Semantic Communication for the Internet of Things

Frameworks, Optimization, and Open Challenges

Samer Lahoud
Faculty of Computer Science, Dalhousie University, Canada
IEEE Senior Member

DALHOUSIE UNIVERSITY

Speaker

Samer Lahoud

Associate Professor and University Research Chair

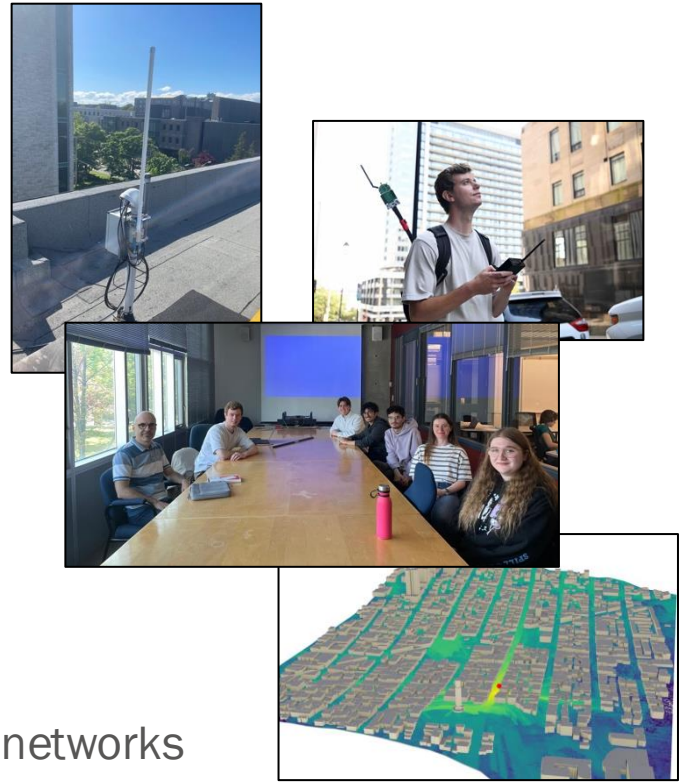
Head of the Systems Cluster

Faculty of Computer Science, Dalhousie University

Research area

Reliable, secure, and resource-efficient wireless networks

SPARK Lab



- Core directions
  - Secure and resilient wireless communications
  - Spectrum-efficient networking and coexistence
  - Interference detection and mitigation
  - Reliable connectivity for constrained IoT systems
  - AI-native and semantic communications
- Selected project directions
  - Seamless transitions between mobile broadband networks
  - IoT threat intelligence across heterogeneous wireless technologies
  - Low-power long-range wireless coexistence in unlicensed spectrum

More information on https://samer.lahoud.fr

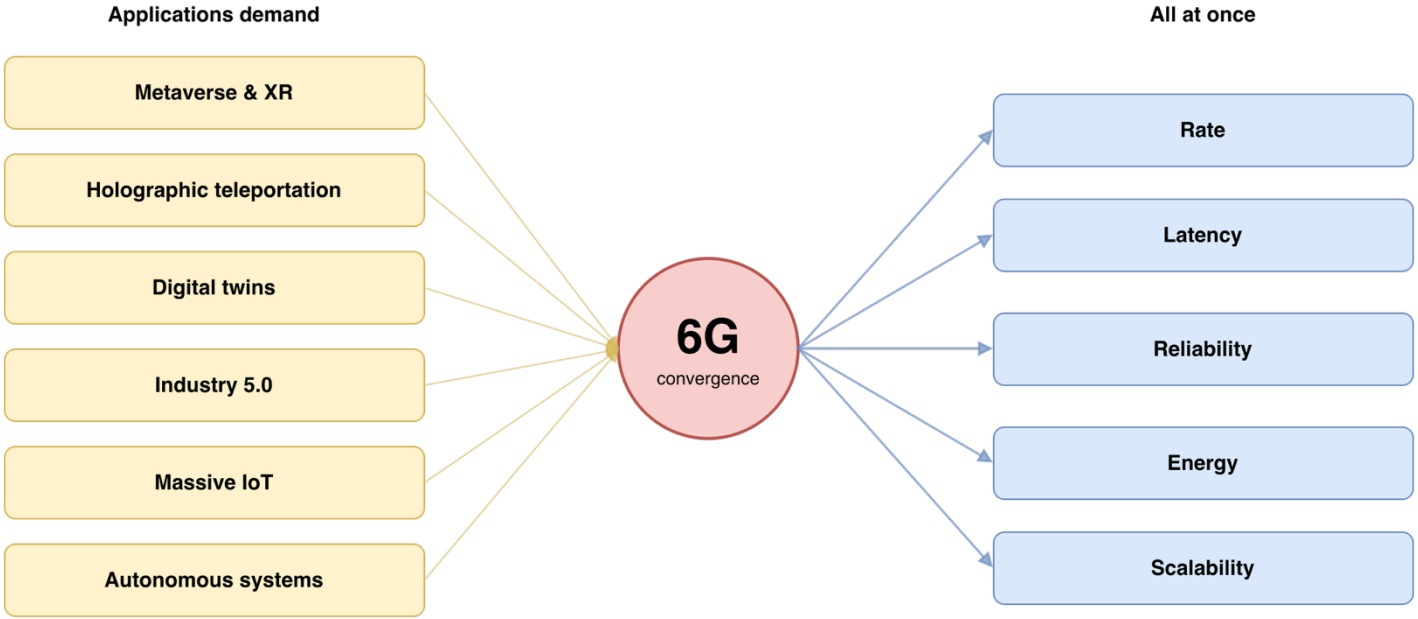
Today’s trajectory

- From bit reliability to semantic effectiveness
- From semantic representation to Deep JSCC
- From centralized models to wireless IoT deployment
- From bit-centric resource allocation to semantic-aware control
- From research prototypes to open challenges

