Task-oriented Communication for Edge AI

From "how to communicate" to "what to communicate"

Jun Zhang

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

Outline

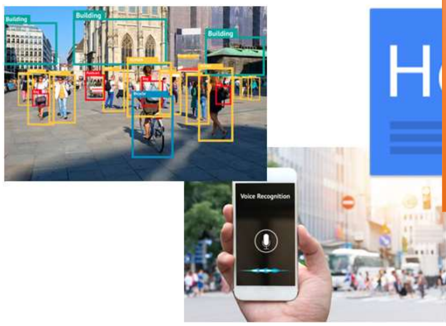
- Background: Task-oriented communication and Edge AI
- Task-oriented communication for edge-assisted inference via information bottleneck (IB)
- Task-oriented communication for cooperative perception via distributed information bottleneck (DIB)
- Conclusions

Task-oriented communication and Edge AI

"When wireless is perfectly applied the whole earth will be converted into a huge brain, which in fact it is, all things being particles of a real and rhythmic whole. We shall be able to communicate with one another instantly, irrespective of distance. Not only this, but through television and telephony we shall see and hear one another as perfectly as though we were face to face, despite intervening distances of thousands of miles; and the instruments through which we shall be able to do this will [fit in a] vest pocket."

https://marionoioso.com/2018/01/08/from-digital-first-to-ai-first/

Mobile Intelligence

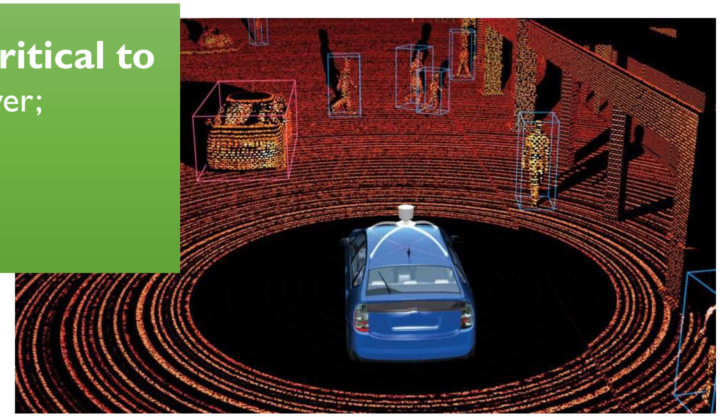


A single device is limited in

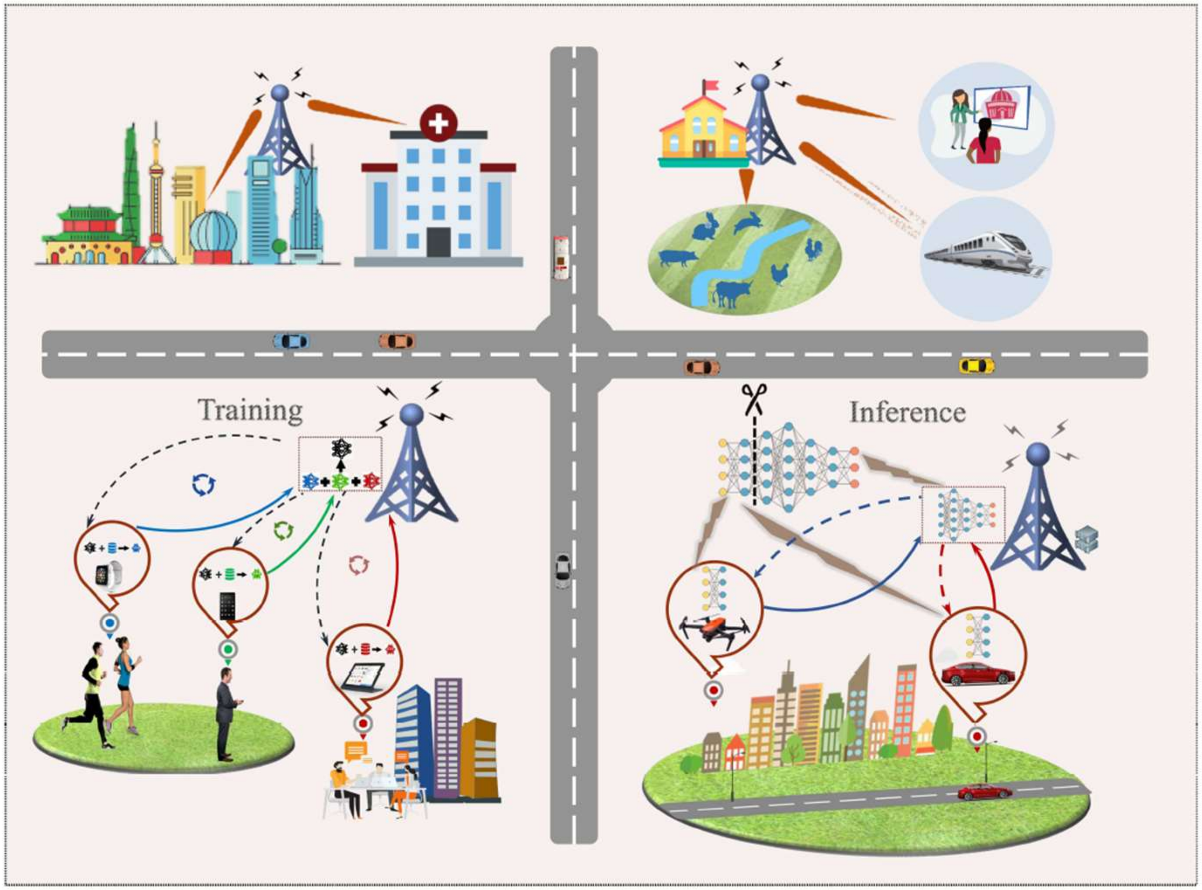
- onboard computing resources;
- limited perception capability;
- limited energy supply.

Effective communication is critical to

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



Edge AI



Y. Shi, K. Yang, T. Jiang, **J. Zhang**, and K. B. Letaief, "Communication-efficient edge AI: Algorithms and systems," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 4, pp. 2167–2191, 4th Quart. 2020.

New communication challenges

- **Enormous volume of data**
  - For example, 4TB sensing data/day for autonomous vehicles
- **Low-latency communication**
  - Millisecond-level latency for safety-critical applications
- **Resource-constrained devices**
  - Limited onboard computation and communication resources

Three levels of communications

Shannon's information theory

Level A
The technical problem

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

Level B
The semantic problem

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

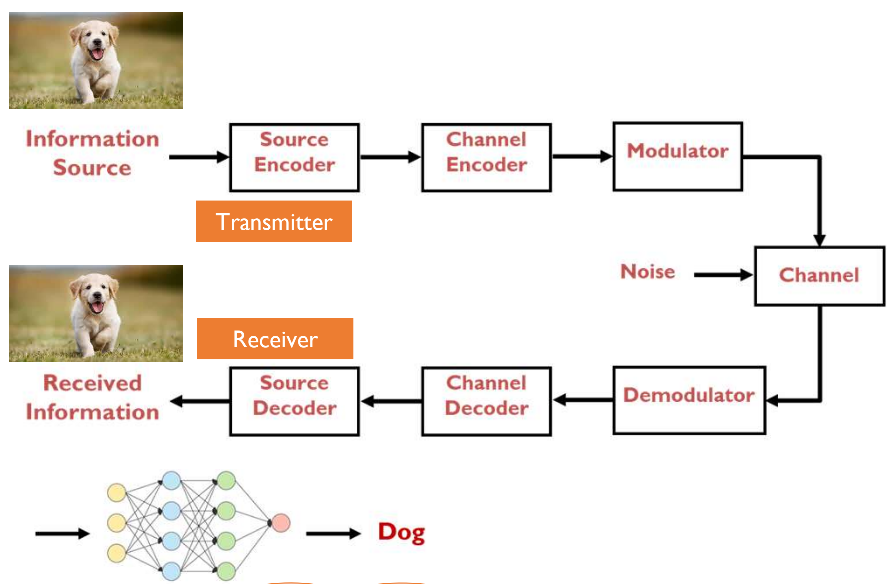
Level C
The effectiveness problem

- 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.

