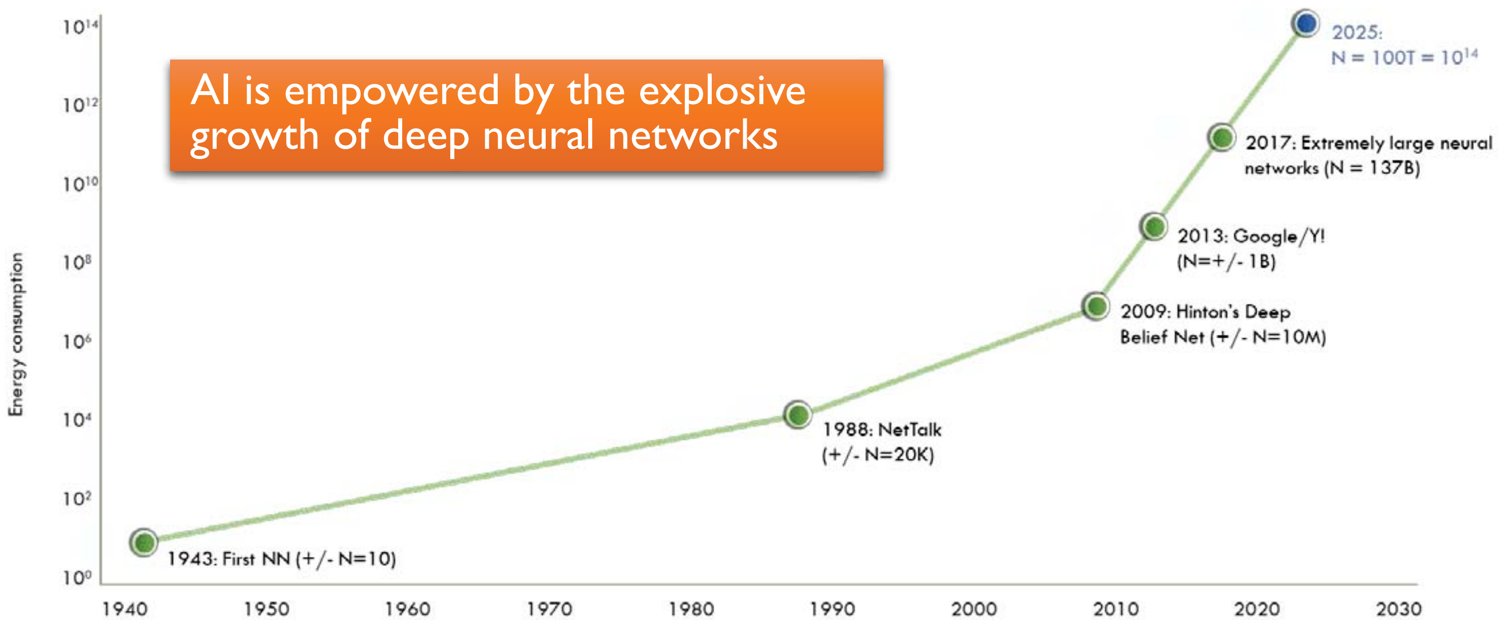
Challenge of edge inference:
Enormous model sizes vs limited onboard computing





https://ai.googleblog.com/2019/05/efficientnet-improving-accuracy-and.html, May 2019

Challenge of edge inference:
Huge energy consumption vs. limited onboard energy



[Max Welling, "Intelligence by the kilowatthour," ICML 2018, Invited Talk.]

~27 minutes

Mobile AI **drains battery rapidly**

~1600 mAh

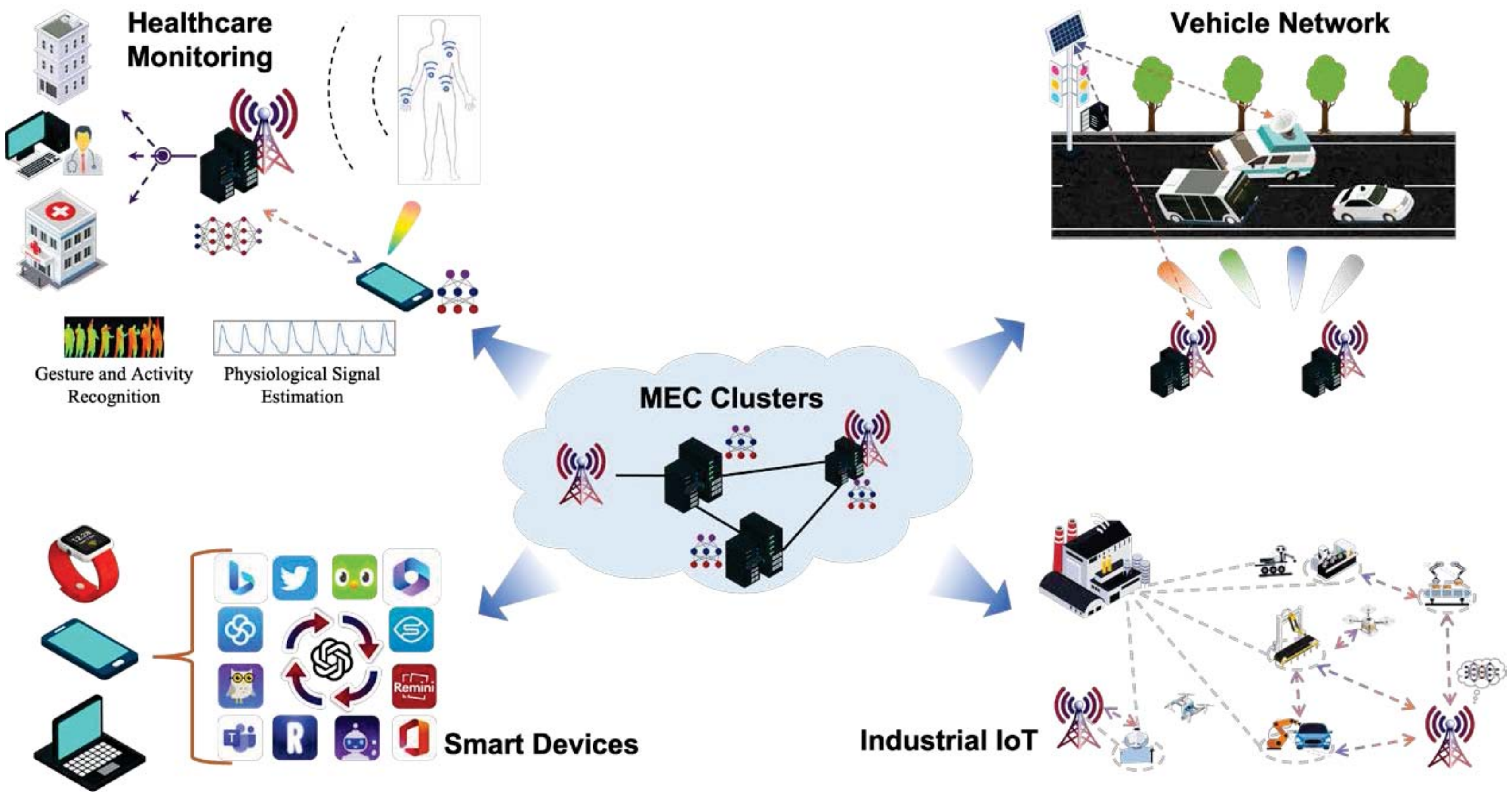
4100 mAh

7250 mAh

41.7 Wh

~90 minutes

Solution: Edge AI



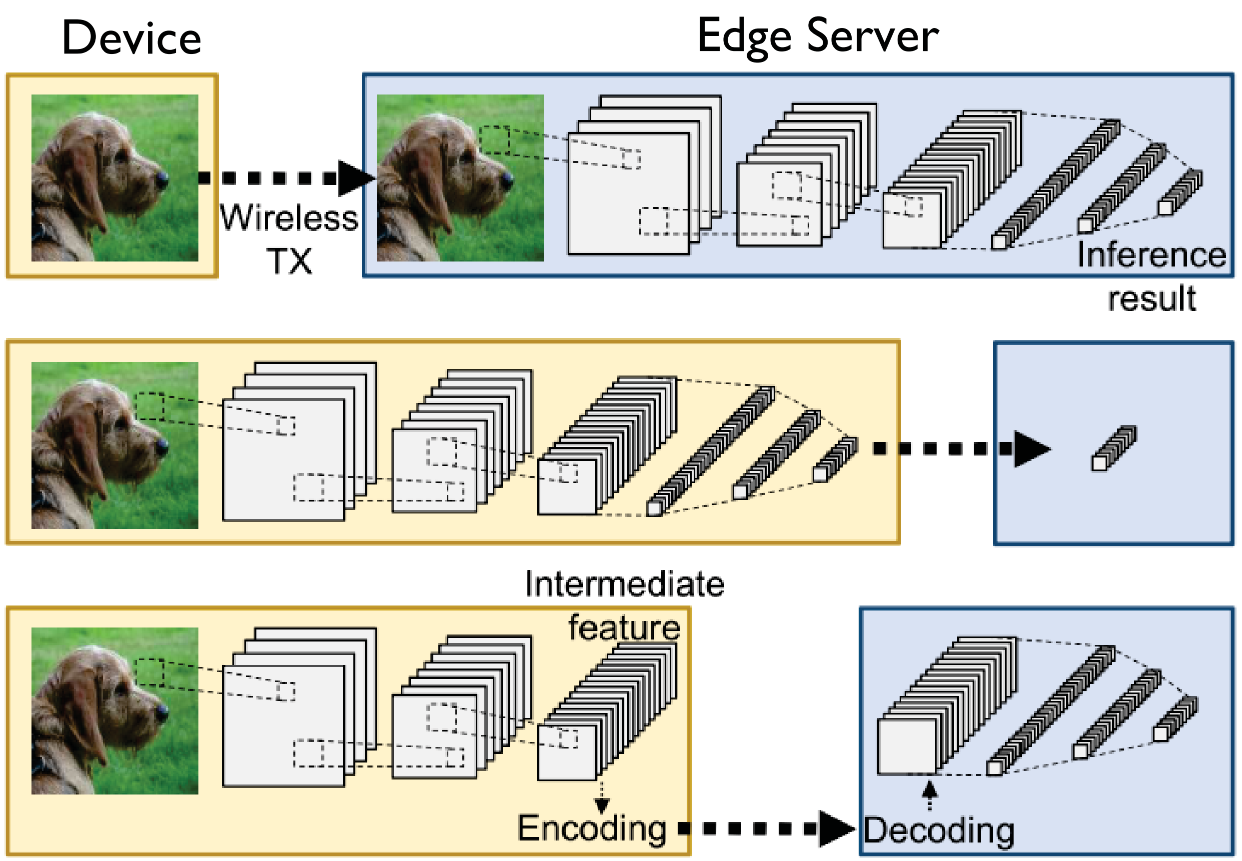
A single device is limited in

- onboard computing resources;
- limited perception capability;
- limited energy supply.

Effective communication to

- access external computing power;
- improve perception capability;
- prolong battery time;
- overcome partial observation.

Solutions for edge inference



Server-based method

- ☒ High communication load
- ☒ Privacy concern

On-device processing

- ☒ High local computation
- ☒ Limited performance

Device-edge co-inference ✓

Balance communication and local computation

New communication problem

- Communication for edge inference (not for data reconstruction)

Rethink communication problems

Shannon's information theory

Level A
The technical problem

- How *accurately* can the symbols of communication be transmitted?

How to communicate?

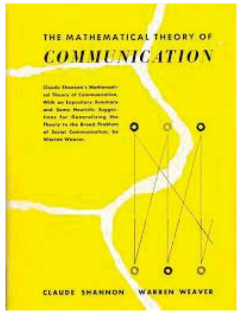
Level B
The semantic problem

- How *precisely* do the transmitted symbols convey the desired meaning?

What to communicate?