Introduction

From bit reliability to semantic effectiveness

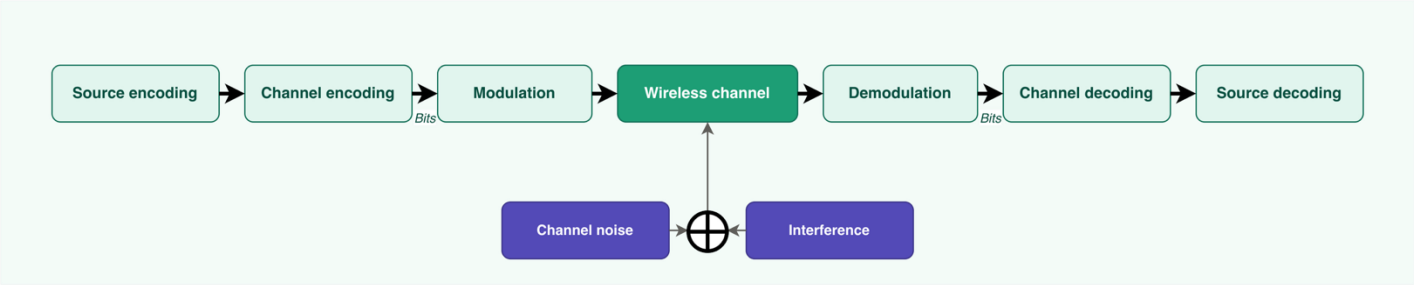
The 6G pressure point



- 6G services combine communication, computation, sensing, control, and intelligence.
- Applications demand rate, latency, reliability, energy efficiency, and scalability at the same time.

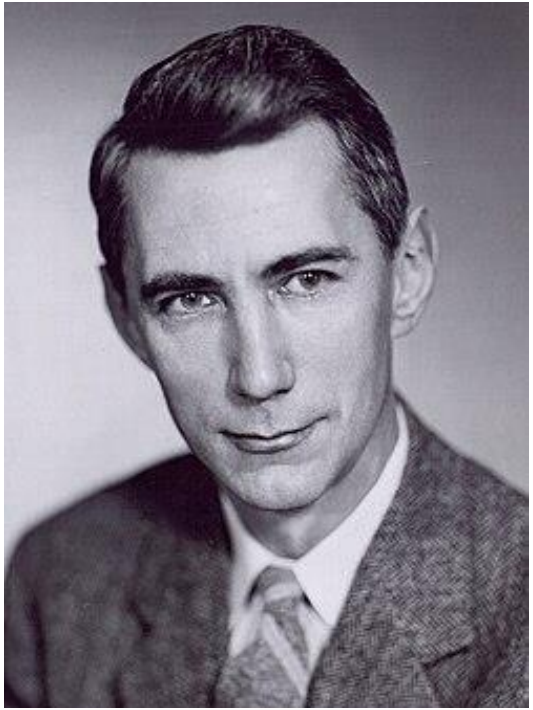
The classical communication model

- How can symbols be transmitted reliably over a noisy channel?
- Classical communication is mainly a reconstruction problem:
  - compress, transmit, recover.
- Encoding captures statistical uncertainty.



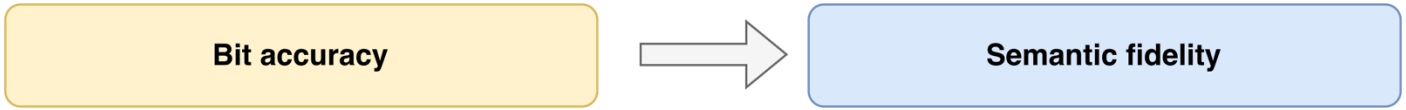
Shannon's 1948 framing

- Source and channel coding can be designed independently under asymptotic conditions.
- This principle drove decades of successful layered design.
- Shannon deliberately excluded semantic aspects from the engineering problem.



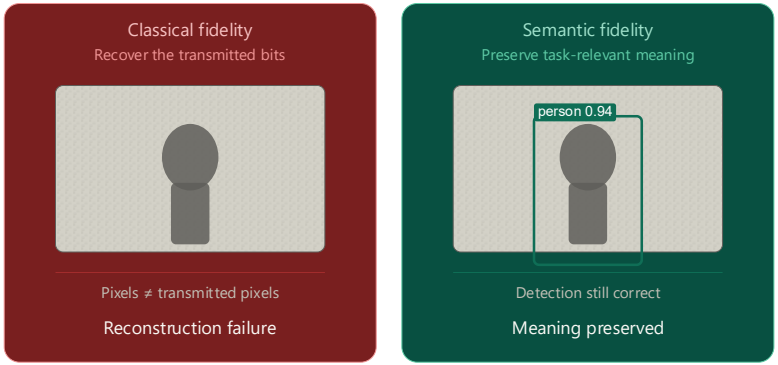
Beyond incremental wireless gains

- More spectrum, antennas, coding, and optimization remain essential.
- Future services also require a new question:
  - What information is relevant to the receiver’s objective?
  - This question opens the path to semantic communication.