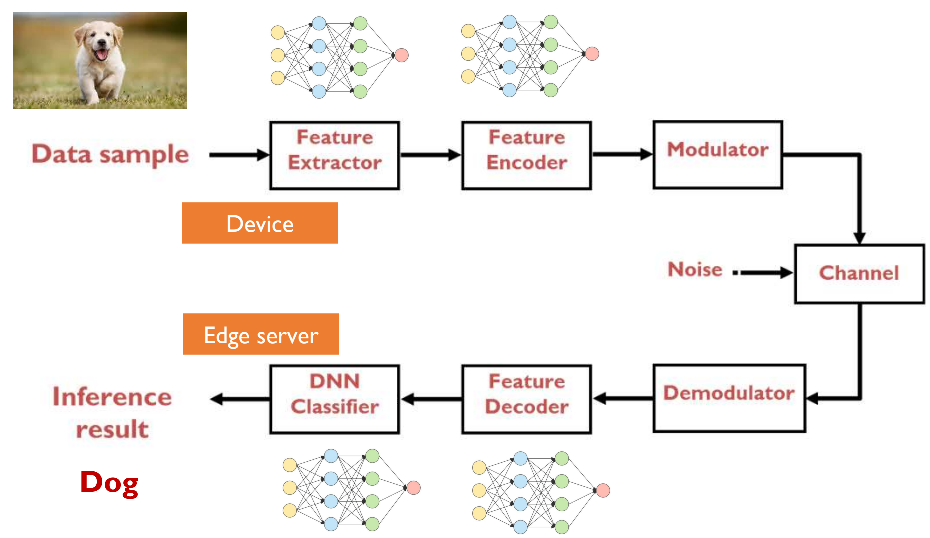
A simplified picture
-- *Data-oriented vs. task-oriented communication*

Data-oriented Communication



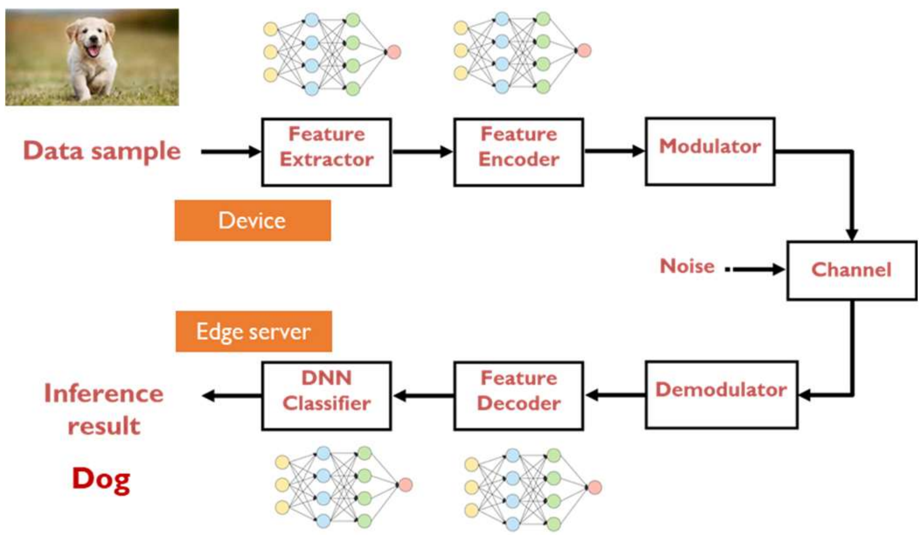
This is how current security camera systems and virtual assistants (e.g., Siri) work!

Task-oriented Communication



Task-oriented communication

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



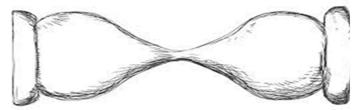
Key design tools
(available only recently)

- Feature encoding via information bottleneck
- End-to-end optimization via deep learning
- Neural architecture optimization for on-device network

Task-oriented communication for edge-assisted inference 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.

The Information Bottleneck (IB) problem
-- An information-theoretical framework for learning



$Y \longrightarrow X \longrightarrow T$

Target **Data** **Representation**

IB strive for *minimality* and *sufficiency* of the latent *T*

- **Minimality**: minimizing amount of information necessary of *X* for the task;
- **Sufficiency**: preserving the information to solve the task (inferring *Y*).

IB problem

$$\begin{aligned} &\inf \quad I(X;T) \\ &\text{subject to:} \quad I(Y;T) \geq \alpha \end{aligned}$$

or

$$\min -I(Y;T) + \beta I(X;T)$$

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

Z. Goldfeld and Y. Polyanskiy, "The information bottleneck problem and its applications in machine learning," IEEE JSAIT, May 2020.

The IB problem

$$\min -I(Y;T) + \beta I(X;T)$$

How well T predicts Y

To promote accuracy

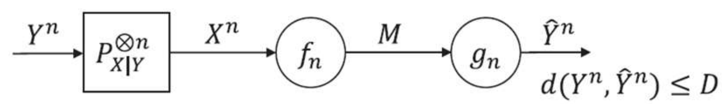
How much T compresses X

To promote generalization

Tradeoff

Preserving "relevant" information vs. finding "compact" representation

- A natural approximate version of **minimal sufficient statistic**.
- Closely related to **remote source coding**.



- Applications of information bottleneck
  - IB theory for deep learning
  - IB as optimization objective (to improve generalization, robustness)

Task-oriented communication vs. Information bottleneck

Task-oriented Commun.

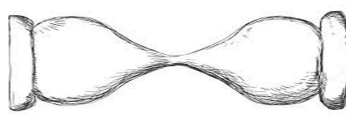
$$X \xrightarrow{\text{Encode}} Z \xrightarrow{\text{Transmission}} \hat{Z} \xrightarrow{\text{Decode}} Y$$

Input data | Transmitted feature | Received feature | Target variable

Dog

Relevance-rate tradeoff

Information Bottleneck



$$X \longrightarrow T \longrightarrow Y$$

Data | **Feature** | **Target**

Task-oriented communication via the IB principle

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

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

How well $\hat{Z}$ predicts Y

To promote accuracy

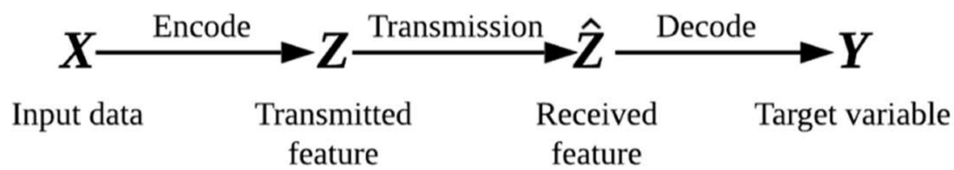
How much $\hat{Z}$ compresses X

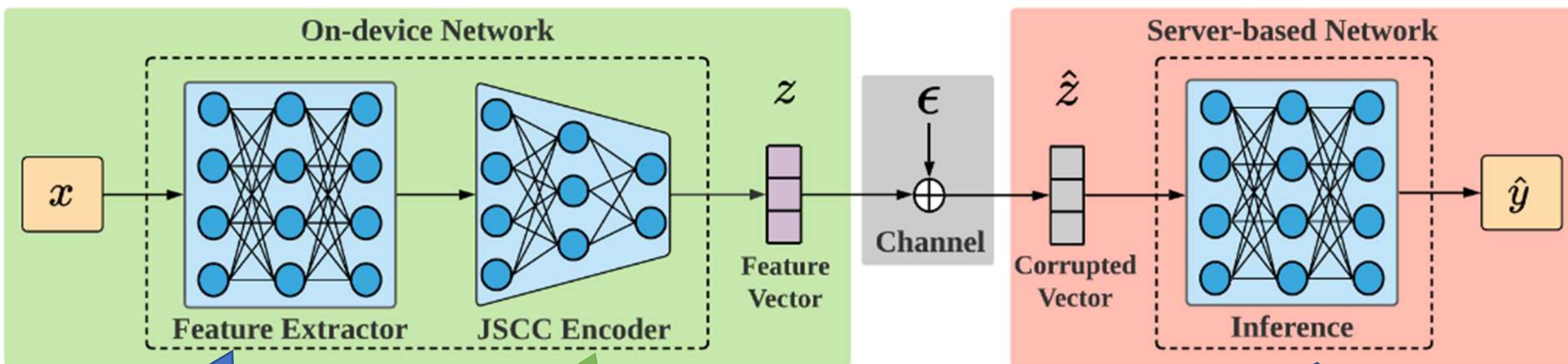
To reduce communication overhead

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

- 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

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

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

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

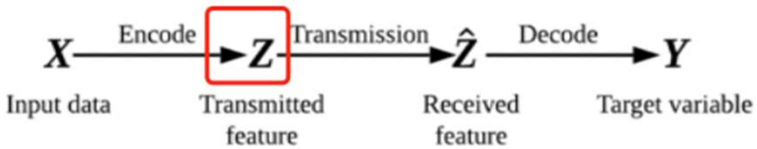
$$\mathcal{L}_{VIB}(\phi,\boldsymbol{\theta}) = \mathbf{E}_{p(\boldsymbol{x},\boldsymbol{y})} \left\{ \mathbf{E}_{p_\phi(\hat{\boldsymbol{z}}|\boldsymbol{x})} \left[-\log q_{\boldsymbol{\theta}}(\boldsymbol{y}|\hat{\boldsymbol{z}}) \right] + \beta D_{KL} \left(p_\phi(\hat{\boldsymbol{z}}|\boldsymbol{x})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