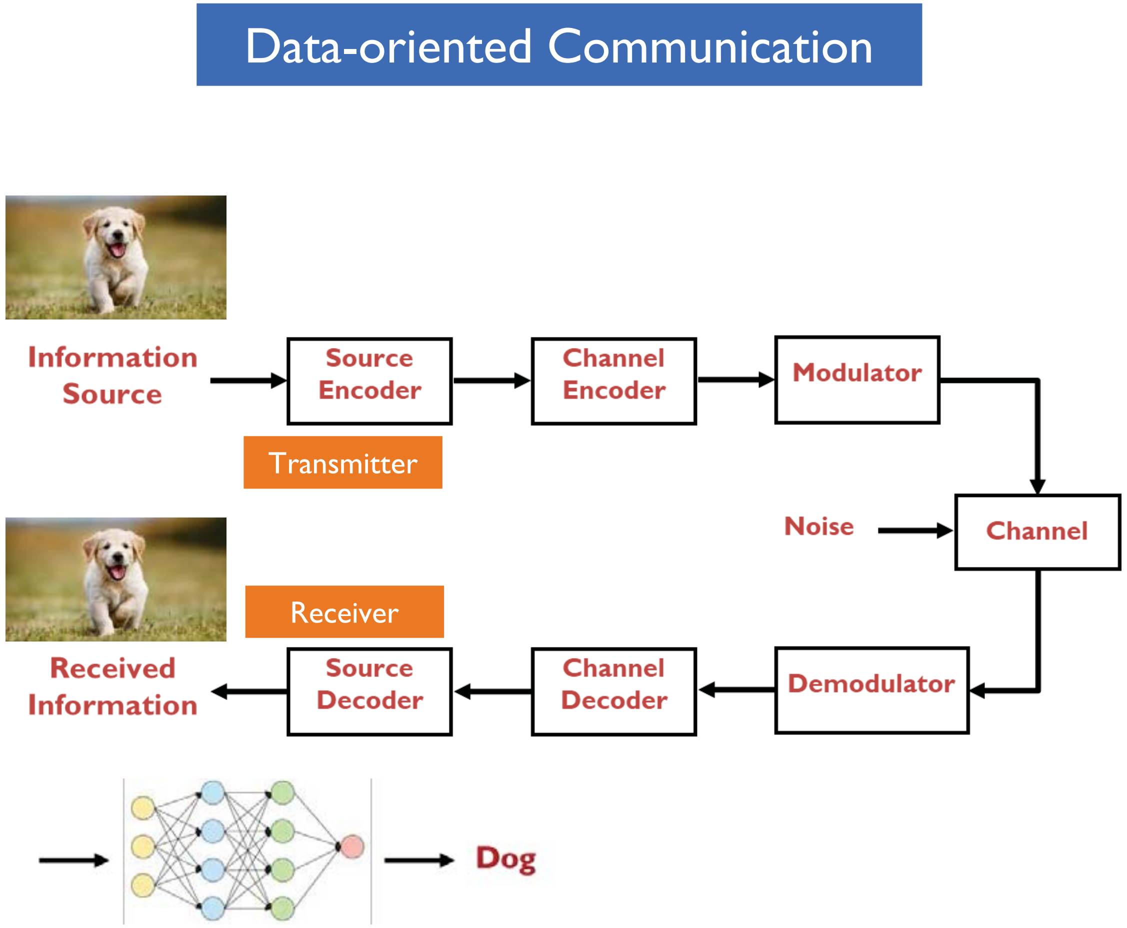
Level C
The effectiveness problem

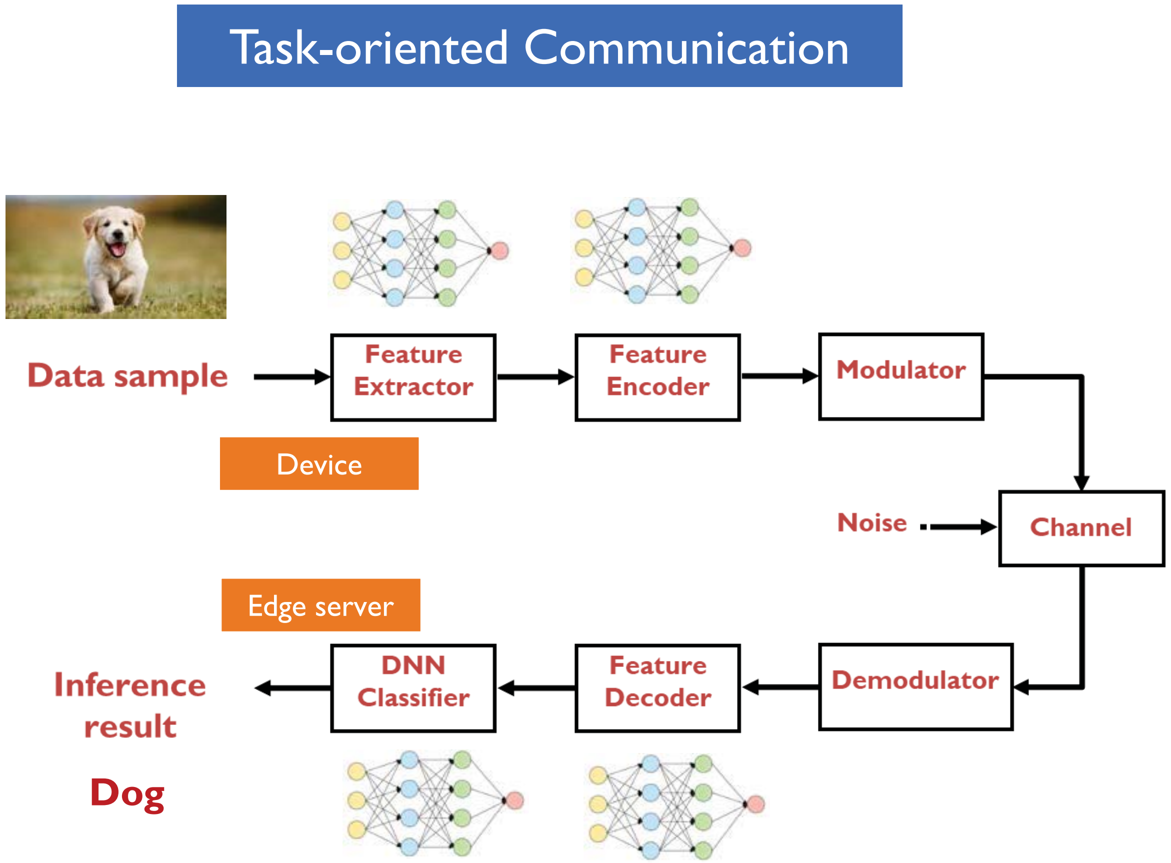
- How *effectively* does the received meaning affect conduct in the desired way?



W. Weaver. Recent contributions to the mathematical theory of communication. In C. E. Shannon and W. Weaver, editors, *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, 1949.

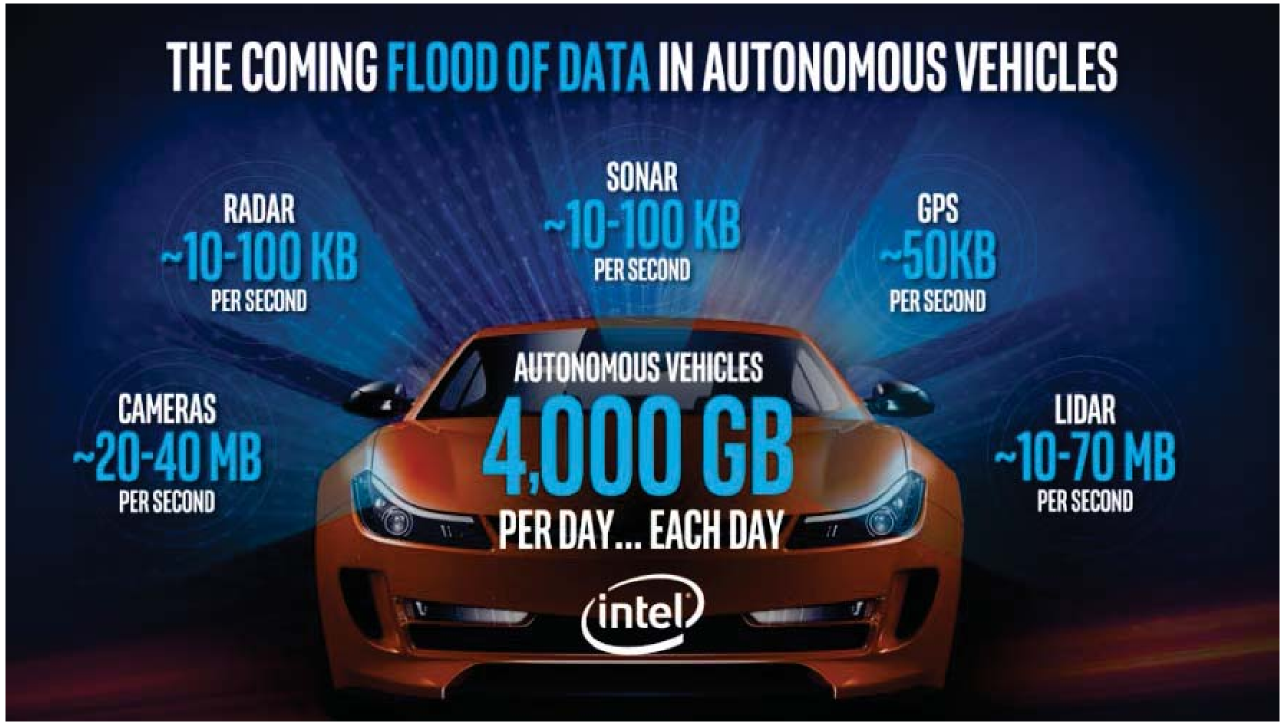
Data-oriented vs. Task-oriented communication





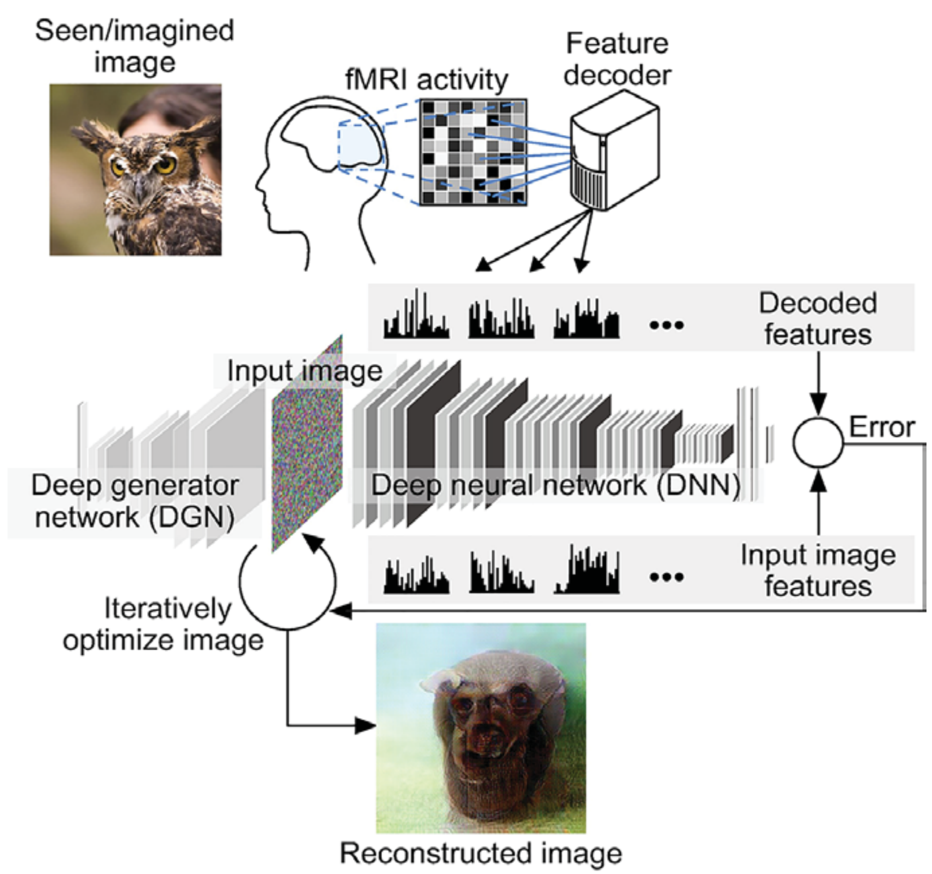
Why do we need task-oriented communication?

- Enormous data volume in emerging applications
  - E.g., robots/self-driving cars with various sensors.



Why do we need task-oriented communication?

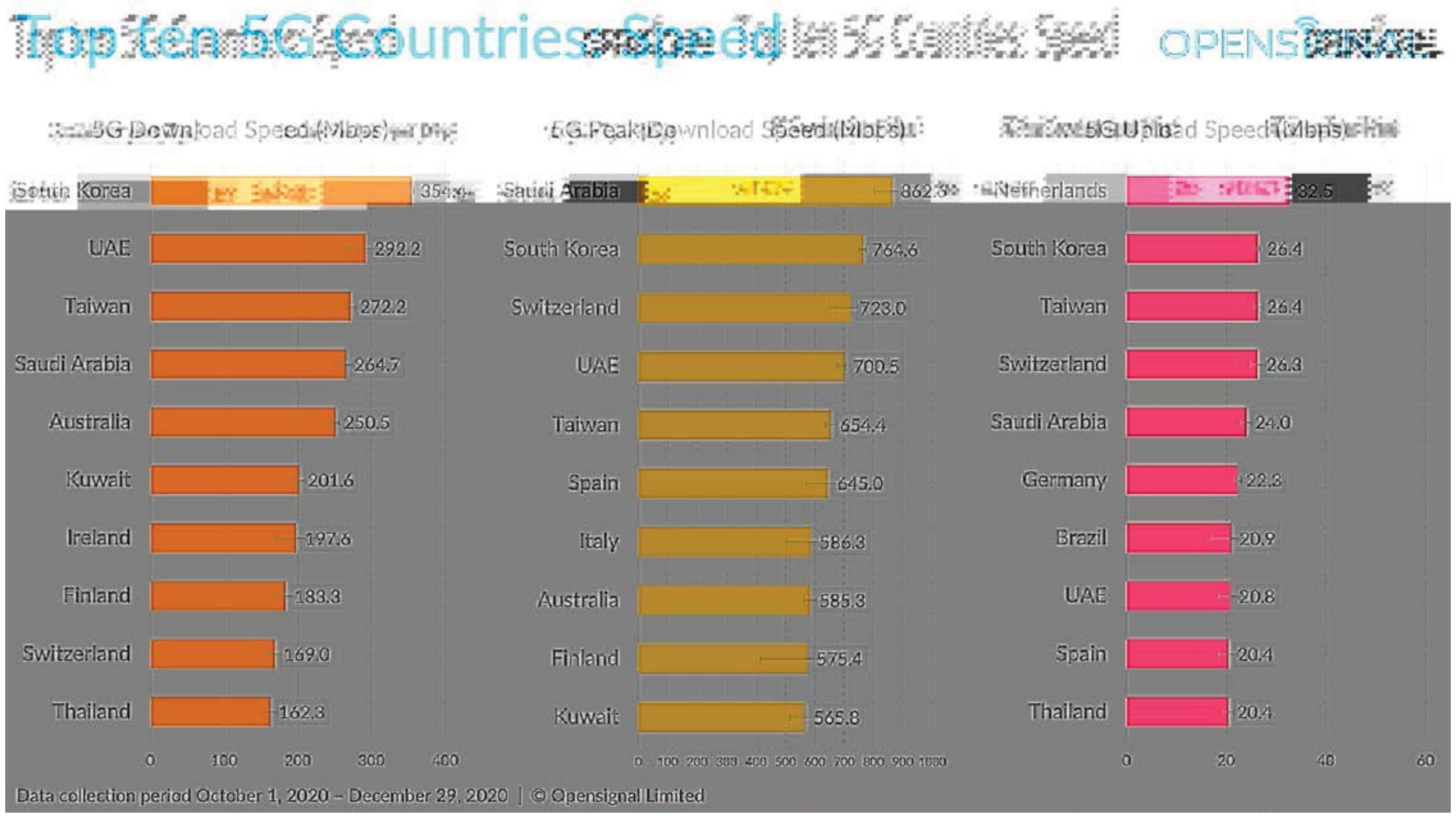
- We do not need to transmit or store everything
  - E.g., sensing data may only ever be "seen" by algorithms and machines that process them.
  - For humans, we do not store high-definition images in our brain:



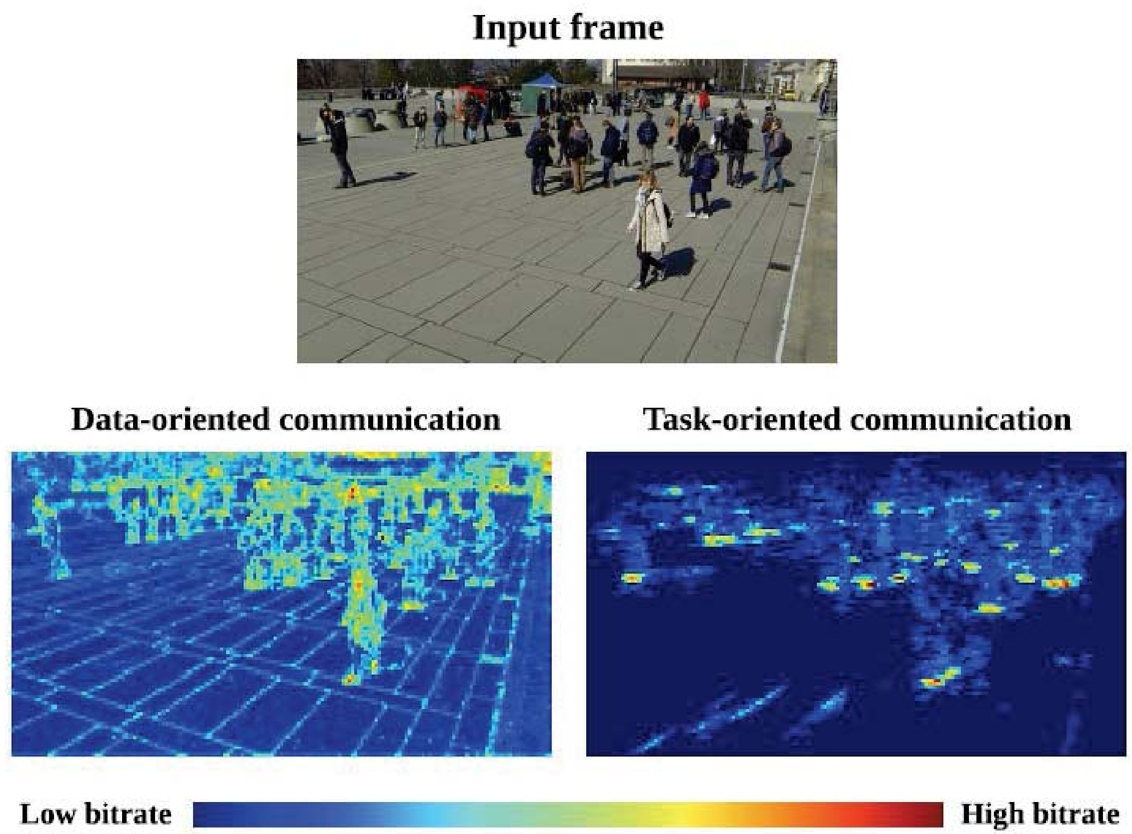
Shen G, Horikawa T, Majima K, Kamitani Y (2019) Deep image reconstruction from human brain activity. PLoS Comput Biol 15(1): e1006633.

Why do we need task-oriented communication?

- Asymmetry in uplink/downlink capacity and traffic
  - Uplink traffic becomes dominant, but uplink capacity lags behind



Example: Multi-camera pedestrian occupancy prediction



Data-oriented communication:
- It allocates many bits to represent the background and ground texture.
- However, these details almost do not influence the performance of the downstream task.

Task-oriented communication:
- It focuses on task-relevant information (e.g., the foot points of pedestrians) and discards the redundancy.
- It substantially reduces the communication overhead and latency.