Semantic fidelity can survive signal distortion

- Classical fidelity: recover the transmitted bits, symbols, or pixels.
- Semantic fidelity: preserve what is needed for interpretation, inference, or action.
- Example: a wireless surveillance camera may lose fine visual detail while still enabling correct detection of an intruder
- Signal distortion is not semantic failure!
  - Semantic failure occurs only when task-relevant meaning is lost.



SemCom in one sentence

- **Semantic communication (SemCom) conveys intended meaning rather than every source bit.**
- It extracts and transmits information that is:
  - meaningful,
  - task-relevant,
  - interpretable at the receiver.
- The objective becomes semantic fidelity, not only bit fidelity.

Weaver's three levels of communication

- Weaver separated communication into three levels:
  - Level A: technical accuracy
    - How accurately can symbols be transmitted?
  - Level B: semantic meaning
    - How precisely do transmitted symbols convey the desired meaning?
  - Level C: effectiveness or conduct
    - How effectively does the received meaning influence conduct?

Why Level A dominated communication design

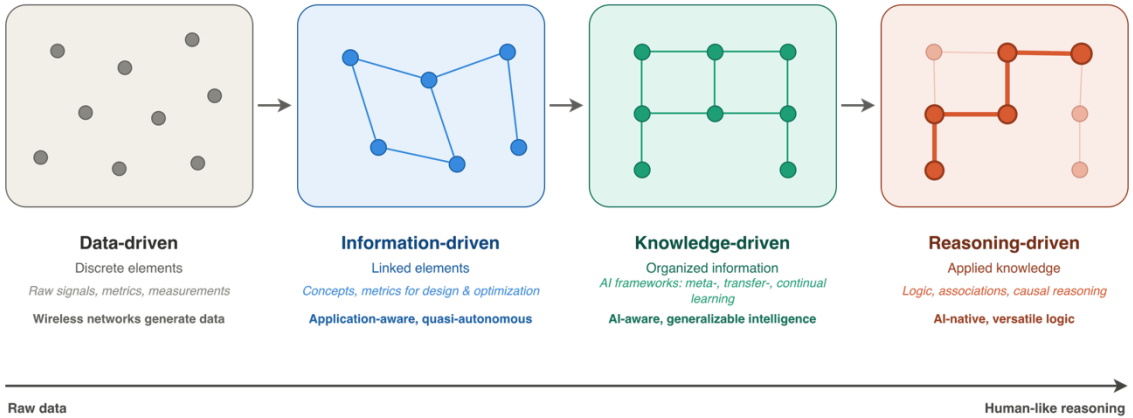
- Meaning was difficult to formalize mathematically.
- The separation principle worked extremely well.
- Layered architectures were modular, scalable, and standardizable.

Why the idea returned

- AI made SemCom practically plausible.
- Modern tools can support:
  - semantic extraction,
  - semantic representation,
  - semantic reconstruction,
  - reasoning with knowledge bases.
- This reopens Weaver's semantic and effectiveness levels.

But AI-for-wireless is not enough

- AI can optimize existing wireless blocks
  - Channel estimation, symbol detection, resource management, etc.
- SemCom requires reasoning over meaning, context, and structure.