Empirical estimation

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

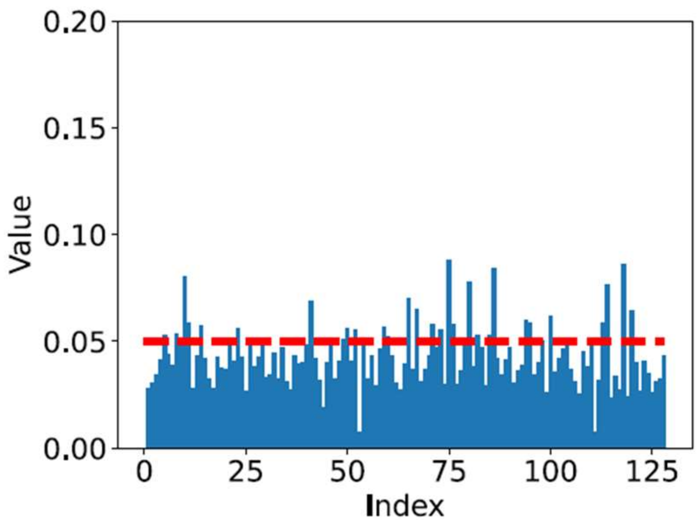
- Variational approximations
  - $p_\phi(\hat{z}|x)$ is defined by the neural network (encoder)
  - $q_\theta(y|\hat{z})$ is a variational distribution to approximate $p(y|\hat{z})$
  - $q(\hat{z})$ is a variational distribution to approximate $p(\hat{z})$

VFE: Output feature sparsification

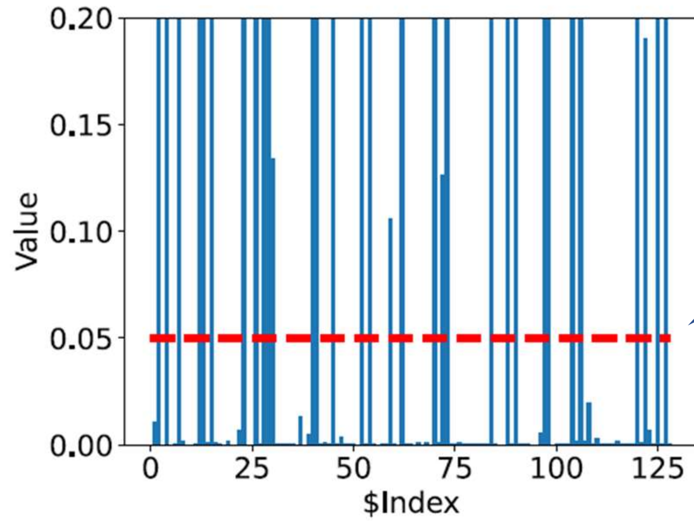


$q(\hat{\boldsymbol{z}})$: log-uniform distribution as the variational prior

$$
\begin{aligned}
-D_{KL}(p_\phi(\hat{z}_i \mid \boldsymbol{x}) \| q(\hat{z}_i)) &= \frac{1}{2}\log\alpha_i - \mathbb{E}_{\epsilon\sim\mathcal{N}(1,\alpha_i)}\log|\epsilon| + \mathrm{C} \\
&\approx k_1 S(k_2 + k_3\log\alpha_i) - 0.5\log\left(1+\alpha_i^{-1}\right) + \mathrm{C}
\end{aligned}
$$



Gaussian prior

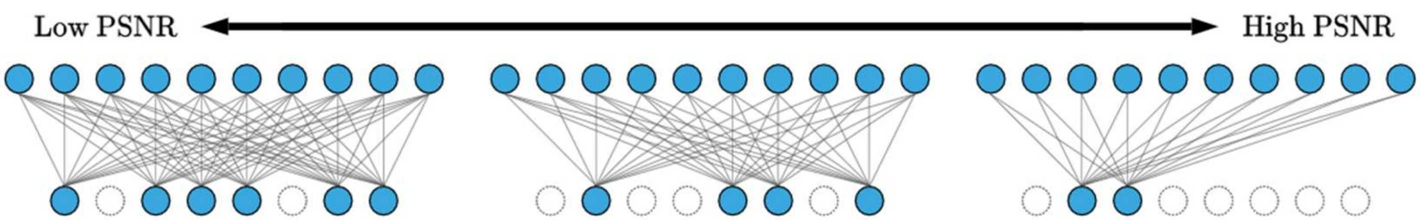


Log-uniform distribution

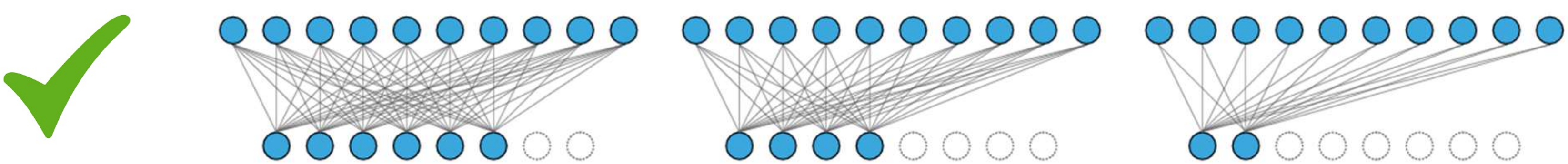
Commun. overhead depends on # active dimensions

Variable-length Variational Feature Encoding (VL-VFE)

- To adapt to channel states: variable-length coding
- To reduce signaling overhead, the coding scheme should be consecutive and monotonic



(a) Random activation



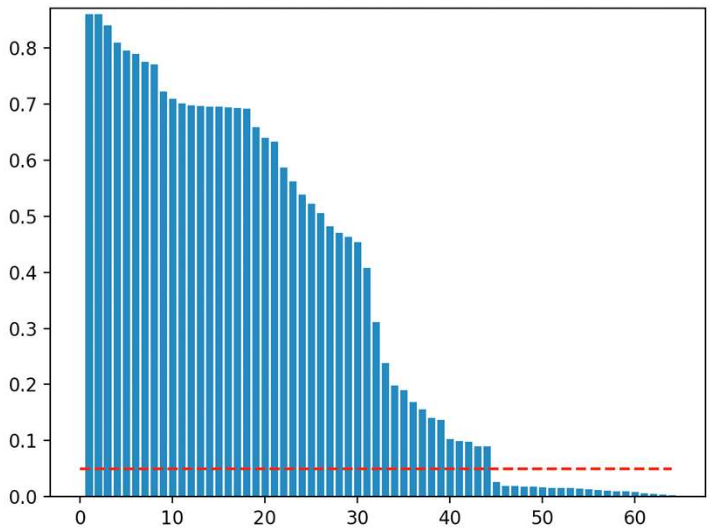
(b) Consecutive activation

VL-VFE: To adapt feature size via selective activation

Dimension importance $\gamma_i(\sigma^2) = \sum_{j=i}^{n} |g_j(\sigma^2)|$ **Soft gate function**