Task-oriented communication system design

- Design goal: To transmit *concise* and *informative* feature with *low-complexity* encoder for *low-latency high-accuracy* inference

Design challenges
- Unknown high-dimensional data distribution
- Intractable task-specific distortion metric
- High computational complexity

Design tools
- End-to-end deep learning
- Variational approximation (to make the objective tractable)
- Neural network architecture optimization

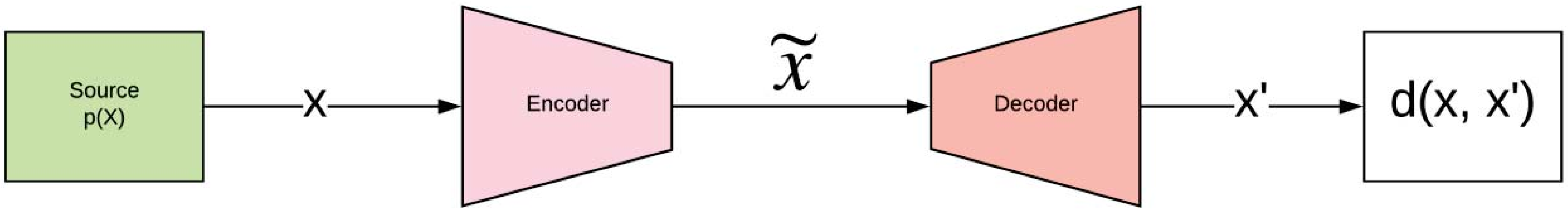
Task-oriented communication via information bottleneck

J. Shao, Y. Mao, and **J. Zhang**, "Learning task-oriented communication for edge inference: An information bottleneck approach," *IEEE J. Select. Areas Commun.*, vol. 40, no. 1, pp. 197-211, Jan. 2022.

J. Shao, Y. Mao, and **J. Zhang**, "Task-oriented communication for multi-device cooperative edge inference," *IEEE Transactions on Wireless Communications*, vol. 11, no. 1, pp. 73-87, Jan. 2023.

Rate-Distortion(R-D) theory

- Rate-Distortion (R-D) theory

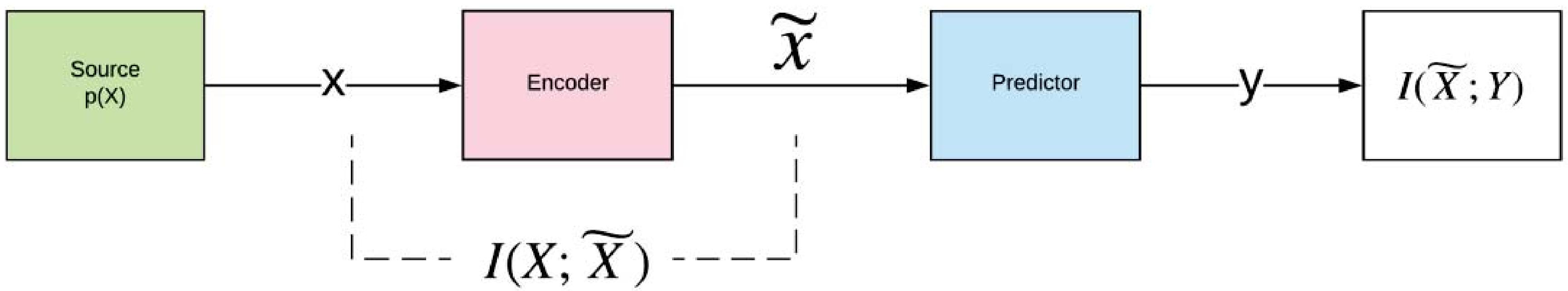


- Problem formulation

$$\min_{p(\widetilde{x}|x)} \underbrace{I(X;\widetilde{X})}_{\text{Rate}} + \beta \underbrace{E[d(X,\widetilde{X})]}_{\text{Distortion}}$$

Information bottleneck

- Information bottleneck extends R-D theory to prediction
  - Measuring the quality of the encoding by its ability to predict another random variable



- Problem formulation
  - The information bottleneck bound characterizes the optimal representations.

$$\min_{p(\widetilde{x}|x)} \underbrace{I(X;\tilde{X})}_{\text{Compression}} - \beta \underbrace{I(\tilde{X};Y)}_{\text{Prediction}}$$

To promote generalization | To promote accuracy

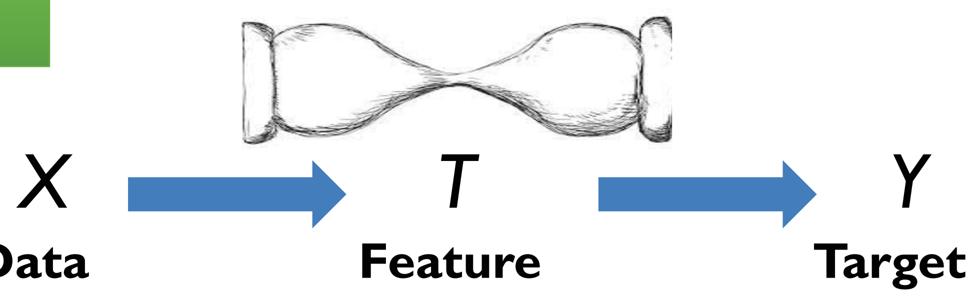
N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," Annu. Allerton Conf. Commun. Control Comput., 1999.

Task-oriented communication vs. Information bottleneck

Task-oriented Commun.

Dog

Information Bottleneck (IB)



Task-oriented communication via the IB principle

$$\underbrace{Y}_{\text{Target}} \leftrightarrow \underbrace{X}_{\text{Input}} \xrightarrow{\text{Encoder}} \underbrace{Z}_{\text{Feature}} \xrightarrow{\text{Channel}} \underbrace{\hat{Z}}_{\text{Rec. Feature}} \xrightarrow{\text{Decoder}} \underbrace{\hat{Y}}_{\text{Ouput}}$$

$$\min \underbrace{-I(\hat{Z}, Y)}_{\text{Distortion}} + \beta \cdot \underbrace{I(\hat{Z}, X)}_{\text{Rate}}$$

- We do not need to recover X from $\hat{Z}$
- $\hat{Z}$ only needs to retain task-relevant information to infer Y

How well $\hat{Z}$ predicts Y

To promote accuracy

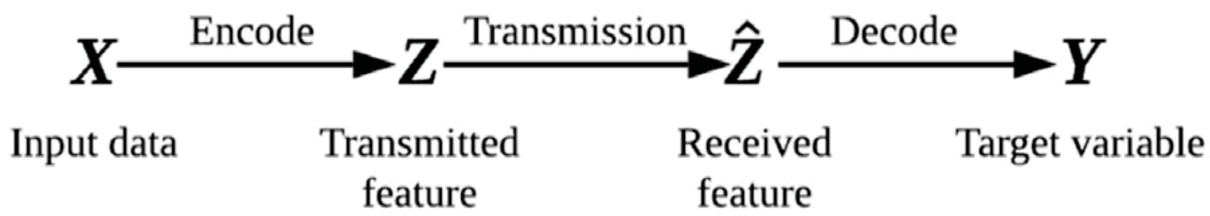
How much $\hat{Z}$ compresses X

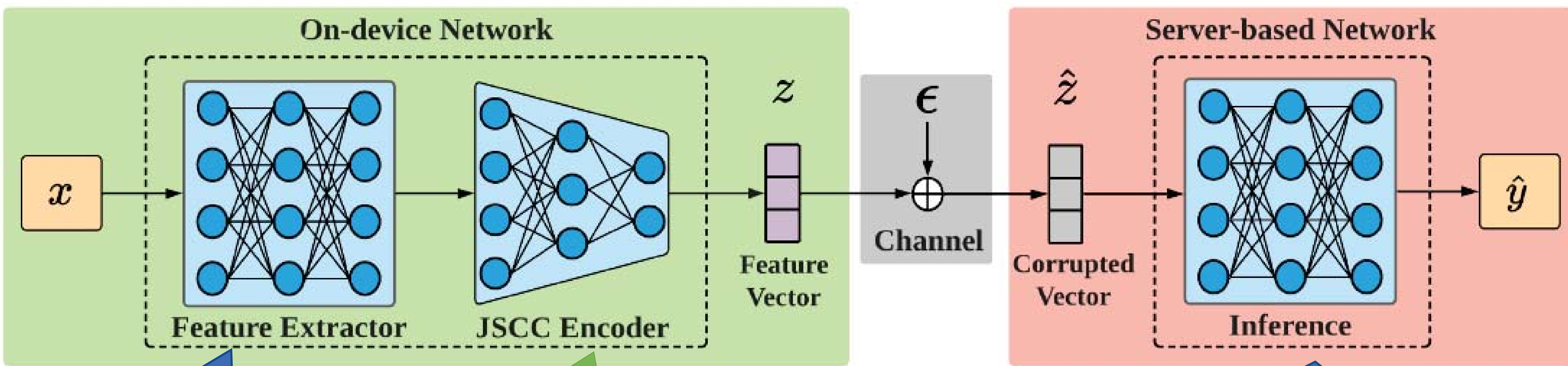
To reduce communication overhead

Relevance-rate tradeoff

- Main design challenges:
  - How to estimate mutual information?
  - How to effectively control communication overhead?
  - How to handle dynamic channel conditions?

Variational Feature Encoding (VFE)





Lightweight feature extractor: to control on-device computation/energy

Joint-source-channel coding (JSCC) encoder: design component, to minimize the output dimension

Powerful server-side network

VFE: Variational approximation

$$\underbrace{Y}_{\text{Target}} \leftrightarrow \underbrace{X}_{\text{Input}} \rightarrow \underbrace{Z}_{\text{Feature}} \rightarrow \underbrace{\hat{Z}}_{\text{Rec. Feature}} \rightarrow \underbrace{\hat{Y}}_{\text{Ouput}}$$

Intractable objective

$$\boxed{-I(Y,\hat{Z}) + \beta I(\hat{Z},X)} = -\int p(\boldsymbol{y} \mid \hat{\boldsymbol{z}}) p(\hat{\boldsymbol{z}}) \log p(\boldsymbol{y} \mid \hat{\boldsymbol{z}}) d\boldsymbol{y} d\hat{\boldsymbol{z}} + \beta \int p_\phi(\hat{\boldsymbol{z}} \mid \boldsymbol{x}) p(\boldsymbol{x}) \log \frac{p_\phi(\hat{\boldsymbol{z}} \mid \boldsymbol{x})}{p(\hat{\boldsymbol{z}})} d\boldsymbol{x} d\hat{\boldsymbol{z}} - H(Y)$$

Variational bound

$$\leq \underbrace{-\int p(\boldsymbol{y} \mid \hat{\boldsymbol{z}}) p(\hat{\boldsymbol{z}}) \log q_\theta(\boldsymbol{y} \mid \hat{\boldsymbol{z}}) d\boldsymbol{y} d\hat{\boldsymbol{z}}}_{\text{Cross-Entropy}} + \underbrace{\beta \int p_\phi(\hat{\boldsymbol{z}} \mid \boldsymbol{x}) p(\boldsymbol{x}) \log \frac{p_\phi(\hat{\boldsymbol{z}} \mid \boldsymbol{x})}{q(\hat{\boldsymbol{z}})} d\boldsymbol{x} d\hat{\boldsymbol{z}}}_{\text{KL-Divergence}} - \underbrace{H(Y)}_{\text{constant}}$$

Variational Information Bottleneck (VIB) objective

$$\mathcal{L}_{VIB}(\boldsymbol{\phi}, \boldsymbol{\theta}) = \mathbf{E}_{p(\boldsymbol{x},\boldsymbol{y})} \left\{ \mathbf{E}_{p_{\boldsymbol{\phi}}(\hat{\boldsymbol{z}}|\boldsymbol{x})} \left[-\log q_{\boldsymbol{\theta}}(\boldsymbol{y}|\hat{\boldsymbol{z}}) \right] + \beta D_{KL} \left(p_{\boldsymbol{\phi}}(\hat{\boldsymbol{z}}|\boldsymbol{x}) \| q(\hat{\boldsymbol{z}}) \right) \right\}.$$

Empirical estimation

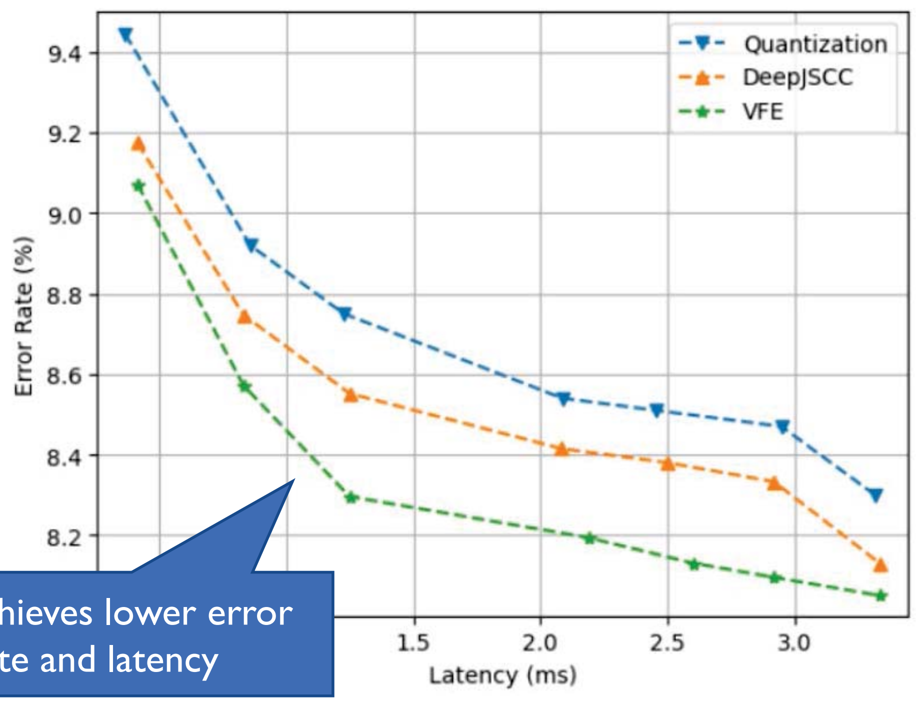
$$\simeq \frac{1}{M} \sum_{m=1}^{M} \left\{ \frac{1}{L} \sum_{l=1}^{L} \left[-\log q_{\boldsymbol{\theta}} \left(\boldsymbol{y}_{\boldsymbol{m}} | \hat{\boldsymbol{z}}_{\boldsymbol{m},\boldsymbol{l}} \right) \right] + \beta D_{KL} \left(p_{\boldsymbol{\phi}} \left(\hat{\boldsymbol{z}} | \boldsymbol{x}_{\boldsymbol{m}} \right) \| q(\hat{\boldsymbol{z}}) \right) \right\}$$

- Variational approximations
  - $p_\phi(\hat{z}|x)$ is defined by the neural network (encoder)
  - $q_\theta(y|\hat{z})$ is a variational distribution to approximate $p(y|\hat{z})$
  - $q(\hat{z})$ is a variational distribution to approximate $p(\hat{z})$