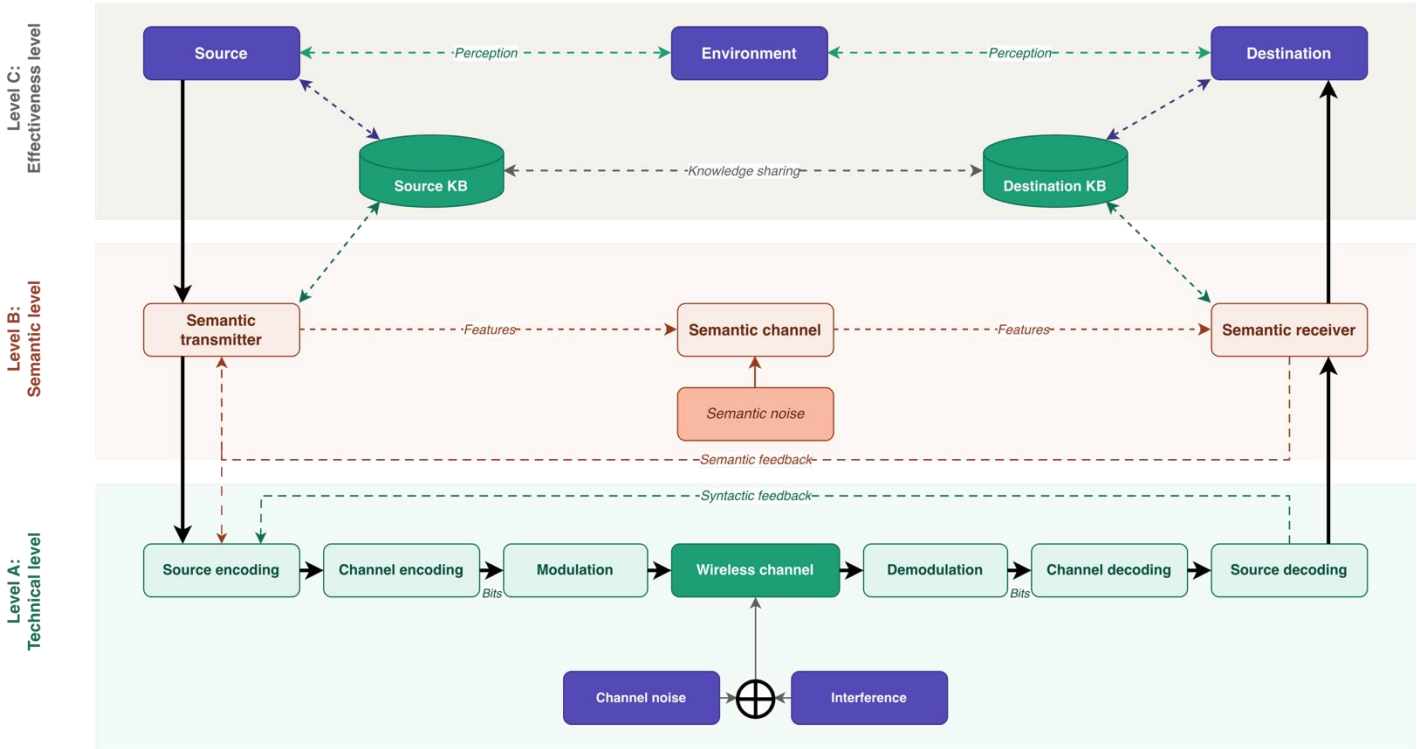


C. Chaccour, W. Saad, M. Debbah, Z. Han and H. Vincent Poor, "Less Data, More Knowledge: Building Next-Generation Semantic Communication Networks," in IEEE Communications Surveys & Tutorials, vol. 27, no. 1, pp. 37-76, Feb. 2025

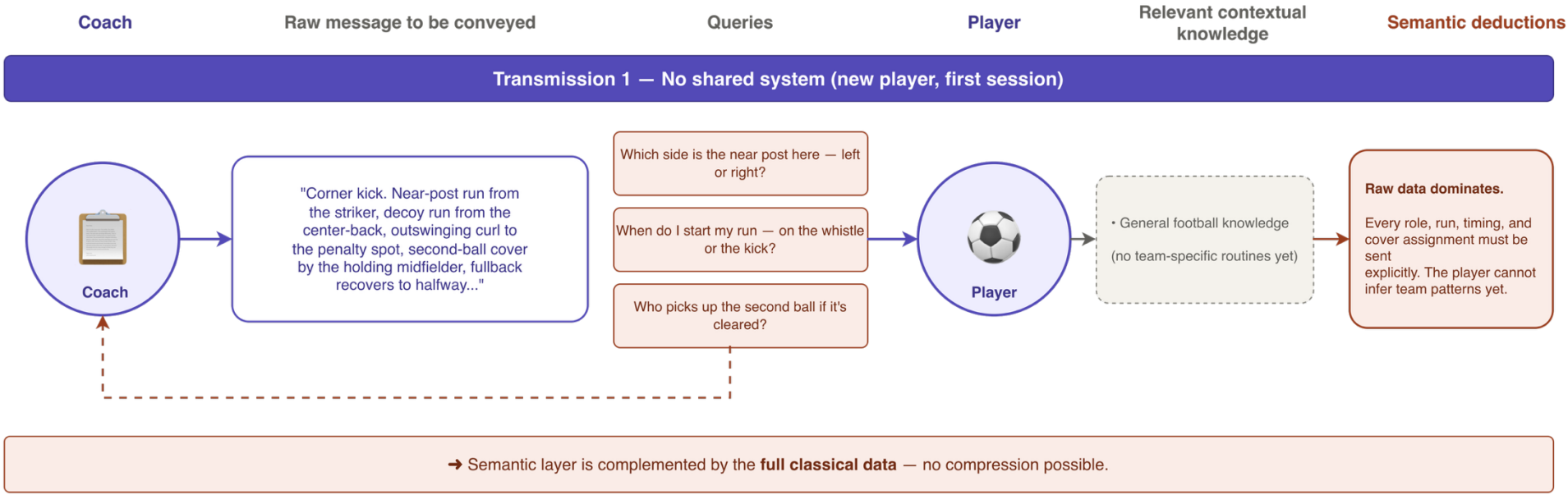
Meaning depends on shared knowledge

- Meaning is not contained in bits alone.
- It depends on:
  - prior knowledge,
  - context,
  - memory of previous exchanges,
  - receiver capabilities,
  - intended action.
- In SemCom, the knowledge base supports semantic extraction, interpretation, and reconstruction.