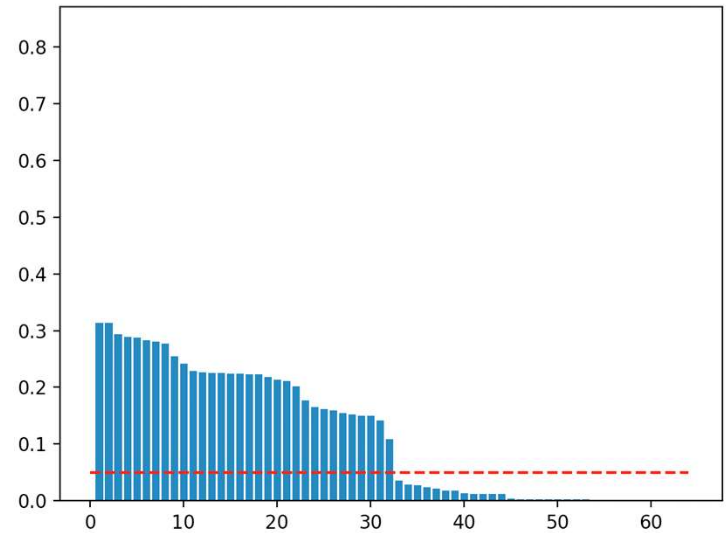
$$\gamma_i(\sigma^2) \geq \gamma_j(\sigma^2), \forall j > i \text{ and } \gamma_i(\sigma^2) \geq \gamma_i(\bar{\sigma}^2), \forall \sigma^2 \geq \bar{\sigma}^2$$

To promote group sparsity



PSNR = 10dB

Activated dimensions: 44



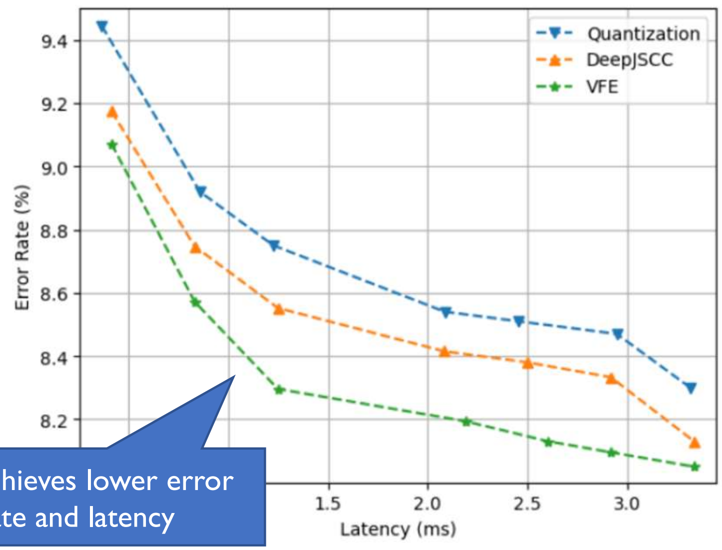
PSNR = 20dB

Activated dimensions: 32

Experiment

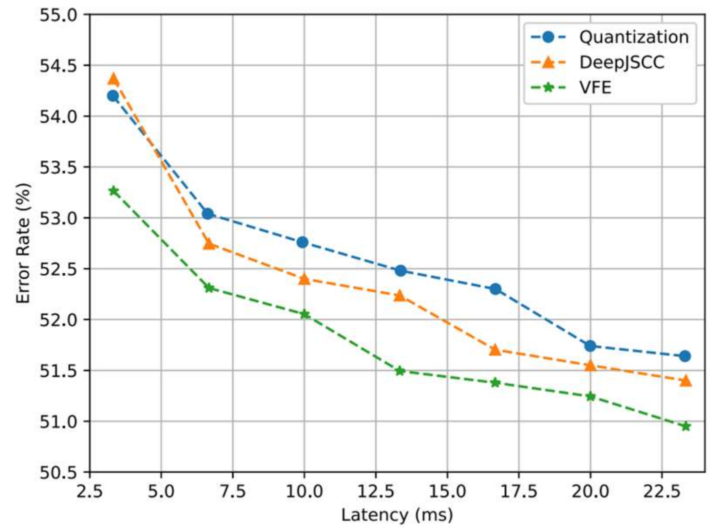
- **Baselines** (data-oriented communication):
  - DeepJSCC (Joint Source-Channel Coding)
  - Learning-based quantization (w/ ideal channel coding)

Rate-distortion on CIFAR-10 dataset



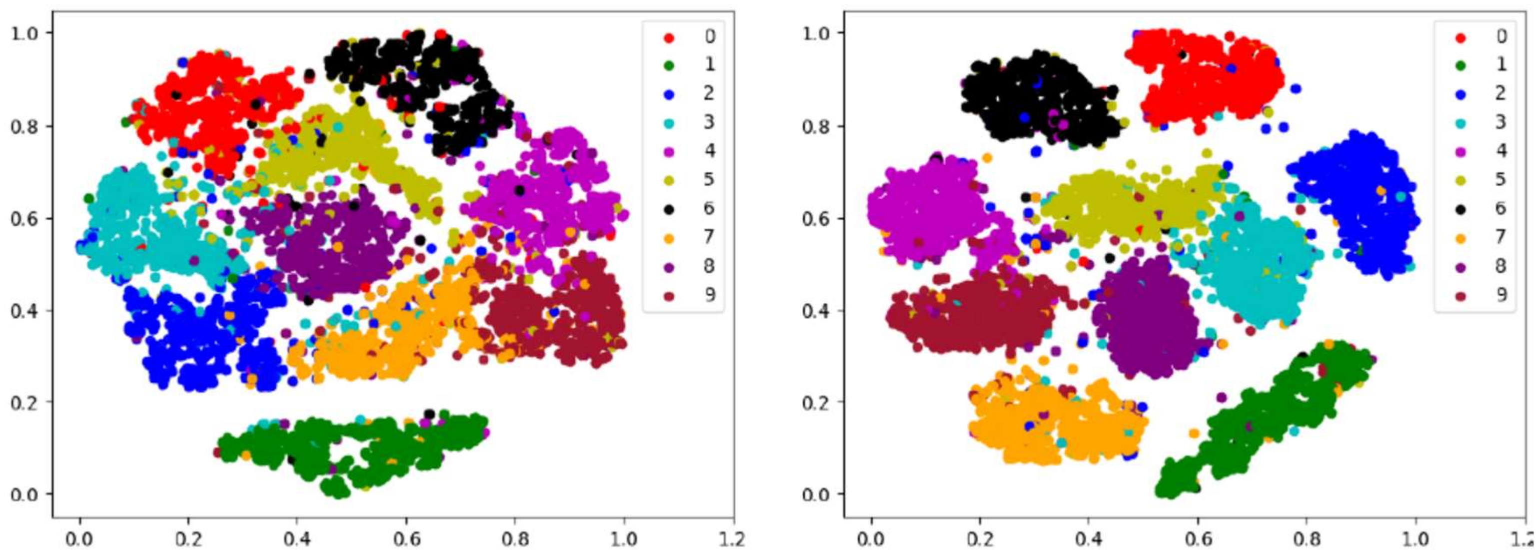
VFE achieves lower error rate and latency

Rate-distortion on Tiny ImageNet dataset



Experiment

- VFE method can better distinguish the data from different classes compared with DeepJSCC.



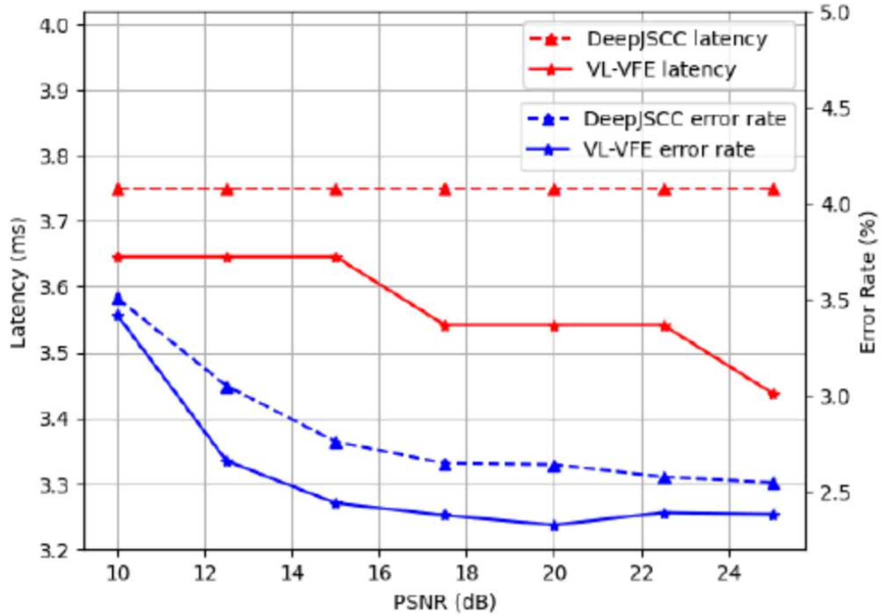
(a) DeepJSCC: Accuracy = 96.77%, dimension n = 24.