Experiment

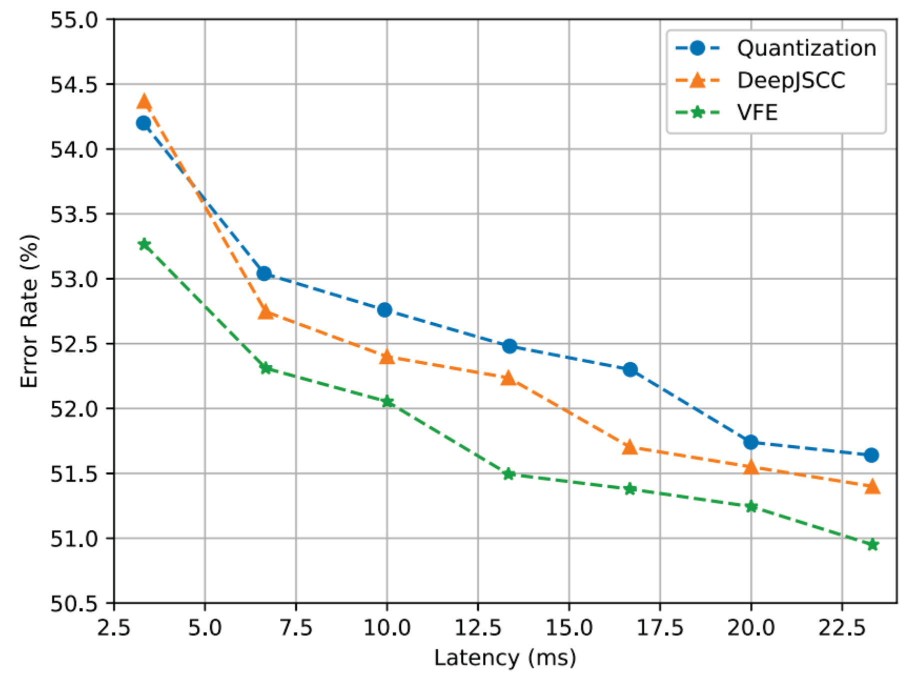
- **Baselines** (data-oriented communication):
  - DeepJSCC (Joint Source-Channel Coding)
  - Learning-based quantization (w/ ideal channel coding)

Rate-distortion on CIFAR-10 dataset



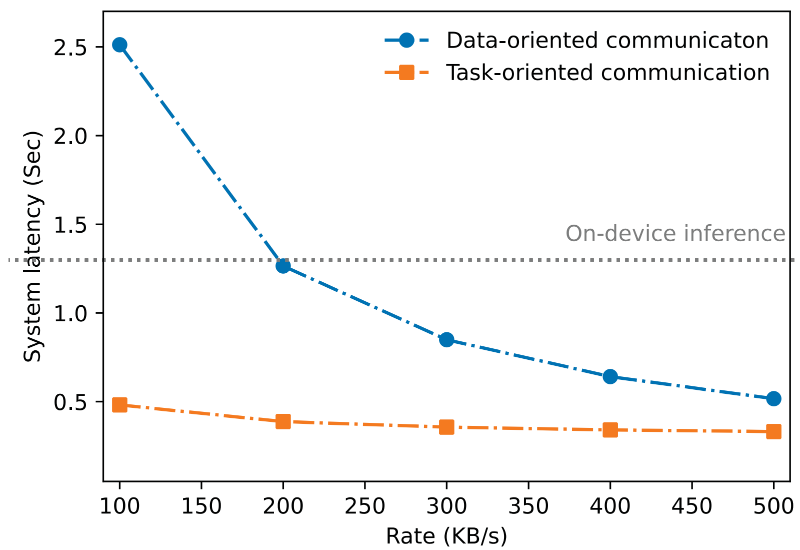
VFE achieves lower error rate and latency

Rate-distortion on Tiny ImageNet dataset



Experiment

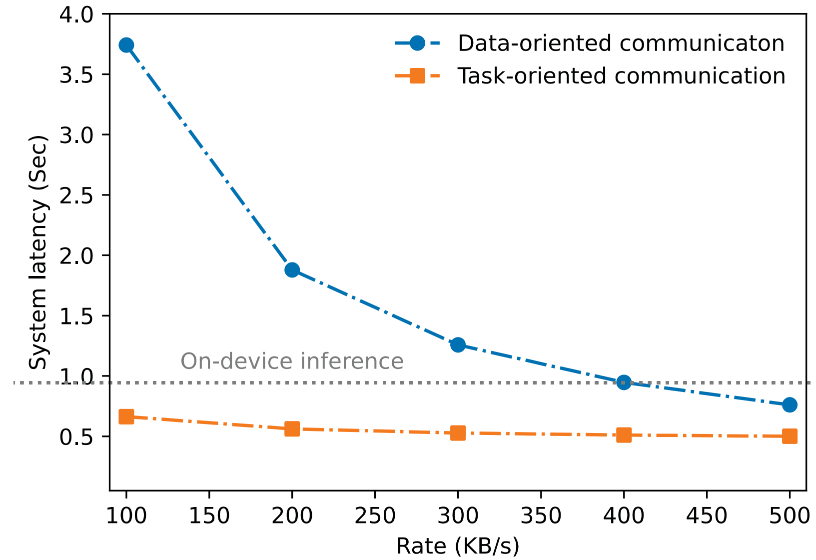
Image captioning



Visual question answering (VQA)

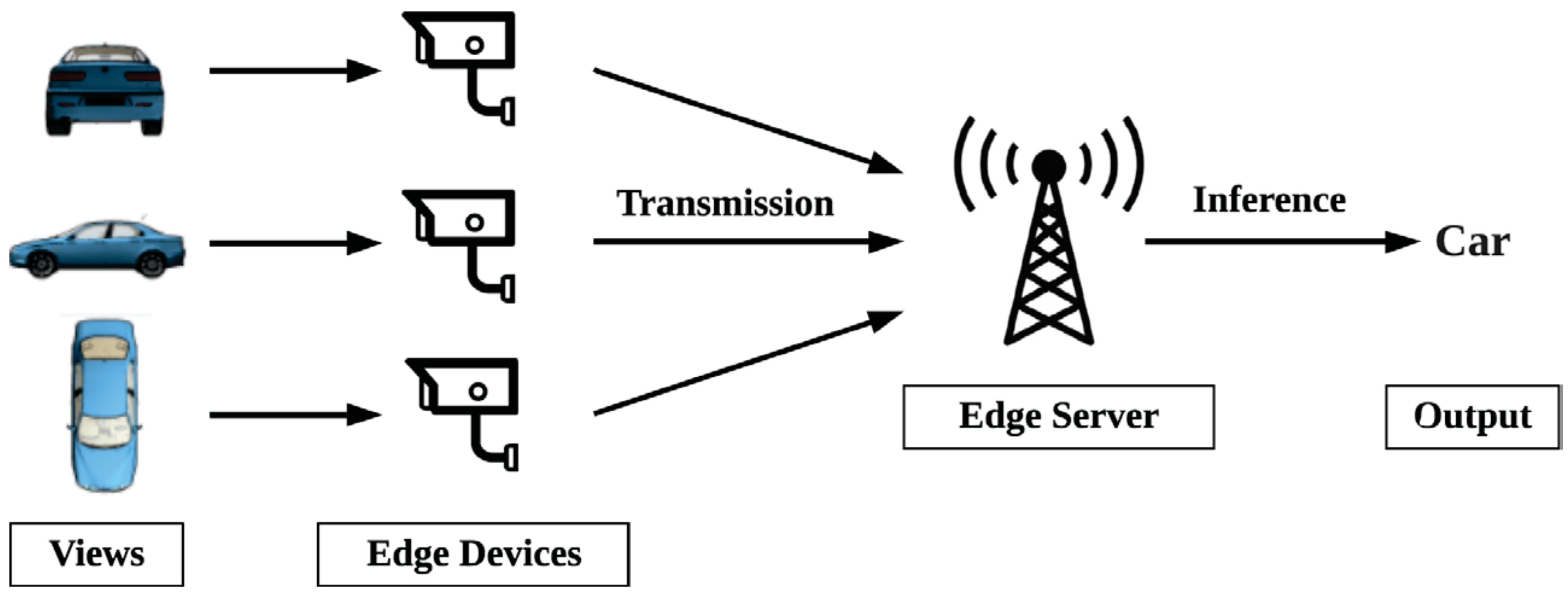
Q: What color are stop lights?

A: Red

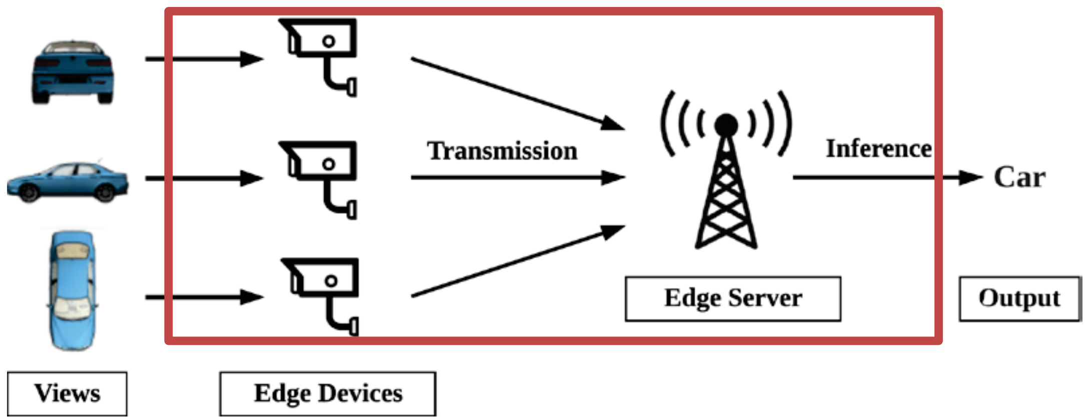


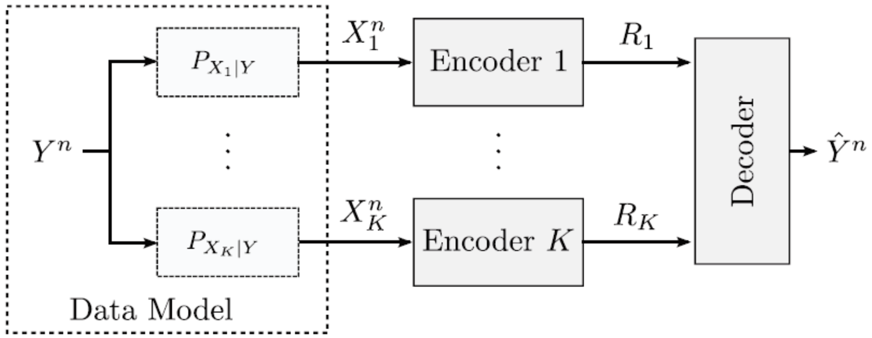
Multi-camera cooperative inference

- Objective: Design an efficient method that can fully exploit the correlation among multiple features in distributed feature encoding.



Cooperative perception vs. Distributed Information Bottleneck (DIB)





Distributed Information Bottleneck (DIB)

Closely related to the distributed Chief Executive Officer (CEO) source coding problem

Proposition. Suppose the input variables $X_k, \forall k = 1, 2, \ldots, K$ are conditional independet given Y. Given the relevance $\Delta = I(Y; Z_{1:K})$, the sum rate

Rate Relevance

$$\sum_{k=1}^{K} R_k \geq \Delta + \sum_{k=1}^{K} [I(X_k; Z_k) - I(Y; Z_k)]$$

Relevance-rate tradeoff

Aguerri, Inaki Estella, and Abdellatif Zaidi. "Distributed variational representation learning." *IEEE Trans. Pattern Anal. Machine Intell.* 120-138, 2019.

Multi-camera cooperative inference

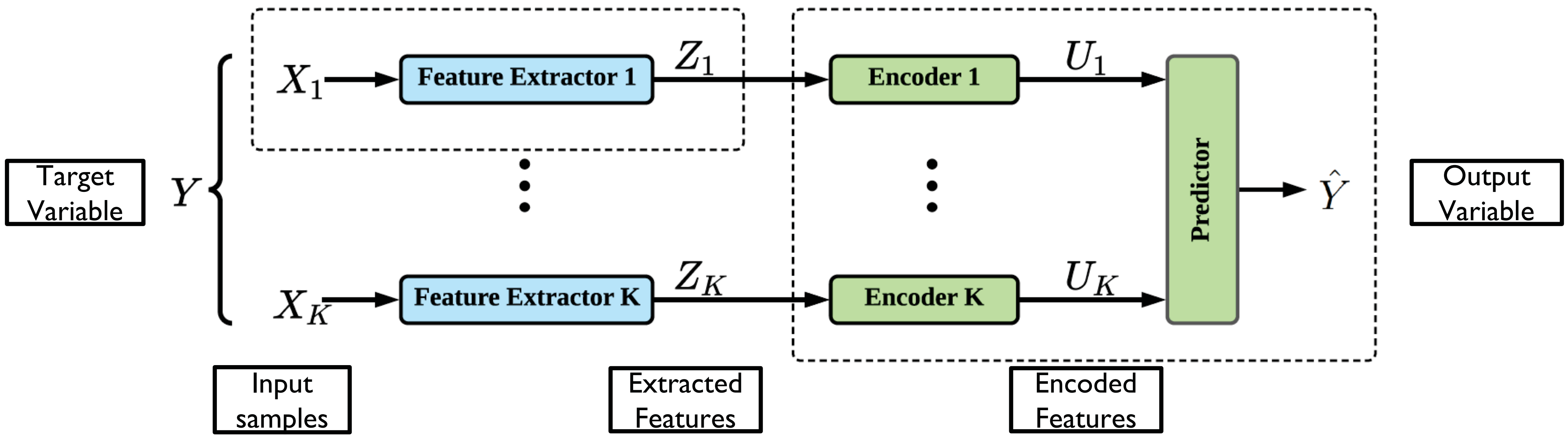
- Probabilistic modeling with *K* devices
- Loss functions

Information Bottleneck (IB)

$$\mathcal{L}_{\mathrm{IB,k}}(\gamma) := H(Y \mid Z_k) + \gamma I(X_k; Z_k)$$

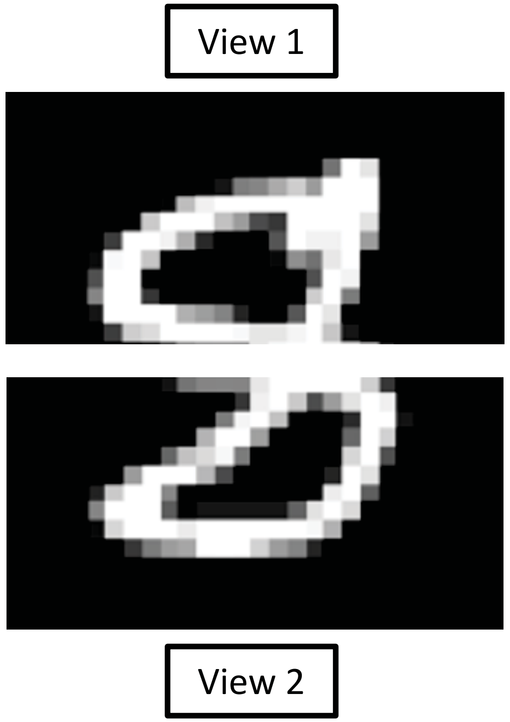
Distributed IB (DIB)

$$\mathcal{L}_{\mathrm{DIB}}(\beta) := H(Y \mid U_{1:K}) + \beta \sum_{k=1}^{K} [H(Y \mid U_k) + I(Z_k; U_k)]$$



Performance evaluation

- Cooperative inference tasks



Two-view MNIST classification

Twelve-view Shape Recognition on ModelNet40 dataset

Experiment

- The accuracy of the cooperative tasks under different bit constraints.

- Task-oriented vs. Data-oriented
  - Two-view MNIST classification task: ~10 bits vs. 1.3 kbits
  - Twelve-view ModelNet40 Shape recognition task: ~200 bits vs. 120 KB

Image Classification

	R_{sum}		
	6 bits	10 bits	14 bits
NN-REG	95.93%	97.49%	97.78%
NN-GBI	96.62%	97.79%	98.02%
eSAFS	96.97%	97.87%	98.05%
CAFS	94.14%	97.43%	97.42%
VDDIB (ours)	97.08%	97.82%	98.06%
VDDIB-SR (T=2) (ours)	**97.13%**	**98.13%**	**98.22%**

Shape recognition

	R_{sum}		
	120 bits	240 bits	360 bits
NN-REG	87.50%	88.25%	89.03%
NN-GBI*	88.82%	—	—
eSAFS	85.88%	87.87%	89.50%
CAFS	86.75%	89.56%	90.67%
VDDIB (ours)	89.25%	90.03%	90.75%
VDDIB-SR (T=2) (ours)	**90.25%**	**91.31%**	**91.62%**

* The GBI quantization algorithm is computationally prohibitive when the number of bits is too large.