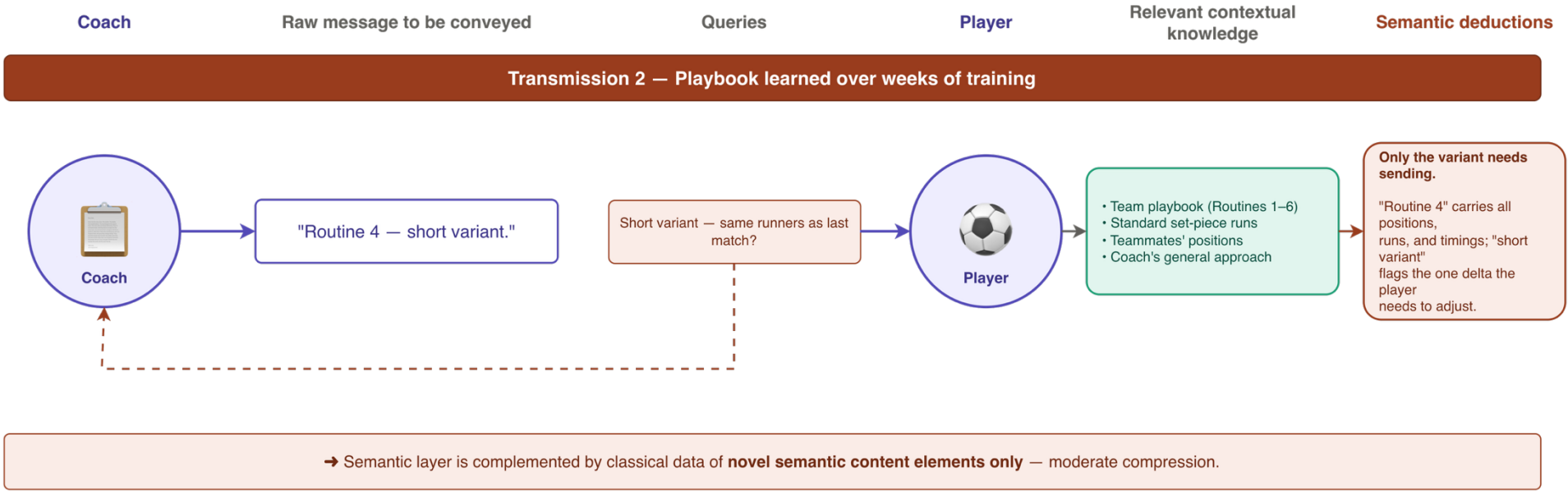
A SemCom architecture



Coach/player example

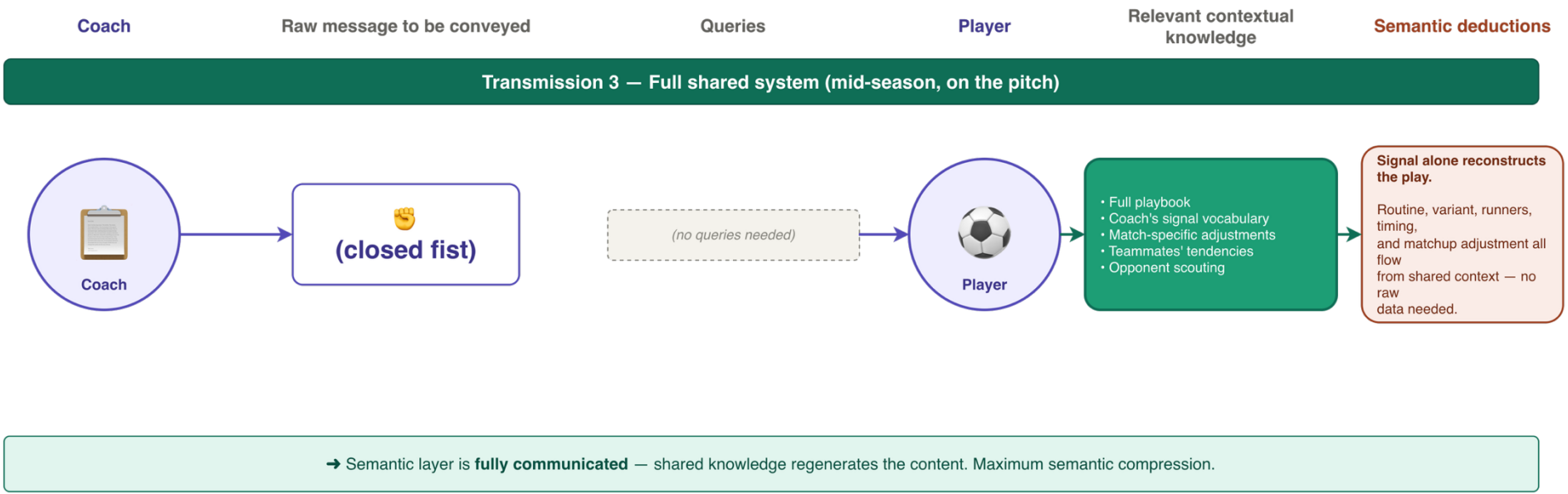


Coach/player example



➜ Semantic layer is complemented by classical data of **novel semantic content elements only** — moderate compression.

Coach/player example



➜ Semantic layer is **fully communicated** — shared knowledge regenerates the content. Maximum semantic compression.

SemCom disambiguation

- SemCom is not data compression
  - it captures structure, meaning, and reusable knowledge.
- SemCom is broader than goal-oriented communication,
  - It can support content delivery, XR and metaverse services, receiver-side generation, knowledge-base growth.
- Semcom is not just Application-aware networking
  - SemCom works deeper: data structure, context, causal/statistical relations, reusable meaning.
  - This enables reasoning across applications and domains.