(b) Proposed VFE: Accuracy = 97.39%, dimension n = 24.

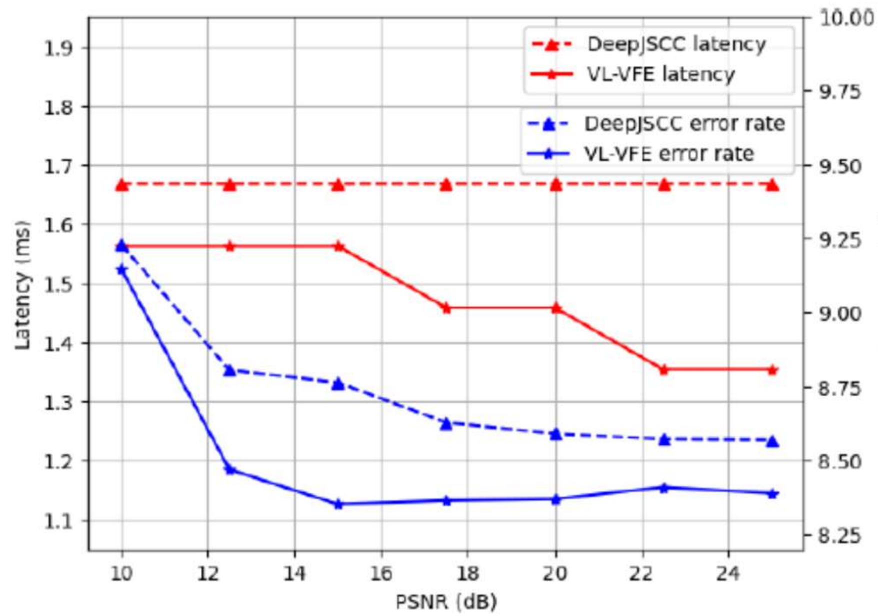
2-dimensional t-SNE embedding of the received feature in the MNIST classification task with PSNR = 20 dB.

Experiment: VL-VFE

- VL-VFE achieves higher accuracy and lower latency compared with DeepJSCC in dynamic channel conditions.



(a) The MNIST classification task



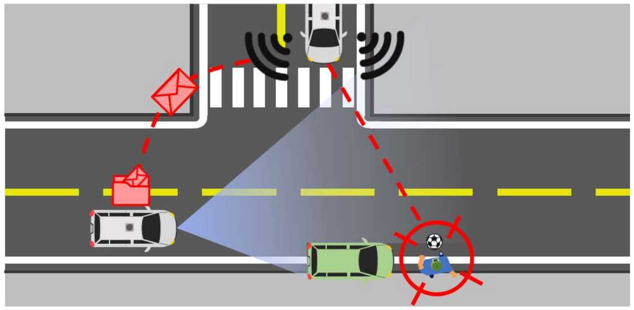
(b) The CIFAR-10 classification task

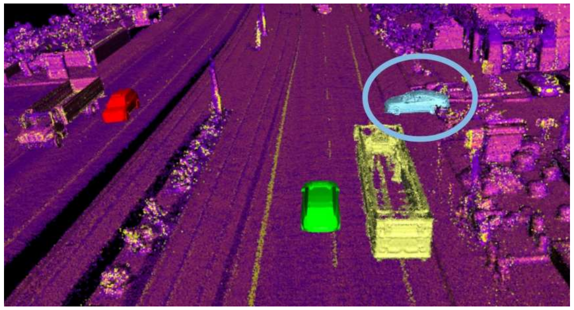
Task-oriented communication for cooperative inference via distributed information bottleneck

J. Shao, Y. Mao, and **J. Zhang**, "Task-oriented communication for multi-device cooperative edge inference," submitted to IEEE Transactions on Wireless Communications.
https://arxiv.org/abs/2109.00172

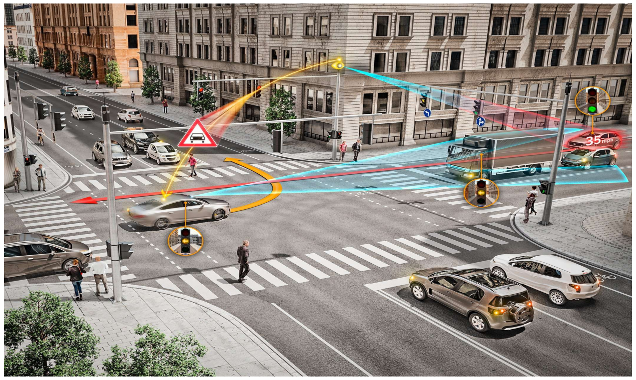
Cooperative perception

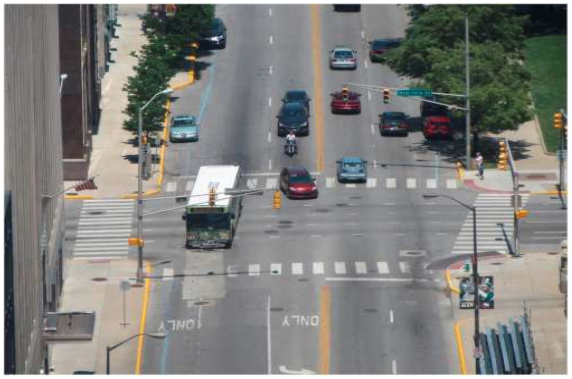
- Cooperative localization, detection, tracking, map generation





Occlusion



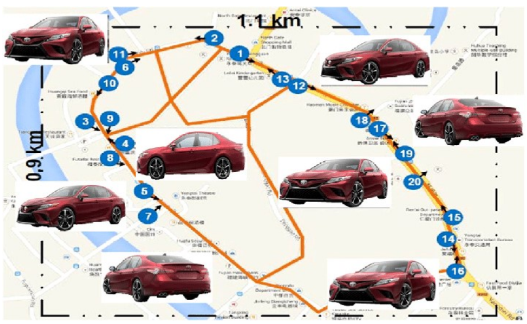


Intersection

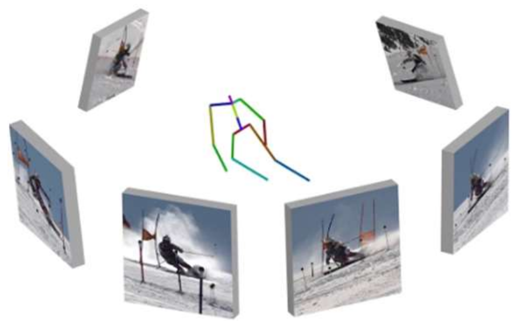
Caillot, Antoine, et al. "Survey on Cooperative Perception in an Automotive Context." *IEEE Transactions on Intelligent Transportation Systems* (2022).

Multi-camera cooperative inference

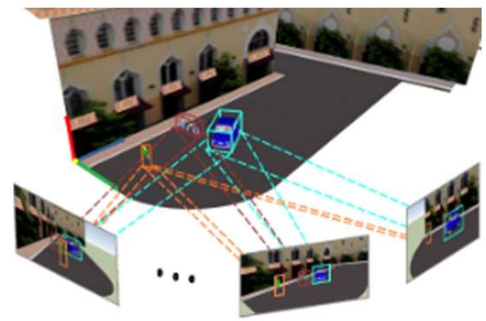
- Cooperation among multiple cameras with distinct views improves sensing capability.