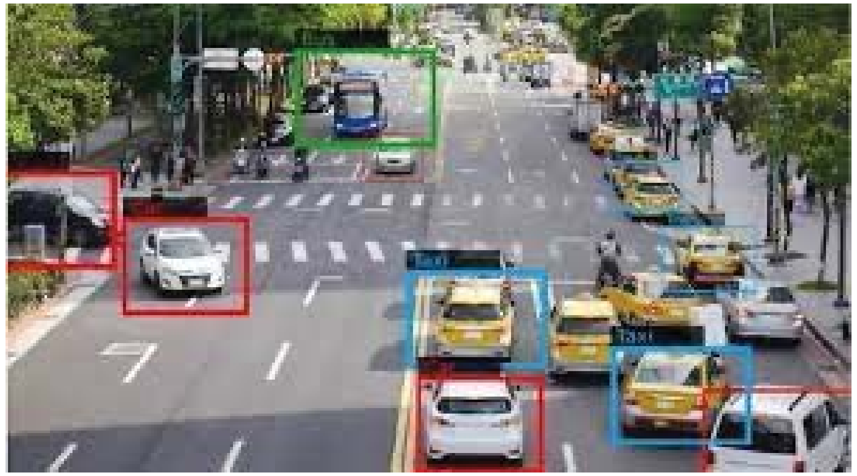
Case study I: Edge video analytics

J. Shao, X. Zhang, and **J. Zhang**, "Task-oriented communication for edge video analytics," *IEEE Transactions on Wireless Communications*, to appear. (https://arxiv.org/abs/2211.14049)

Edge video analytics

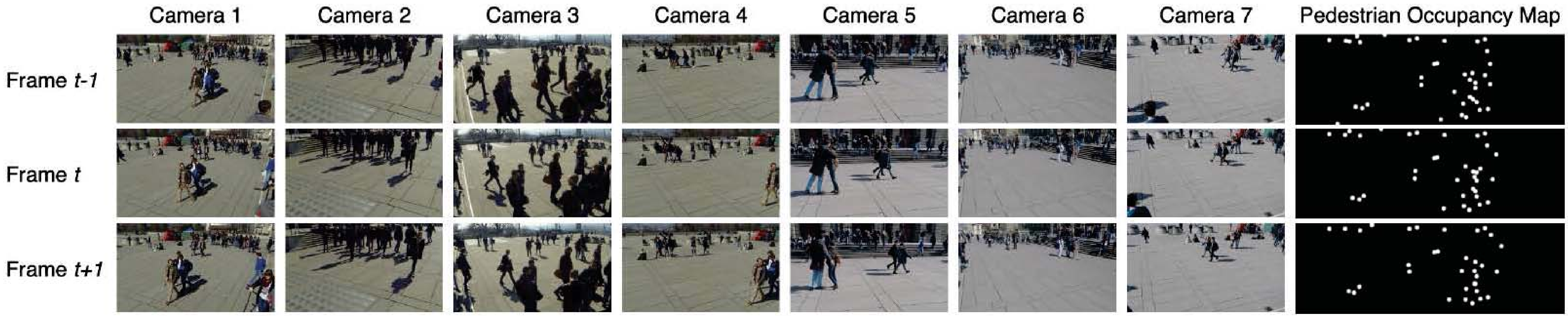
- More and more cameras and video data at the edge

- Powerful AI models for visual data



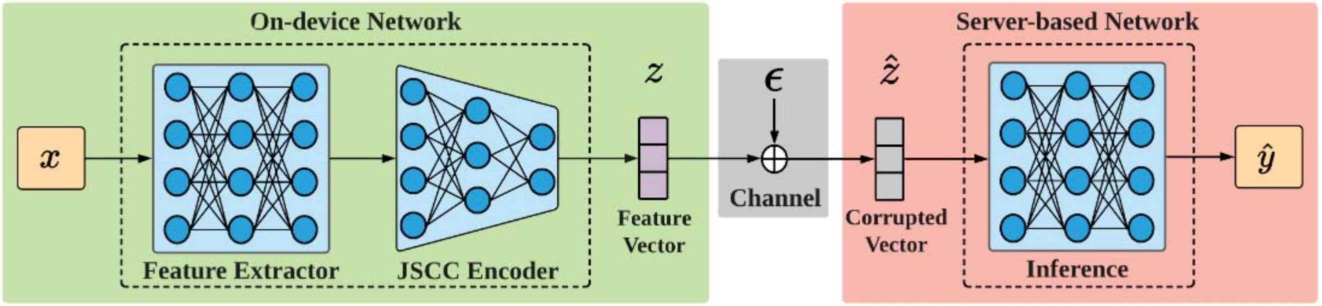
An example of edge video analytics

- Challenges in edge video analytics:
  - How to effectively exploit the **temporal dependence** among frames.
  - How to effectively leverage the **spatial correlation** among cameras.

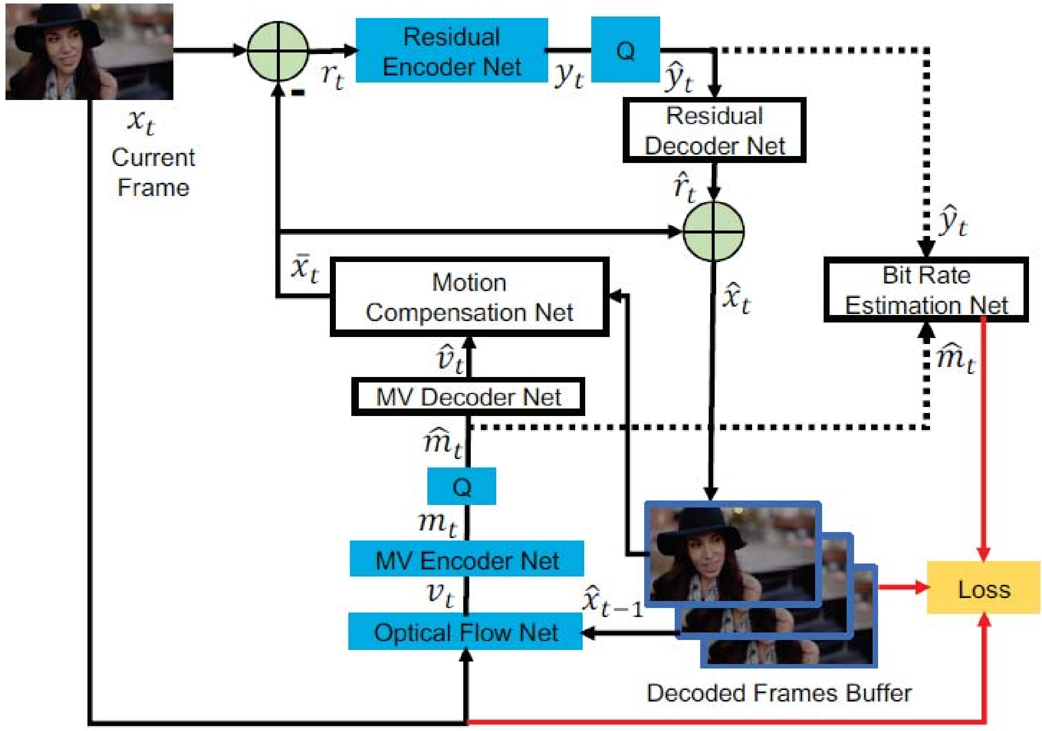


Existing methods

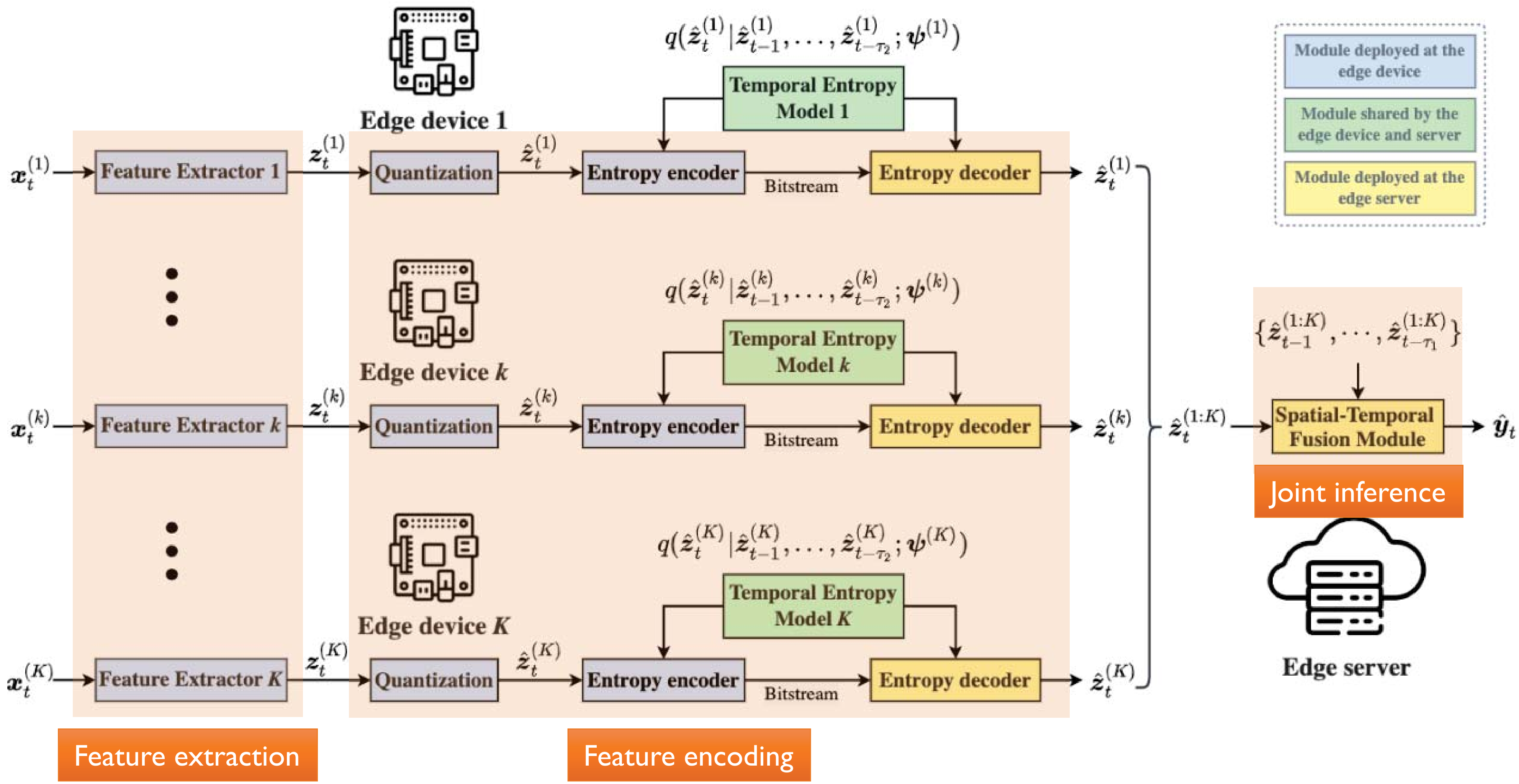
- **VFE**
  - Task-oriented
  - Only for images



- **DVC** (Deep video compression)
  - Efficient in extracting temporal correlation
  - But data-oriented

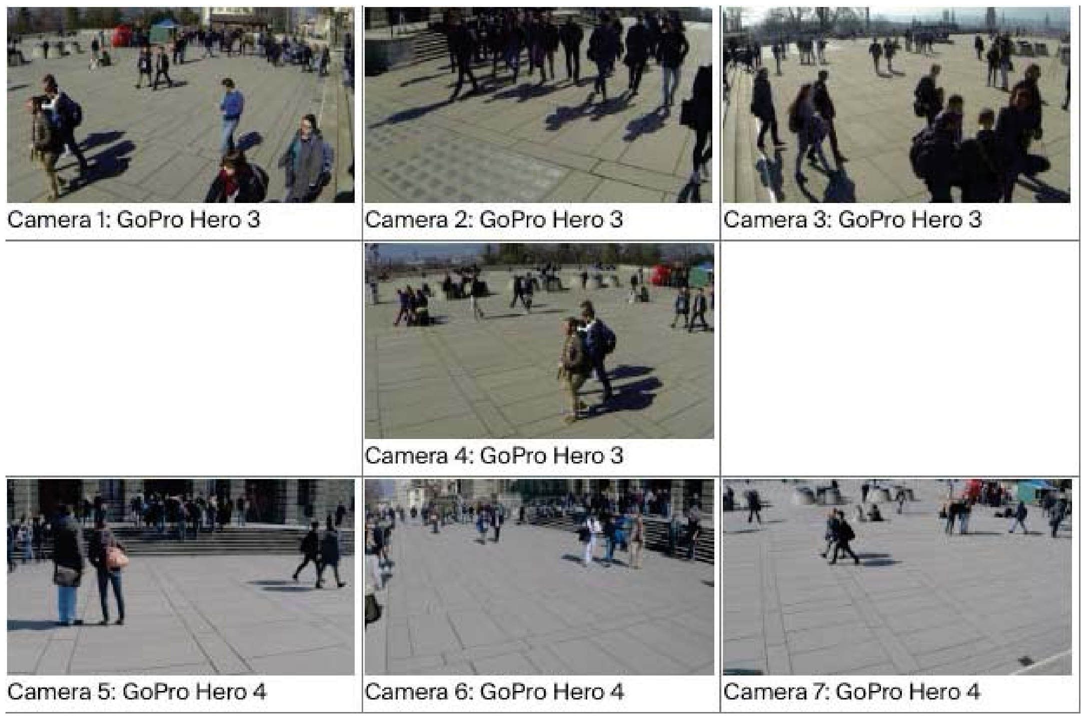


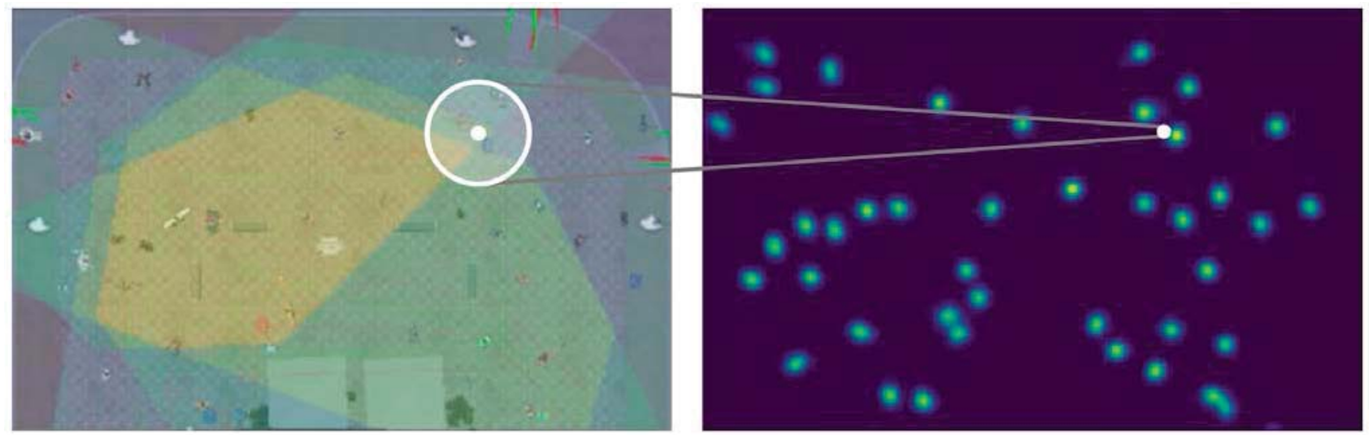
Proposed method



Experimental results

- Multi-camera pedestrian occupancy prediction (Wildtrack dataset)





ground plane (bird's eye view)