A fragmented research landscape

- SemCom is growing fast, but the field is not yet unified.
- Open issues remain around:
  - definitions,
  - semantic representations,
  - semantic metrics,
  - scalable architectures,
  - practical deployment.
- This motivates a structured tutorial view.

A taxonomy by communication objective

- Semantic-oriented communication
  - Reconstruct the meaning of the source.
- Goal- or task-oriented communication
  - Transmit what serves a downstream task.
- Semantic-aware communication
  - Use semantics to assist networking and resource-control decisions.

A taxonomy by enabling paradigm

- Theory of Mind
  - Shared knowledge, intent inference, concise interaction.
- Generative AI
  - Compact descriptions, semantic reconstruction, receiver-side generation.
- Deep JSCC
  - Learned semantic features transmitted robustly over noisy channels.

Why IoT is a demanding proving ground

- Why SemCom fits IoT
  - IoT traffic is often task-driven.
  - Devices communicate to support monitoring, inference, control, and coordination.
  - Sending only task-relevant meaning can reduce communication load.
- Why IoT is challenging
  - Massive scale and sporadic traffic.
  - Tight energy and bandwidth budgets.
  - Data locality and privacy.
  - Intermittent connectivity.
  - Device and channel heterogeneity.

Guiding questions of the tutorial

- What should be represented and transmitted: data, features, meaning, or task-relevant information?
- How can semantic models be trained when IoT data remain distributed?
- How can SemCom adapt to channels, devices, and traffic without full retraining?
- How should resources be allocated when semantic quality, latency, and energy matter?

Tutorial roadmap

- Part 1: SemCom foundations and Deep JSCC for IoT
- Part 2: Training and operating semantic encoders in IoT
- Part 3: Semantic-aware resource allocation
- Part 4: Open challenges and research directions

Part 1

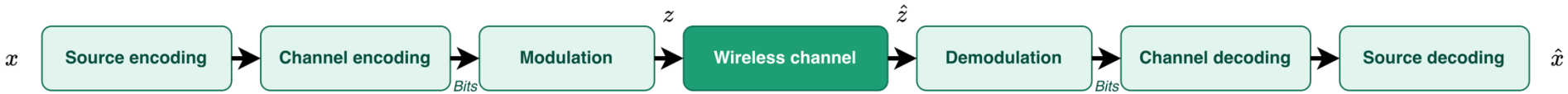
SemCom Foundations and Deep JSCC for IoT

Why start with image transmission?

- Image-rich services appear in autonomous systems, AR/VR, sensing, and IoT.
- They require high payload efficiency under latency, bandwidth, and energy constraints.
- They provide a concrete setting to compare conventional and semantic communications.
- This example as well as the corresponding results are adapted from:
  - E. Bourtsoulatze, D. B. Kurka and D. Gündüz, "Deep Joint Source-channel Coding for Wireless Image Transmission," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2019, pp. 4774-4778.
  - This will be denoted as Bourtsoulatze et al., ICASSP 2019.

Conventional digital communication pipeline

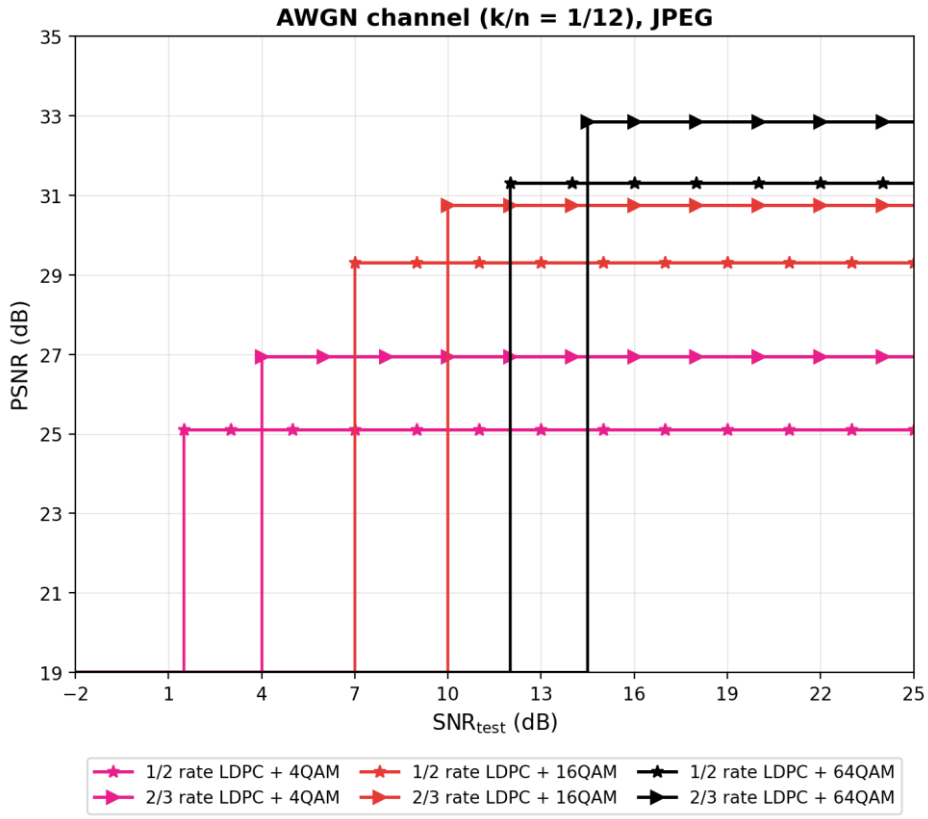
- Source encoding: compression, e.g., JPEG, JPEG2000.
- Channel encoding: error correction, e.g., LDPC, turbo codes.
- Modulation: digital mapping, e.g., BPSK, QAM.