Vehicle Re-identification



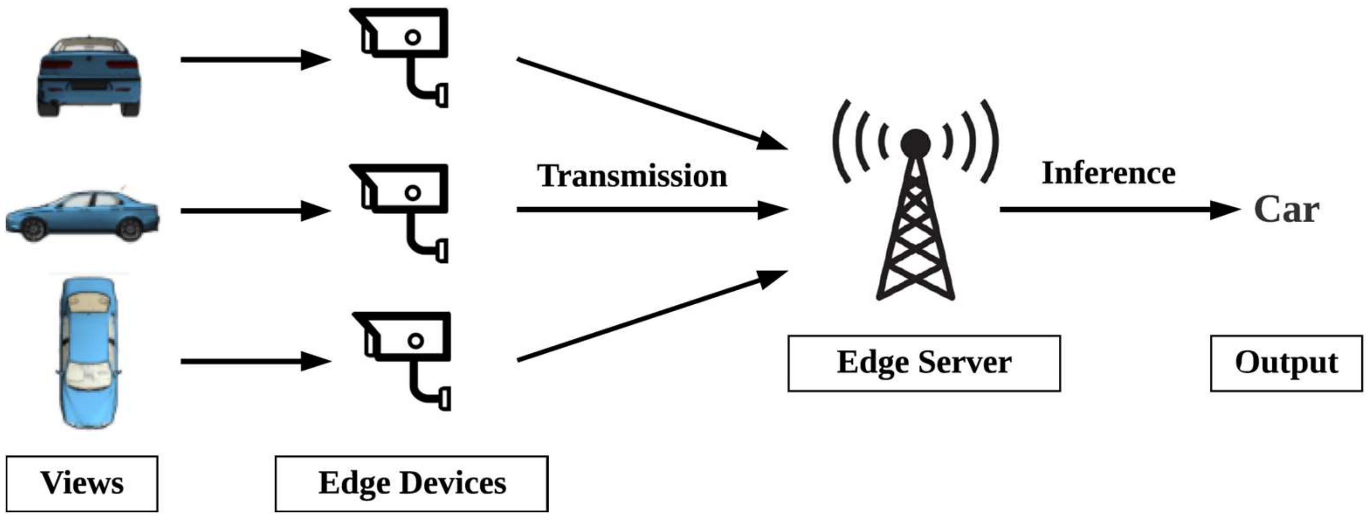
Pose Estimation



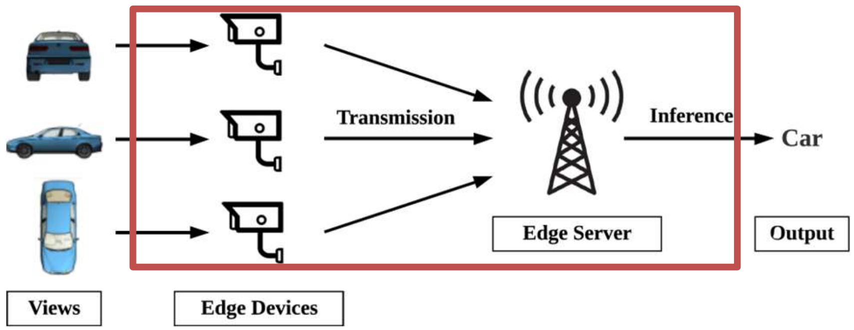
3D Localization

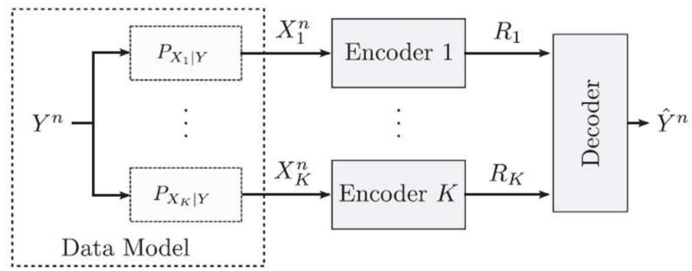
Multi-camera cooperative inference

- 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

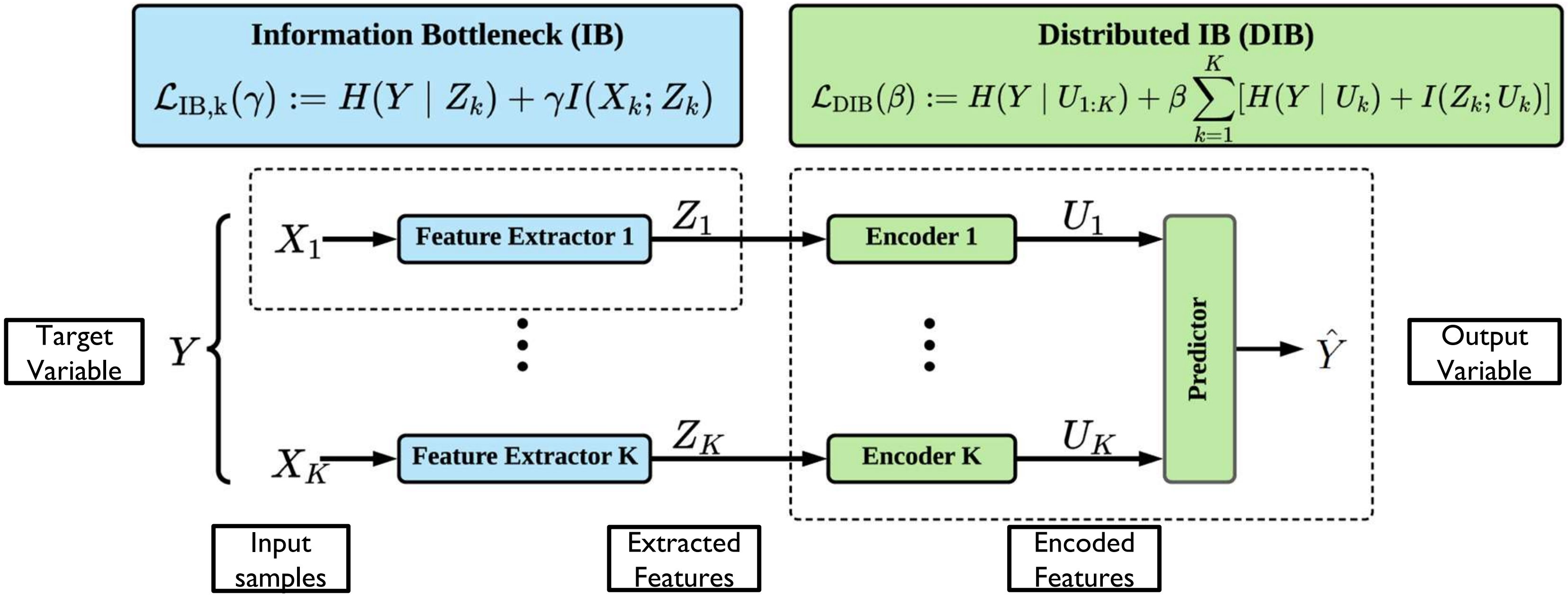
Rate Relevance

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

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



Task-relevant feature extraction via IB

$$Y \leftrightarrow X_k \leftrightarrow Z_k$$

Information Bottleneck (IB)

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

- Ideally, if extracted features are minimal and sufficient:

$$\underbrace{I(X_k; Z_k) = \overbrace{I(Y; X_k)}^{\textbf{Sufficiency}}}_{\textbf{Minimality}} \overbrace{= I(Y; Z_k)}, \;\; k \in \{1, \dots, K\}.$$

$$p(Z_{1:K}|Y) = \prod_{k=1}^{K} p(Z_k|Y)$$

Conditional independence

DIB theorem: optimal rate-relevance tradeoff

Distributed feature encoding via DIB

- Rate-relevance tradeoff via the DIB objective

Proposition 1. *(Distributed Information Bottleneck [16]) Suppose the extracted features Z_k for $k \in \{1, \ldots, K\}$ are conditionally independent given the target variable Y. Each (Δ_β, R_β) with $\beta \geq 0$ is an optimal rate-relevance tuple, i.e., there exists no relevance $\Delta \geq \Delta_\beta$ given the sum rate constraint $R_{\text{sum}} = R_\beta$, where*

$$\Delta_\beta = I(Y; U^*_{1:K}), \quad R_\beta = \Delta_\beta + \sum_{k=1}^{K} [I(Z_k; U^*_k) - I(Y; U^*_k)], \tag{7}$$

and the encoded features $U^*_{1:K}$ are obtained by minimizing the following distributed information bottleneck (DIB) objective:

$$\min_{\{p(\boldsymbol{u}_k|\boldsymbol{z}_k)\}_{k=1}^K} \boxed{\mathcal{L}_{\text{DIB}}(\beta) := H\left(Y|U_{1:K}\right) + \beta \sum_{k=1}^{K} \left[H\left(Y|U_k\right) + I\left(Z_k; U_k\right)\right].} \tag{8}$$

- Main design challenges:
  - How to effectively control communication overhead?
  - How to estimate mutual information?
  - How to compensate the performance loss due to approximations?