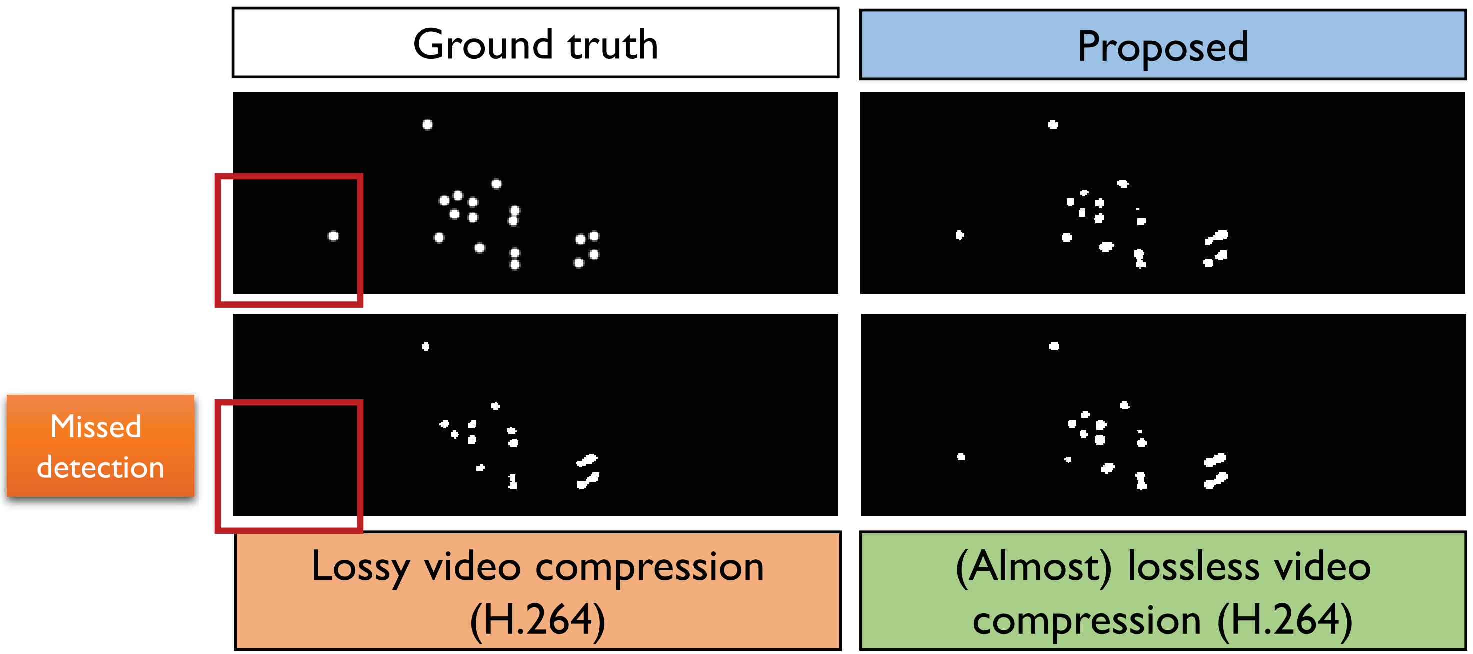
Output: pedestrian occupancy map

- Chavdarova, Tatjana, Pierre Baqué, Stéphane Bouquet, Andrii Maksai, Cijo Jose, Timur Bagautdinov, Louis Lettry, Pascal Fua, Luc Van Gool, and François Fleuret. "Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5030-5039. 2018.

Utility vs. communication overhead

- **Metric**:
  - Cost: Communication overhead per frame.
  - Performance: Multi-object detection accuracy.
- **Output**: Pedestrian occupancy map
- **Baseline**: Video coding (H.264)

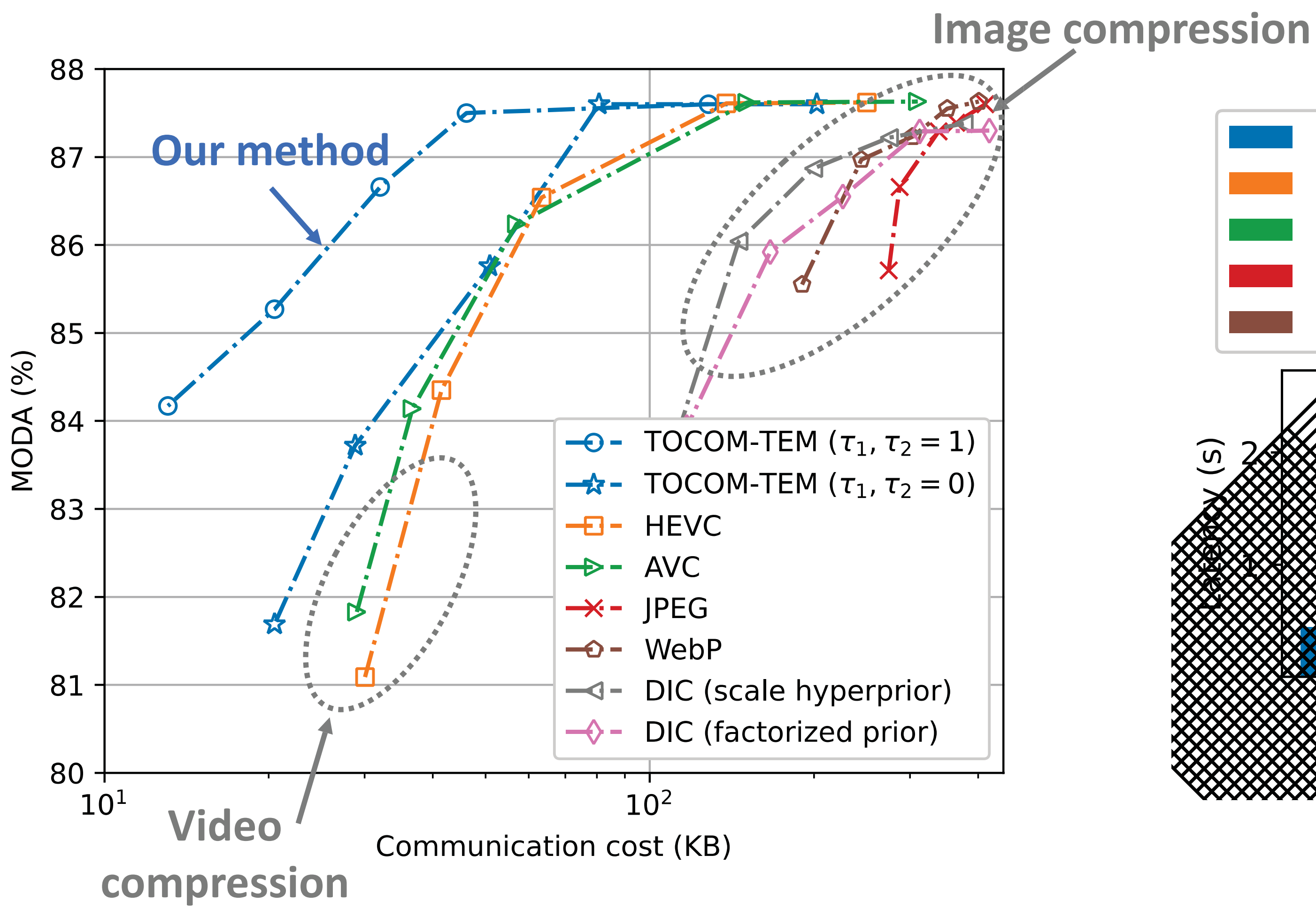
- Lossless video compression achieves performance comparable to our task-oriented communication scheme, but it results in significantly higher communication overhead.
- While lossy video compression can match the communication cost of our method, it comes at the cost of performance degradation.

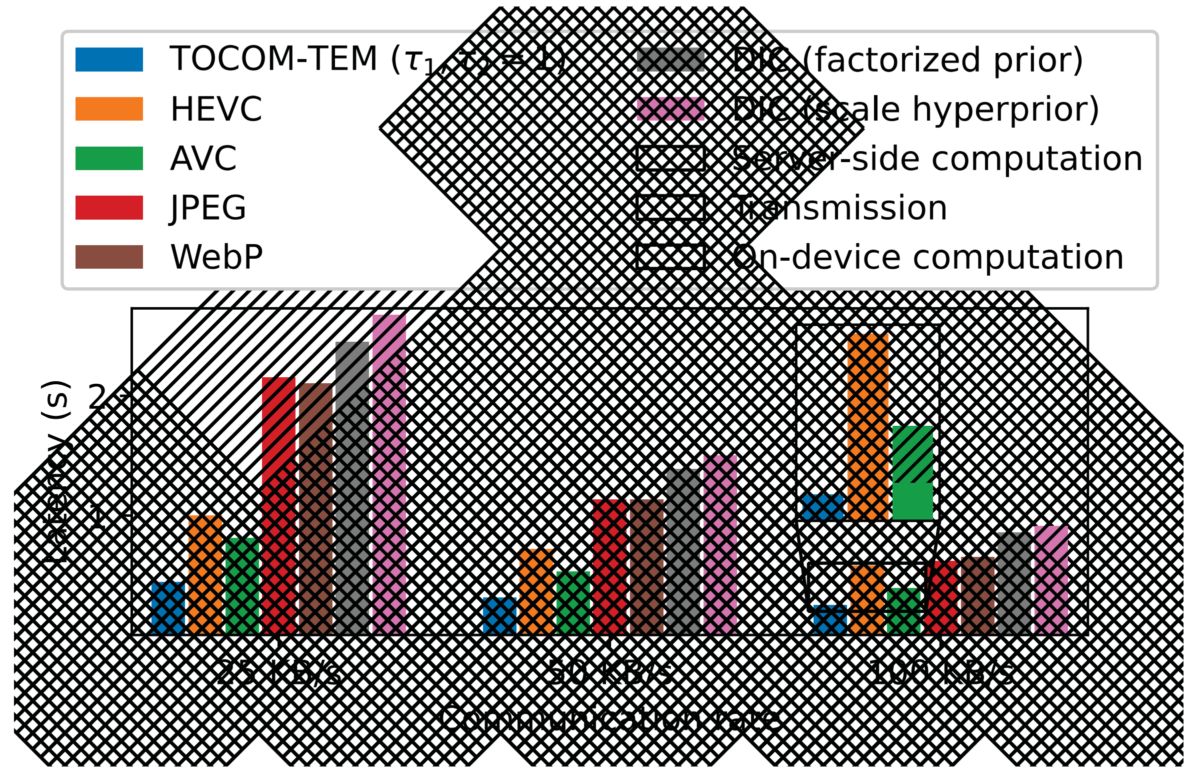


Method	Communication cost per frame (KB) ↓	Multi-object detection accuracy (%) ↑
Task-oriented communication (Proposed)	6.1	87.3
H.264 (almost lossless)	614.6	87.3
H.264 (lossy)	6.1	84.6

Utility vs. communication overhead

- Metric: Multi-object detection accuracy (MODA).

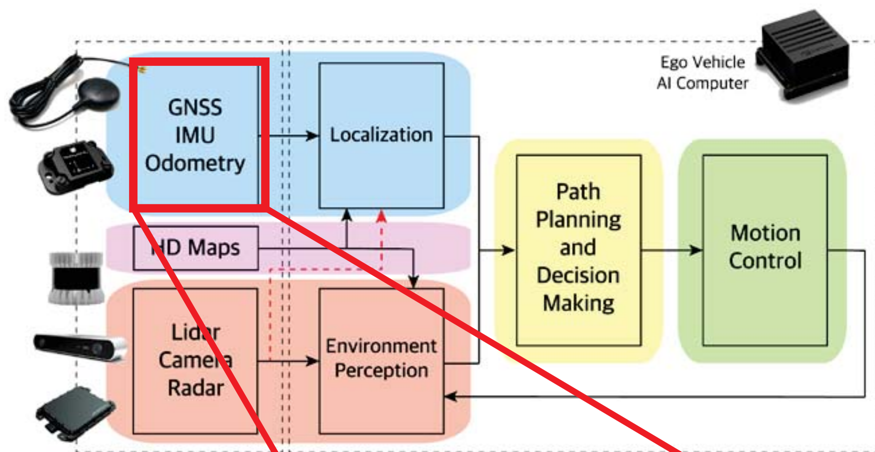


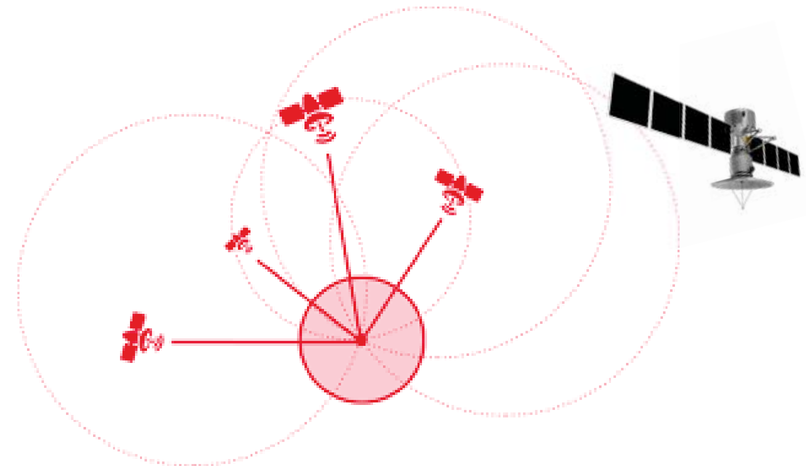


Case study II: Localization for autonomous driving

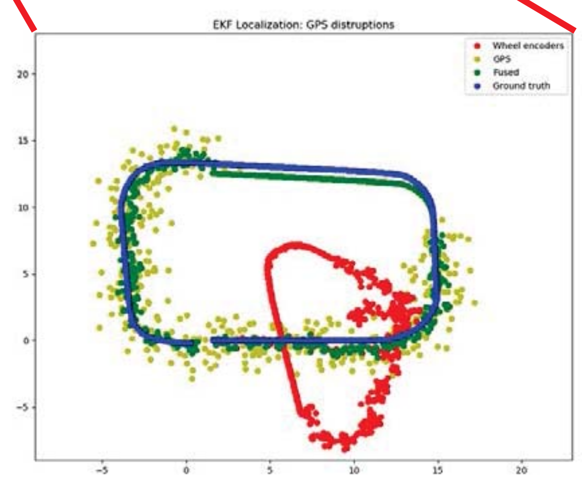
B. Liu, J. Zhang, "ParaLoc: A Communication-Adaptive Parallel System for Real-Time Localization in Infrastructure-Assisted Autonomous Driving," in preparation.

Self-driving needs absolute localization





Red: IMU
Yellow: GPS
Greed: Fused
Blue: Ground Truth



Absolute localization e.g. GPS is important to correct the cumulative error of the on-board sensors.

Challenges of GPS-based solution:
- Affected by **weather** conditions.
- Unable to provide localization in **indoor** environments.
- Difficult to estimate the **uncertainty** of localization.

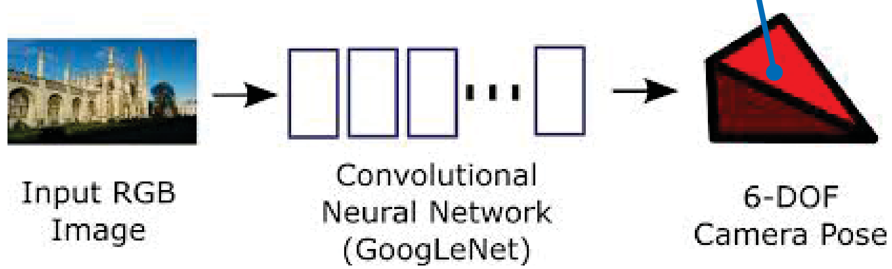
Is there any alternative or complementary way to GPS for absolute localization?

How about RSU + Deep Learning → Absolute localization ?