Bourtsoulatze et al., ICASSP 2019.

The *cliff* effect

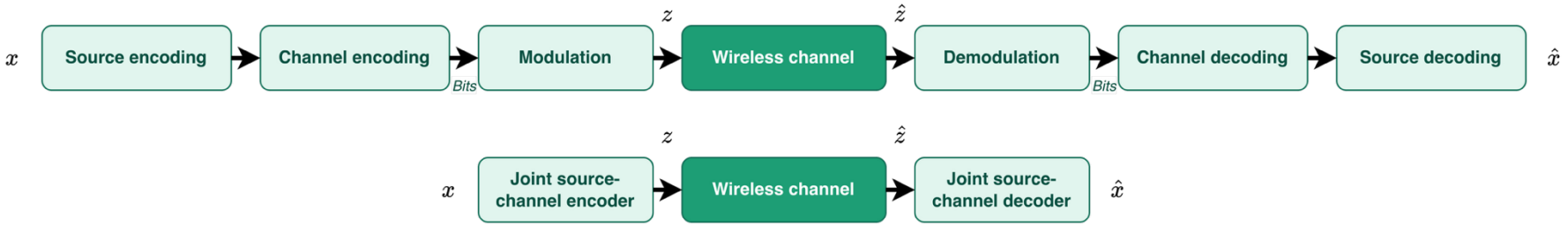
- Digital schemes are optimized for a target SNR.
- Below the design threshold, reconstruction quality can collapse sharply.
- Above the threshold, quality often saturates.



Bourtsoulatze et al., ICASSP 2019.

Introducing Deep JSCC

- End-to-end communication with an autoencoder architecture.
- Direct mapping from source samples to complex channel symbols.
- Joint optimization of source representation and channel robustness.
- No explicit intermediate bitstream.



Bourtsoulatze et al., ICASSP 2019.

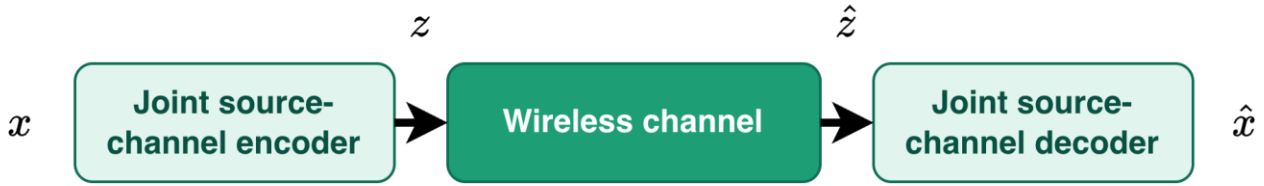
Why JSCC now?

- SemCom asks what information should be preserved.
- JSCC asks how this information should cross the channel.
- Deep JSCC makes this question trainable end to end.

Bourtsoulatze et al., ICASSP 2019.

Deep JSCC system model

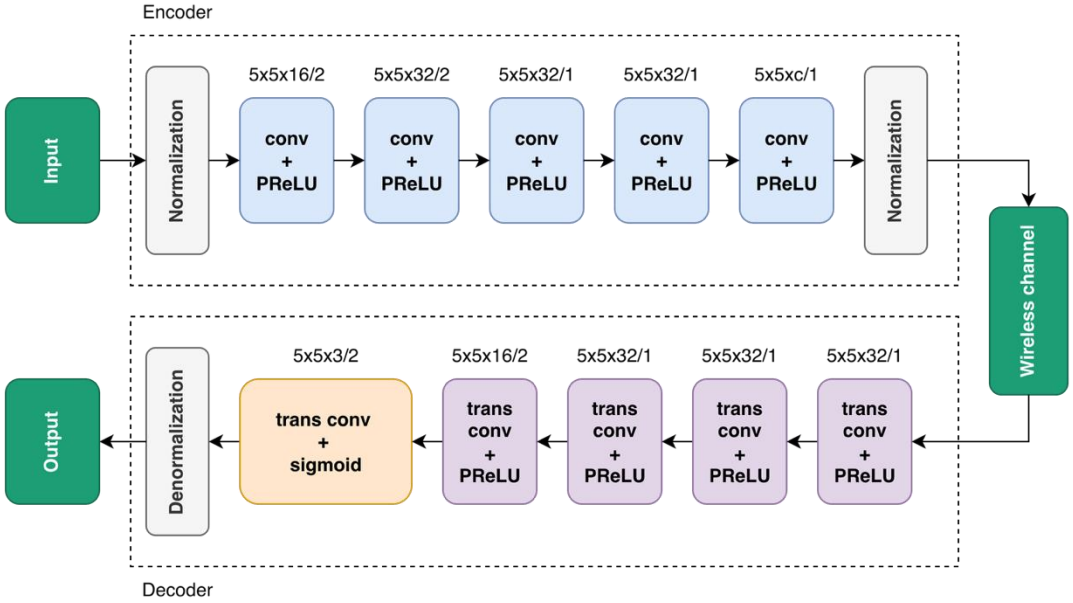
- Input Image $x \in R^n$
- Encoder $f_\theta : R^n \rightarrow C^k$
- Channel $\eta : C^k \rightarrow C^k$
- Decoder $g_\phi : C^k \rightarrow R^n$



Bourtsoulatze et al., ICASSP 2019.