Distributed Deterministic Information Bottleneck (DDIB)

- **DIB objective**

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

The minimality is only satisfied in the asymptotic limit

- **DDIB objective**

$$\mathcal{L}_{\mathrm{DDIB}}(\beta) := H(Y \mid U_{1:K}) + \beta \sum_{k=1}^{K} [H(Y \mid U_k) + R_{\mathrm{bit}}(U_k)]$$

Enable fine control of communication overhead, and instantaneous edge inference for each input sample

Proposed method: Variational DDIB (VDDIB)

- Using variational inference to estimate the intractable (entropy) terms.

$$\mathcal{L}_{\mathrm{DDIB}}(\beta) := H(Y \mid U_{1:K}) + \beta \sum_{k=1}^{K} [H(Y \mid U_k) + R_{\mathrm{bit}}(U_k)]$$

$$\Downarrow$$

$$\begin{aligned}\mathcal{L}_{\mathrm{VDDIB}}(\beta; \boldsymbol{\phi}, \boldsymbol{\psi}) := \mathbf{E}_{p_\theta(\boldsymbol{z}_{1:K}, \boldsymbol{y})} \Big\{ & - \log p_{\psi_0}(\boldsymbol{y} \mid \boldsymbol{u}_{1:K}) \\ & + \beta \Big\{ \sum_{k=1}^{K} - \log p_{\psi_k}(\boldsymbol{y} \mid \boldsymbol{u}_k) + \sum_{k=1}^{K} R_{\mathrm{bit}}(\boldsymbol{u}_k) \Big\} \Big\}\end{aligned}$$

<u>Variational distributions:</u>

$$p_{\psi_0}(\boldsymbol{y}|\boldsymbol{u}_{1:K}) \propto \exp(-\ell(\boldsymbol{y}, \hat{\boldsymbol{y}}(\boldsymbol{u}_{1:K}; \psi_0))),$$

$$p_{\psi_k}(\boldsymbol{y}|\boldsymbol{u}_k) \propto \exp(-\ell(\boldsymbol{y}, \hat{\boldsymbol{y}}(\boldsymbol{u}_k; \psi_k))), k \in \{1, \ldots, K\}$$

$$\mathcal{L}_{\mathrm{DIB}}(\beta) \leq \mathcal{L}_{\mathrm{DDIB}}(\beta) \leq \mathcal{L}_{\mathrm{VDDIB}}(\beta; \boldsymbol{\phi}, \boldsymbol{\psi}).$$

Minimizing the VDDIB objective may not result in the optimal rate-relevance tradeoff due to the approximations

Introduce a selective retransmission (SR) mechanism to further reduce the communication overhead caused by the redundancy among the extracted features.

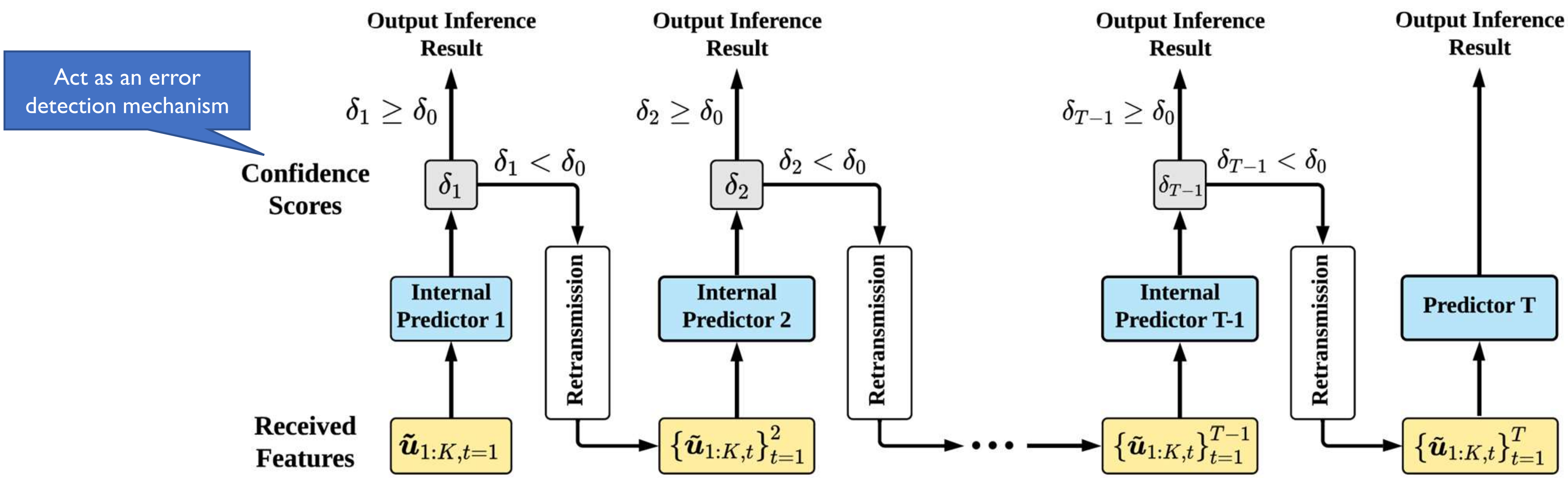
Selective retransmission mechanism

- Retransmission mechanisms
  - Error detection + retransmission requests
  - E.g., ARQ, HARQ

- Selective retransmission
  - The edge server selectively activates the edge devices to retransmit their encoded features based on the informativeness of the received features.
  - The mechanism consists of a stopping policy and an attention module.

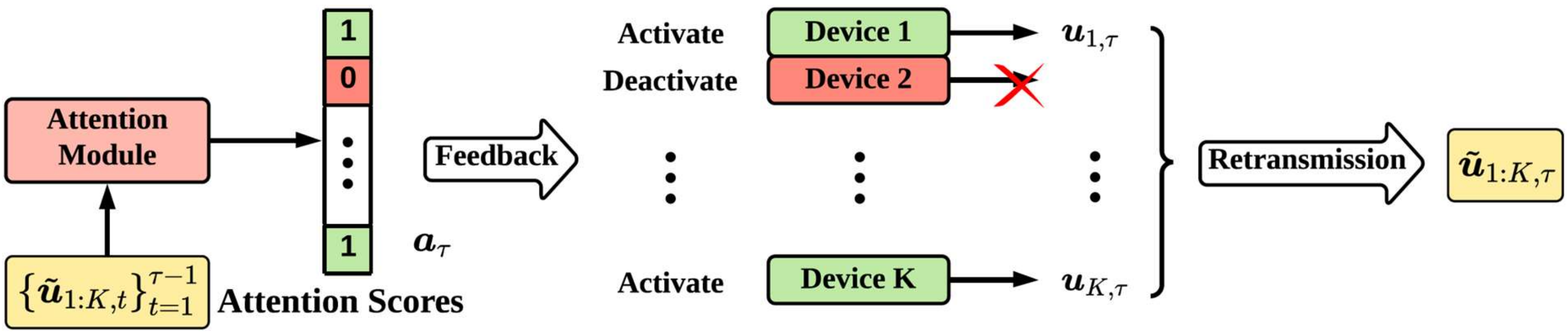
Selective retransmission mechanism

- Stopping policy
  - Each edge device is allowed to transmit the encoded feature with a maximum number of *T* attempts.
  - Once the received features are sufficient to output a confident result, the remaining retransmission attempts can be saved.



Selective retransmission mechanism

- Attention module
  - Select the most informative features to retransmit based on the attention scores.