Mobile devices

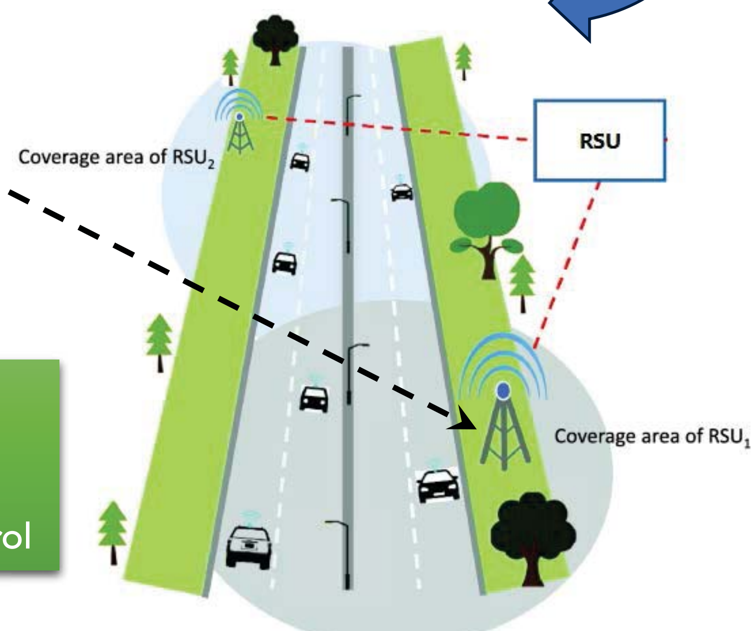
DNN models have limitations regarding **reliability**, **high computational resource requirements**, and **inference latency**, which prevent their direct deployment in real-world autonomous driving applications.

Absolute localization



RSUs enable deep learning models to be practical for self-driving.

Deep learning Advantages :
- High precision.
- Rapid development.
- Easy to deploy and update.



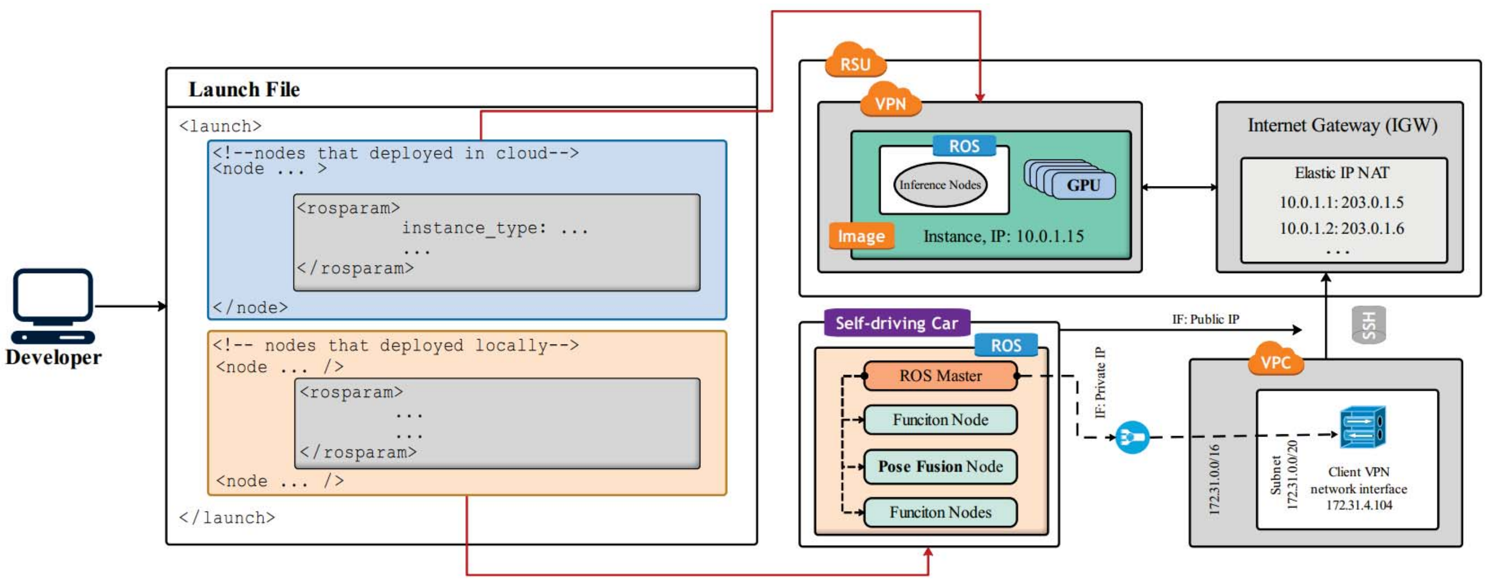
RSU Advantages :
- Minimal weather impact
- Ability to provide services indoors
- Convenient for maintenance and control

Communication

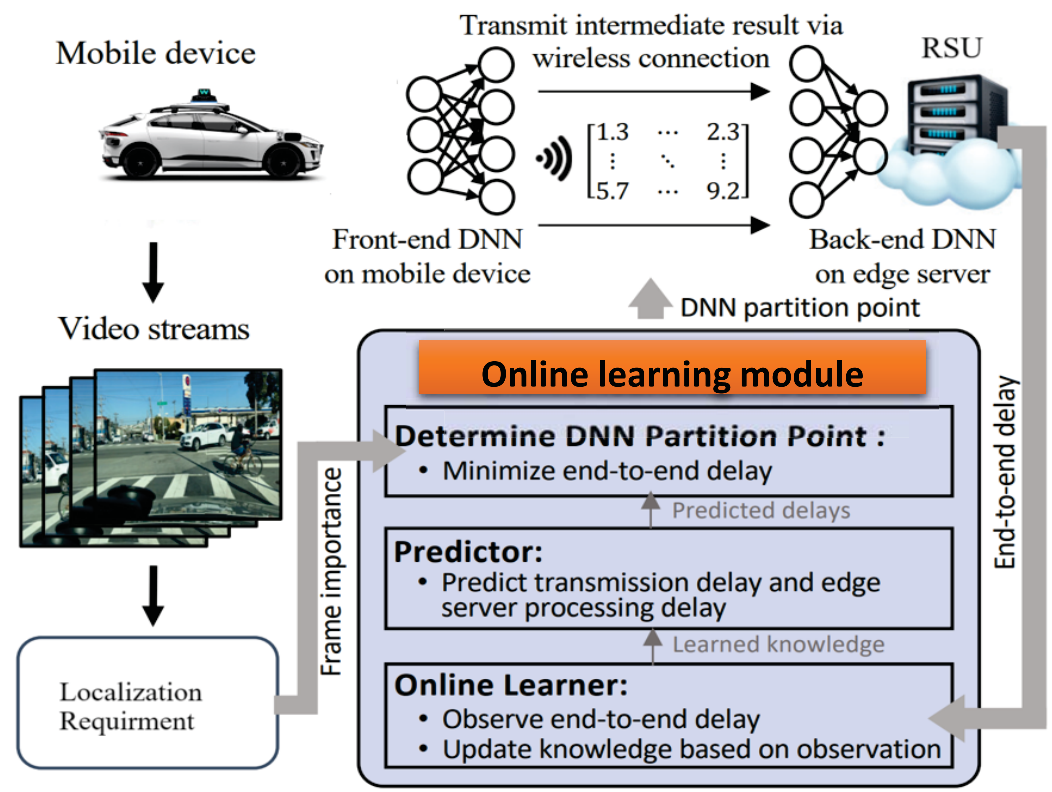
Absolute localization

System design and implementation

ROS based parallel system



Task-oriented communication



Performance

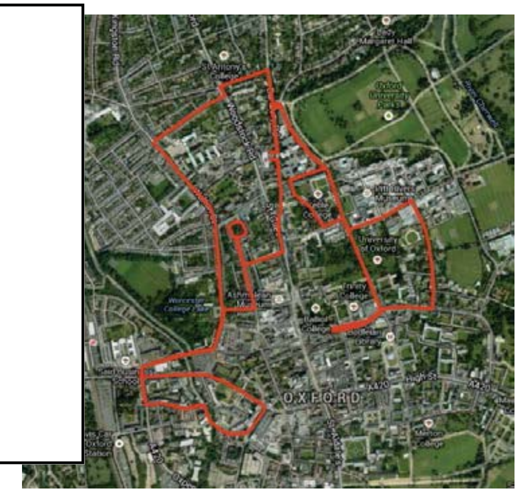
× 16 speed

Traditional

Deep learning model

Uncertainty-aware

Frames



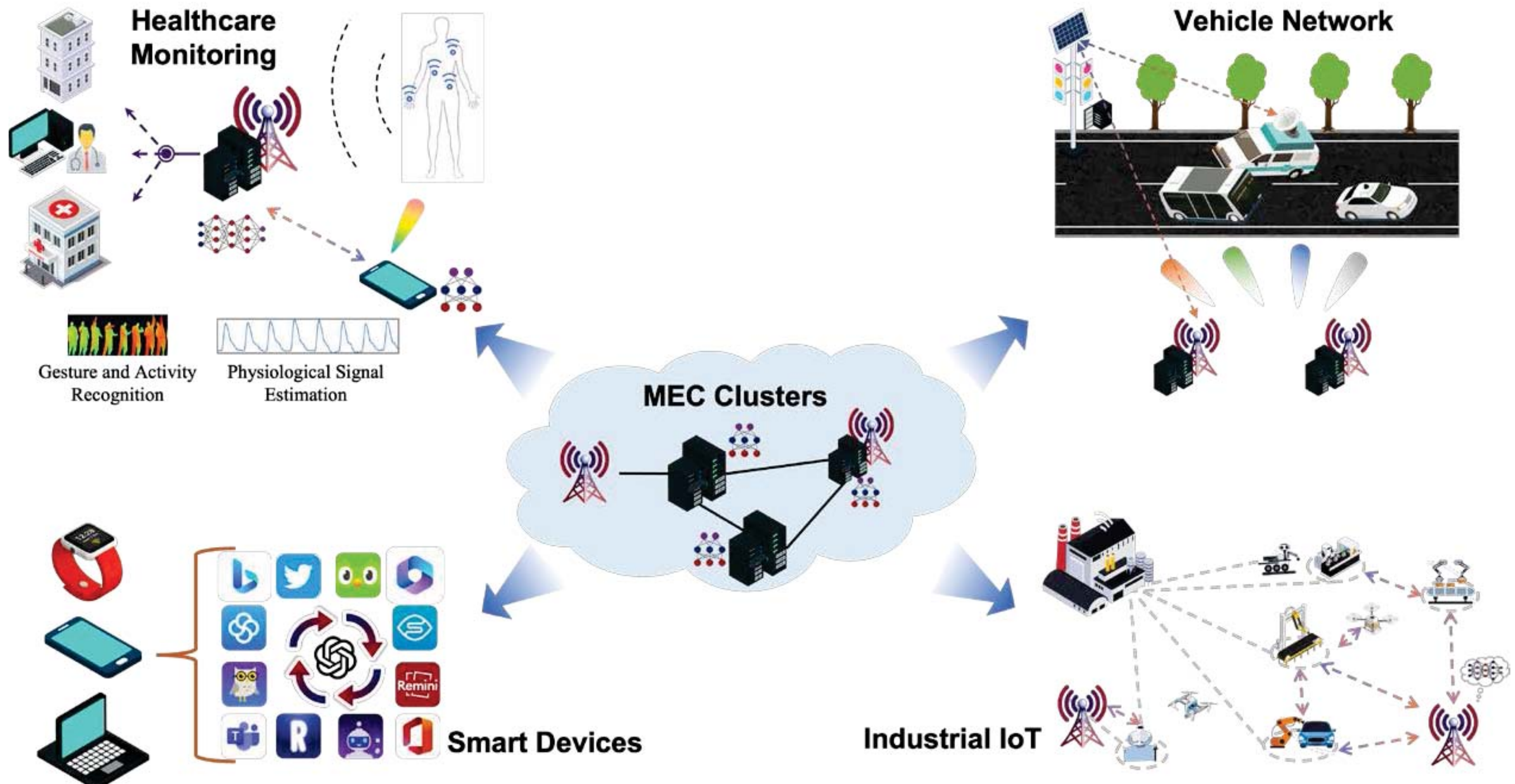
Oxford 10km

Case study III: EdgeGPT for autonomous edge AI

Y. Shen, J. Shao, X. Zhang, Z. Lin, H. Pan, D. Li, **J. Zhang**, and K. B. Letaief, "Large language models empowered autonomous edge AI for connected intelligence," *IEEE Commun. Mag.*, to appear. (https://arxiv.org/abs/2307.02779)

Vision of Edge AI

- Edge AI offers a promising solution for **connected intelligence** by allowing data collection, processing, transmission, and consumption at the network edge.

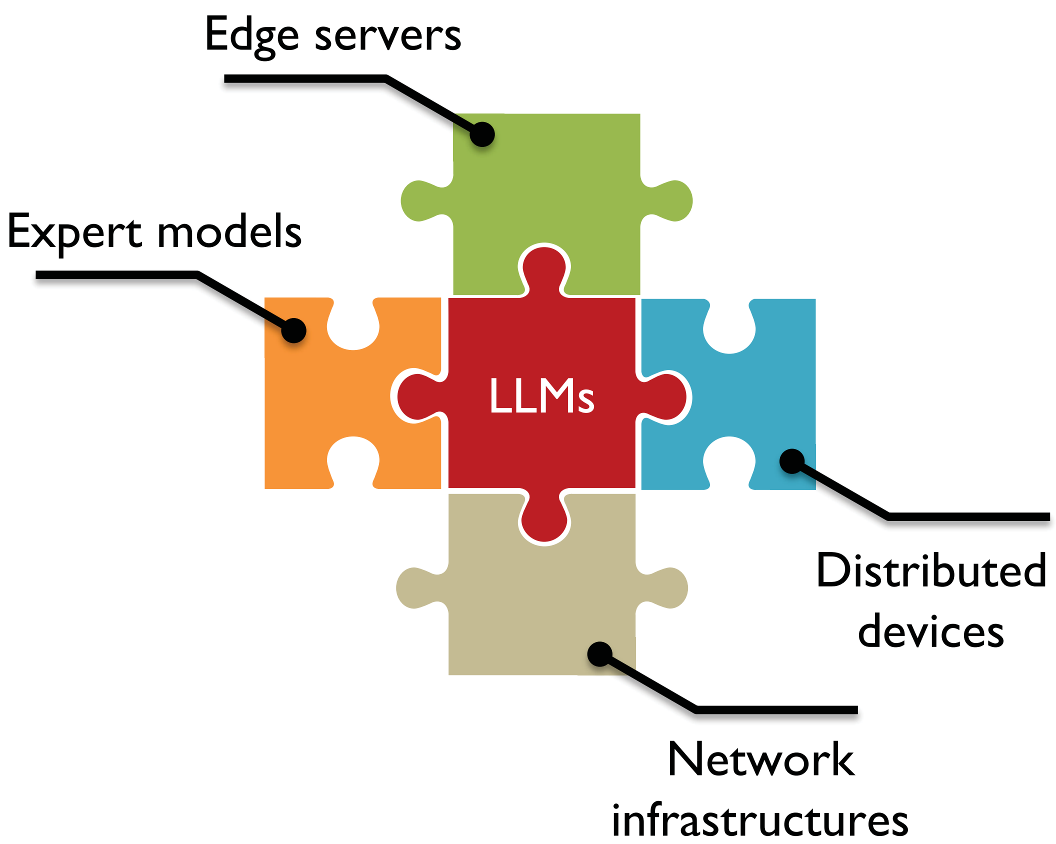


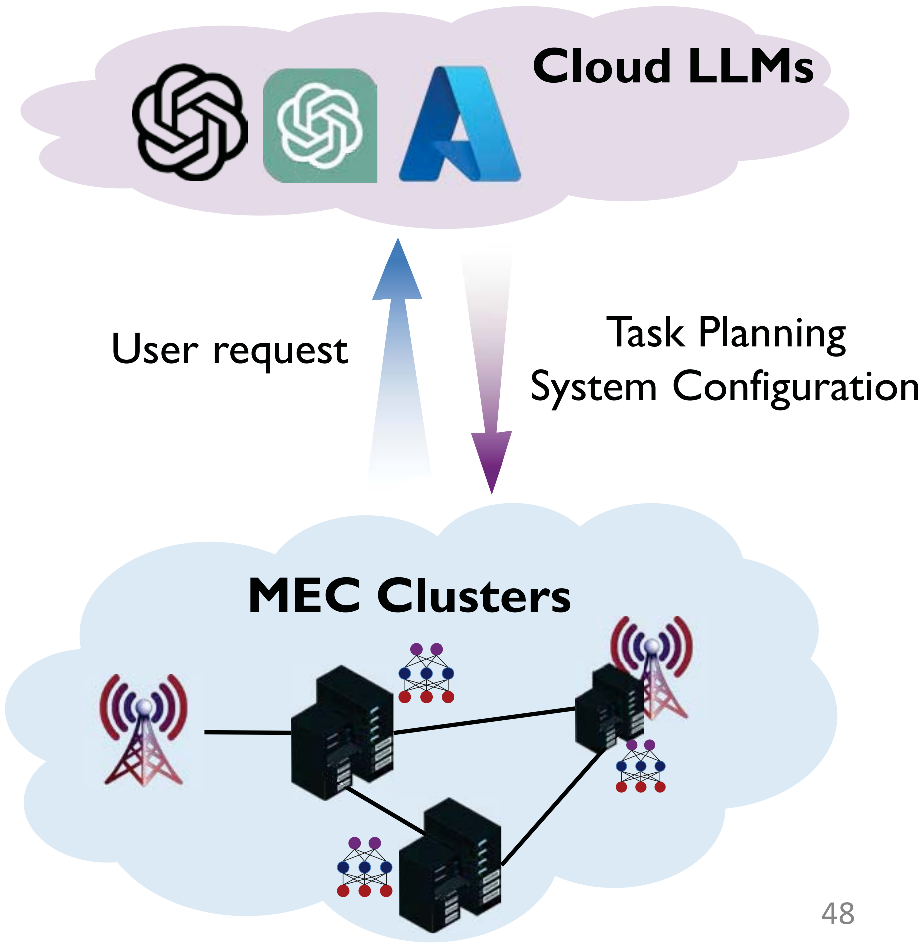
High system complexity

- To effectively meet the evolving demands and new requests of users, it is crucial for distributed devices, edge servers, expert AI models, and network infrastructures to work together seamlessly.