VDDIB with Selective Retransmission Mechanism (VDDIB-SR)

- **VDDIB-SR loss function**

$$
\begin{aligned}
\mathcal{L}_{\mathrm{VDDIB}}(\beta;\boldsymbol{\phi},\boldsymbol{\psi}) :=\mathbf{E}_{p_\theta(\boldsymbol{z}_{1:K},\boldsymbol{y})}\Big\{&-\log p_{\psi_0}(\boldsymbol{y}\mid\boldsymbol{u}_{1:K}) \\
&+\beta\left\{\sum_{k=1}^{K}-\log p_{\psi_k}(\boldsymbol{y}\mid\boldsymbol{u}_k)+\sum_{k=1}^{K}R_{\mathrm{bit}}(\boldsymbol{u}_k)\right\}\Big\}
\end{aligned}
$$

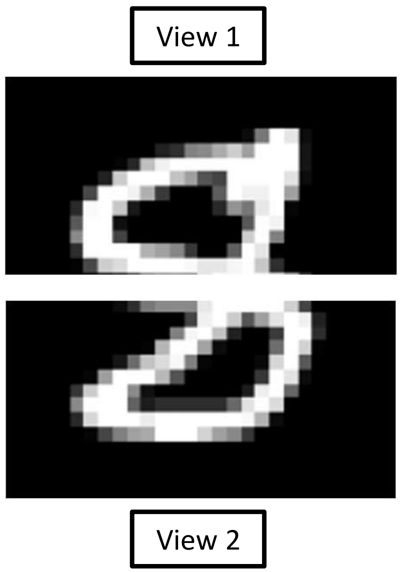
Account for T predictors

Communication cost by the SR mechanism

$$
\begin{aligned}
\mathcal{L}_{\mathrm{VDDIB-SR}}\left(\beta,T;\tilde{\boldsymbol{\phi}},\tilde{\boldsymbol{\psi}},\{\boldsymbol{\psi}_k\}_{k=1}^{K}\right) :=\mathrm{E}_{p_\theta(\boldsymbol{z}_{1:K},y)}\Bigg\{&\frac{1}{T}\sum_{\tau=1}^{T}-\log p_{\tilde{\psi}_\tau}\left(\boldsymbol{y}\mid\{\tilde{\boldsymbol{u}}_{1:K,t}\}_{t=1}^{\tau}\right) \\
&+\beta\left\{\sum_{k=1}^{K}-\log p_{\psi_k}(\boldsymbol{y}\mid\boldsymbol{u}_k)+\sum_{k=1}^{K}\sum_{t=1}^{T}R_{\mathrm{bit}}(\tilde{\boldsymbol{u}}_{k,t})\right\}\Bigg\}
\end{aligned}
$$

Performance evaluation

- Cooperative inference tasks



Two-view MNIST classification

Twelve-view Shape Recognition on ModelNet40 dataset

Performance evaluation

- The accuracy of the cooperative tasks under different bit constraints.
- Data-oriented communication leads to
  - 1.3 kbits overhead with 98.6% accuracy in the MNIST classification task.
  - 120 KB overhead with 92% accuracy in the shape recognition task.

MNIST 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%**

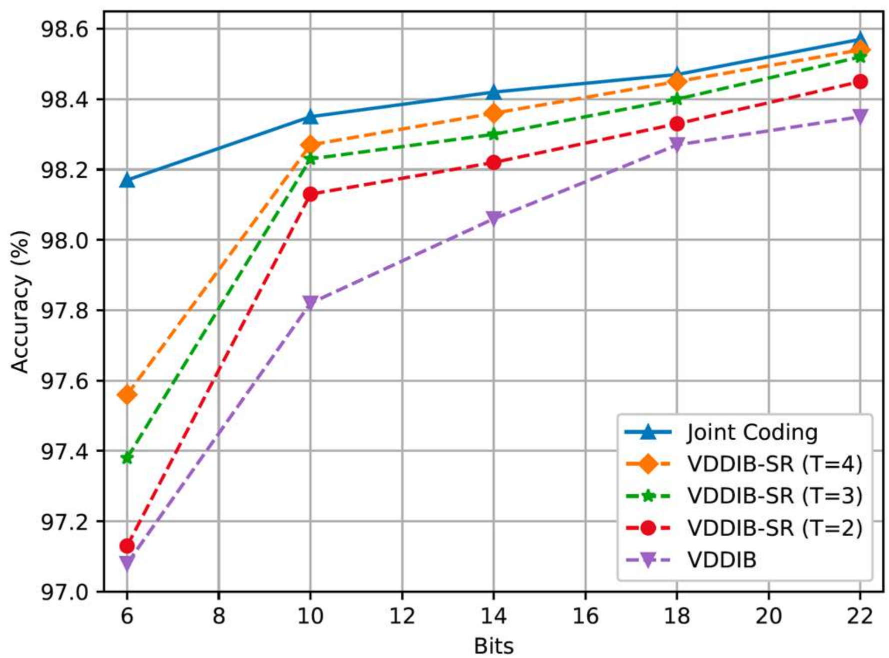
Shape Recognition

	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.

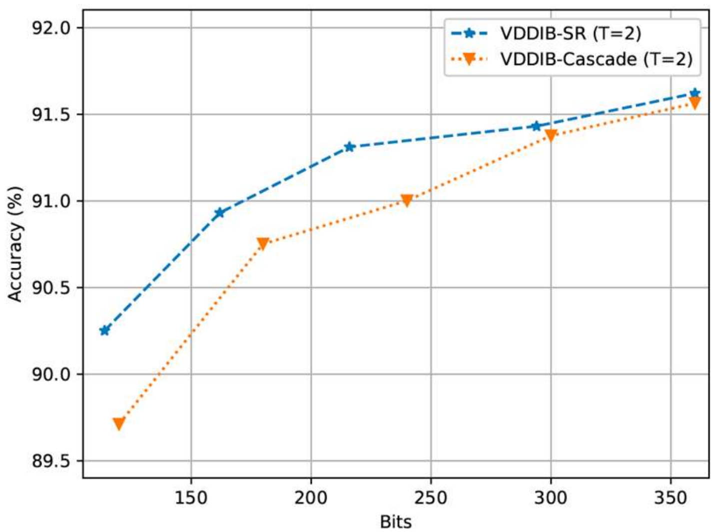
Ablation study

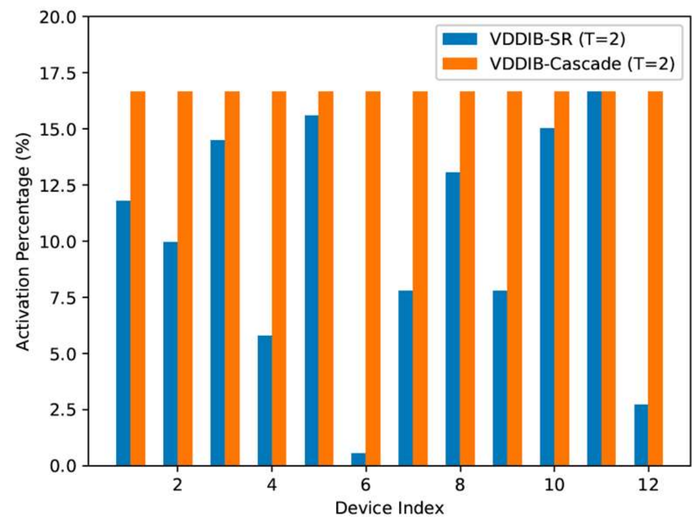
- Impact of the maximum transmission attempts *T.*
  - The performance of the VDDIB-SR method improves *with T.*



Ablation study

- Impact of the attention module
  - We propose a baseline method that removes the attention module denoted as VDDIB-Cascade for comparison.





	accuracy	bits
VDDIB-SR	91.25%	216
VDDIB-Cascade	90.88%	240

Conclusions

Conclusions

- Task-oriented communication
  - Shift from "how to transmit" to "what to transmit"
- Task-oriented communication for Edge AI
  - Edge-assisted inference via information bottleneck
  - Cooperative perception via distributed information bottleneck
- Information theory is still our guide
  - Rate-distortion theory
  - Distributed source coding theory

References

- J. Shao, **J. Zhang**, "BottleNet++: An end-to-end approach for feature compression in device-edge co-inference systems," *IEEE Int. Conf. Commun. (ICC) Workshop on Edge Machine Learning for 5G Mobile Networks and Beyond*, Jun. 2020.
- J. Shao, **J. Zhang**, "Communication-computation trade-off in resource-constrained edge inference," *IEEE Commun. Mag.*, Dec 2020.
- J. Shao, H. Zhang, Y, Mao, and **J. Zhang**, "Branchy-GNN: a device-edge co-inference framework for efficient point cloud processing," *ICASSP 2021*.
- X. Zhang, J. Shao, Y. Mao, and **J. Zhang**, "Communication-computation efficient device-edge co-inference via AutoML," *IEEE GLOBECOM 2021*.
- 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," submitted to *IEEE Trans. Wireless Communications*. (https://arxiv.org/abs/2109.00172)

Thank you!

- For more details

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