To enable autonomous Edge AI via LLMs

- Idea: To utilize a cloud server with LLMs for task planning and system configuration, adapting to user requests.





Autonomous task planning and model selection

- Available tasks, models, and datasets for evaluation:
  - Image classification, ViT model, ImageNet dataset
  - Image caption, blip-image-captioning-base, COCO Karpathy dataset
  - Visual question answering, blip-vqa-base, VQA v2 dataset

Request: What kind of animal is in the image?

{Task planning: image classification, Selected model: ViT model}
Output: Dog

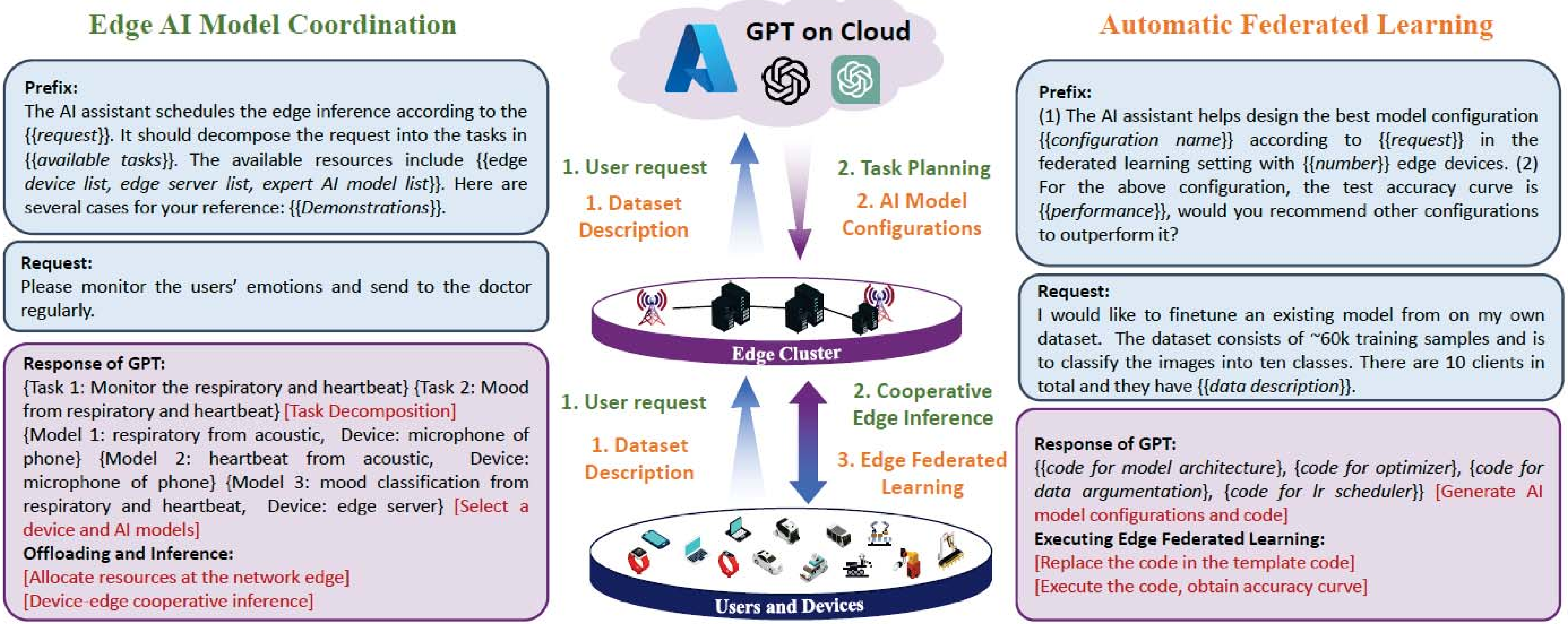
Request: Briefly describe this image.

{Task planning: image caption, Selected model: blip-image-captioning-base}
Output: A horse carrying a large load of hay and two people sitting on it.

Request: Tell me what the mustache is made of in this image.

{Task planning: Visual question answering, Selected model: blip-vqa-base}
Output: Bananas

EdgeGPT



Edge AI Model Coordination
Prefix:
The AI assistant schedules the edge inference according to the {{request}}. It should decompose the request into the tasks in {{available tasks}}. The available resources include {{edge device list, edge server list, expert AI model list}}. Here are several cases for your reference: {{Demonstrations}}.
Request:
Please monitor the users' emotions and send to the doctor regularly.
Response of GPT:
{Task 1: Monitor the respiratory and heartbeat} {Task 2: Mood from respiratory and heartbeat} [Task Decomposition]
{Model 1: respiratory from acoustic, Device: microphone of phone} {Model 2: heartbeat from acoustic, Device: microphone of phone} {Model 3: mood classification from respiratory and heartbeat, Device: edge server} [Select a device and AI models]
Offloading and Inference:
[Allocate resources at the network edge]
[Device-edge cooperative inference]
GPT on Cloud
1. User request
1. Dataset Description
2. Task Planning
2. AI Model Configurations
Edge Cluster
1. User request
1. Dataset Description
2. Cooperative Edge Inference
3. Edge Federated Learning
Users and Devices
Automatic Federated Learning
Prefix:
(1) The AI assistant helps design the best model configuration {{configuration name}} according to {{request}} in the federated learning setting with {{number}} edge devices. (2) For the above configuration, the test accuracy curve is {{performance}}, would you recommend other configurations to outperform it?
Request:
I would like to finetune an existing model from on my own dataset. The dataset consists of ~60k training samples and is to classify the images into ten classes. There are 10 clients in total and they have {{data description}}.
Response of GPT:
{{code for model architecture}, {code for optimizer}, {code for data argumentation}, {code for lr scheduler}} [Generate AI model configurations and code]
Executing Edge Federated Learning:
[Replace the code in the template code]
[Execute the code, obtain accuracy curve]

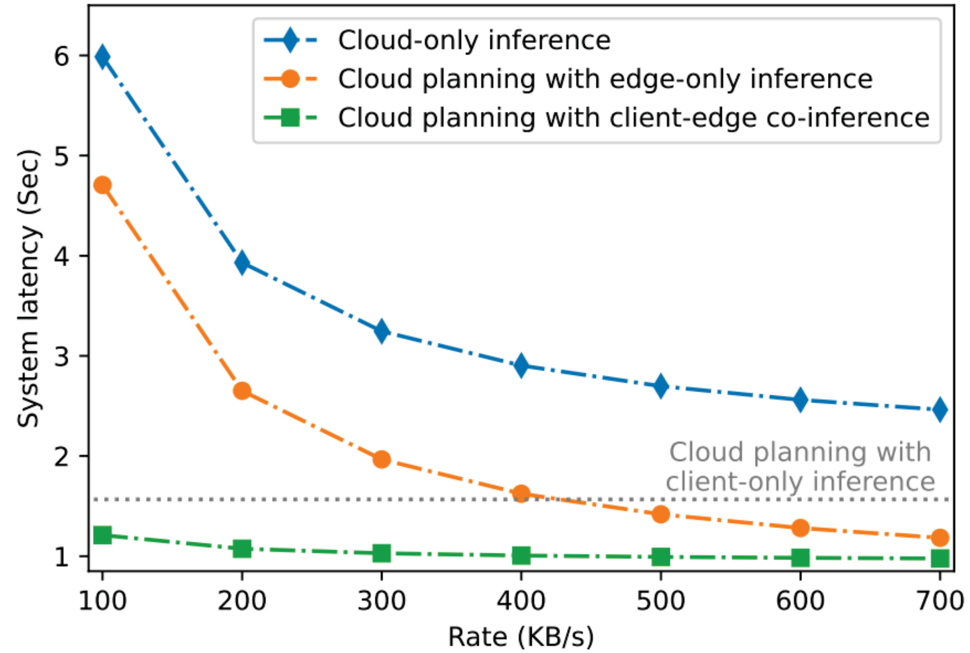
Performance: Automatic edge inference

Automatic task planning

Model	Acc↑	F1↑	Latency↓
GPT-3 350M	32.93%	33.95%	0.37 sec
GPT-3 6.7B	40.24%	42.31%	0.55 sec
GPT-3 175B	68.89%	74.70%	0.45 sec
GPT-3 175B IT	**84.44%**	**85.39%**	0.58 sec
Zero-shot Classification	36.59%	36.59%	**0.28 sec**

Large models are needed

End-to-end latency



Communication cost and performance in edge inference

	Method	Image classification Cost↓	Image classification Accuracy↑	Image caption Cost↓	Image caption BLEU↑	Image caption CIDEr↑	Image caption SPICE↑	Visual question answering Cost↓	Visual question answering Test-dev↑	Visual question answering Test-std↑
Data-oriented	Edge-only inference with lossless data compression	224.41 KB	**84.16%**	342.15 KB	39.62	**133.04**	23.72	503.50 KB	**77.38**	**77.50**
Data-oriented	Edge-only inference with lossy data compression	33.86 KB	82.83%	19.77 KB	39.13	131.12	23.45	21.55 KB	76.75	76.73
Task-oriented	Client-edge co-inference	**32.83 KB**	84.02%	**19.50 KB**	**39.83**	132.92	**23.74**	**21.54 KB**	77.30	77.40

Conclusions

Conclusions

- Task-oriented communication
  - Shift from "how to communicate" to "what to communicate"
- Task-oriented communication for Edge AI
  - Edge-assisted inference via information bottleneck
  - Cooperative perception via distributed information bottleneck
- Interesting applications
  - Edge video analytics
  - Edge-assisted localization
  - …

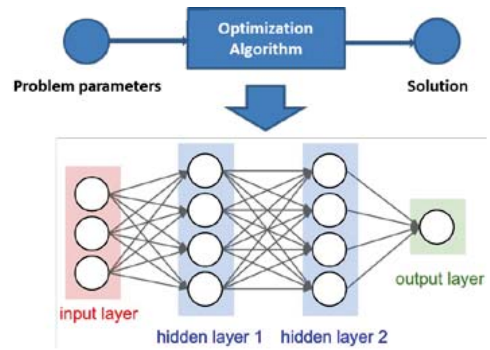
References

- J. Shao, **J. Zhang**, "Communication-computation trade-off in resource-constrained edge inference," *IEEE Commun. Mag.*, vol. 58, no. 12, pp. 20–26, Dec. 2020.
- J. Shao, Y. Mao, **J. Zhang**, "Learning task-oriented communication for edge inference: An information bottleneck approach," *IEEE J. Select. Areas Commun.*, vol. 40, no. 1, pp. 197-211, Jan. 2022.
- J. Shao, Y. Mao, and **J. Zhang**, "Task-oriented communication for multi-device cooperative edge inference," *IEEE Trans. Wireless Communications*, vol. 11, no. 1, pp. 73-87, Jan. 2023.
- J. Shao, X. Zhang, and **J. Zhang**, "Task-oriented communication for edge video analytics," *IEEE Transactions on Wireless Communications,* to appear.
- Y. Shen, J. Shao, X. Zhang, Z. Lin, H. Pan, D. Li, **J. Zhang**, and K. B. Letaief, "Large language models empowered autonomous edge AI for connected intelligence," *IEEE Commun. Mag.*, to appear. (https://arxiv.org/abs/2307.02779)

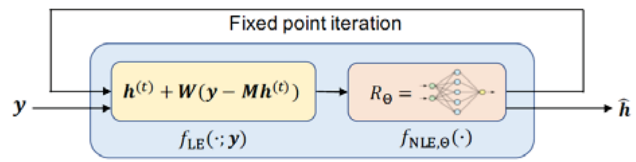
Overview of my research

AI4COM

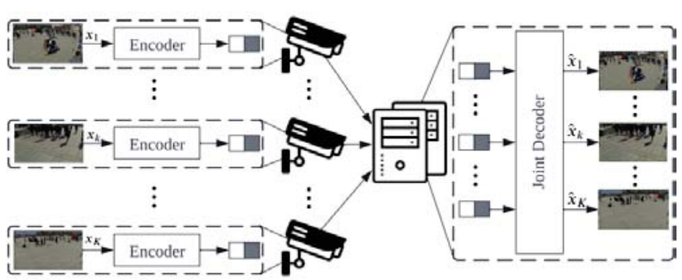
Learning to optimize
- When the problem scale is enormous



Learning to estimate
- When the system is intractable to model

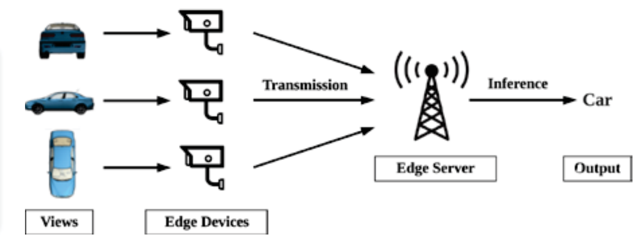


Learning to compress
- When the source model and distortion metric is intractable

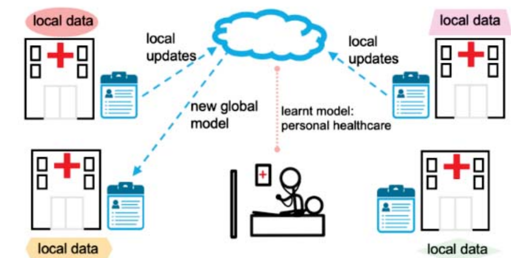


COM4AI

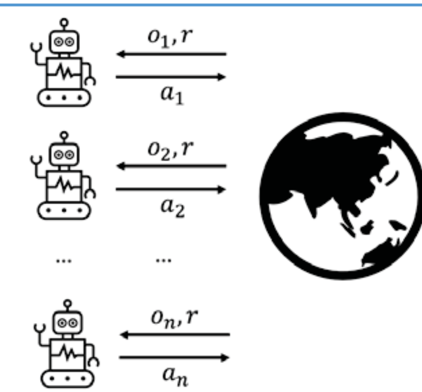
Edge inference
- To resolve limited onboard resources
- Key problem: *what to communicate*?



Privacy-preserving collaborative learning
- To resolve data islands problem
- Key problem: *what to share*?



Cooperative multi-agent system
- To overcome limitation of a single agent
- Key problem: *what, when to communicate*?



Thank you!

- For more details

  https://eejzhang.people.ust.hk/