Encoder architecture

- CNN-based semantic channel encoder.
- Convolutional layers extract hierarchical features.
- Final layers map features to complex channel symbols.
- Normalization enforces the transmit power constraint.



Bourtsoulatze et al., ICASSP 2019.

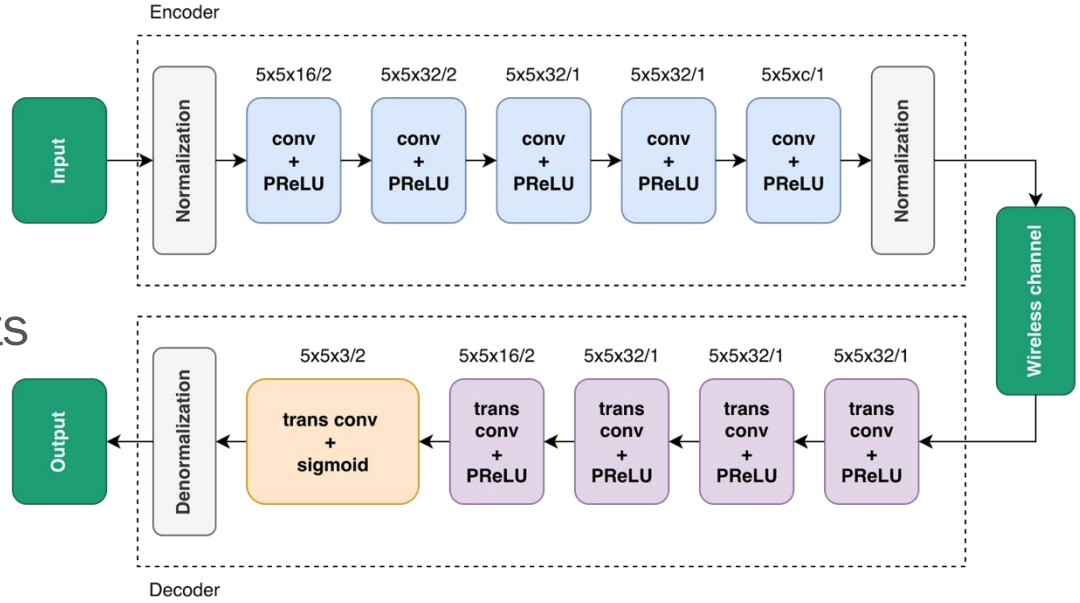
Differentiable channel layer

- The wireless channel is inserted as a non-trainable layer.
- It must be differentiable for backpropagation.
- Typical models: AWGN and slow Rayleigh fading.
- Random channel realizations expose the model to channel variability.

Bourtsoulatze et al., ICASSP 2019.

Decoder architecture

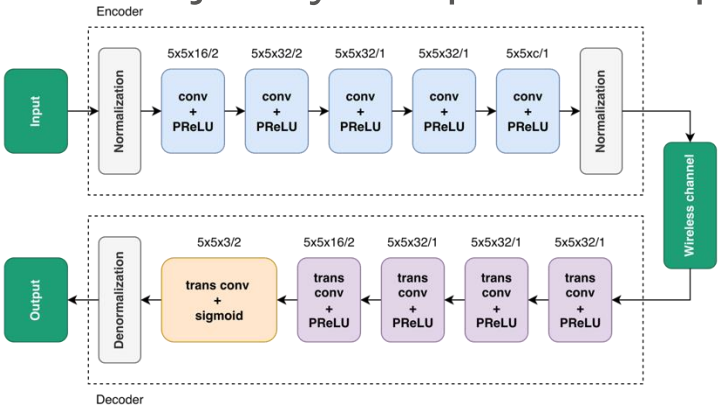
- Transposed convolutions reconstruct the image.
- The decoder mirrors the encoder structure.
- Final activation maps outputs to valid pixel values.



Bourtsoulatze et al., ICASSP 2019.

Training objective

- Loss: mean squared error for reconstruction.
- Optimizer: Adam.
- Training samples random channel realizations across mini-batches.
- The learned encoder must jointly compress and protect the source.



Bourtsoulatze et al., ICASSP 2019.

Bandwidth compression ratio

- n : source dimension, e.g., $3 \times h \times w$ for color images
- k : transmitted complex channel symbols.
- $\frac{k}{n}$: number of complex channel uses per source dimension.
  - Smaller $\frac{k}{n}$ means more aggressive compression
- Target rate is controlled through the final encoder width.
  - Different target rates typically require training or configuring different encoder widths.

Bourtsoulatze et al., ICASSP 2019.

Evaluation methodology

- Datasets: CIFAR-10 and Kodak.
- Baselines: JPEG/JPEG2000 with channel coding.
- Metrics: PSNR and SSIM.
- Channel conditions: AWGN and fading channels.

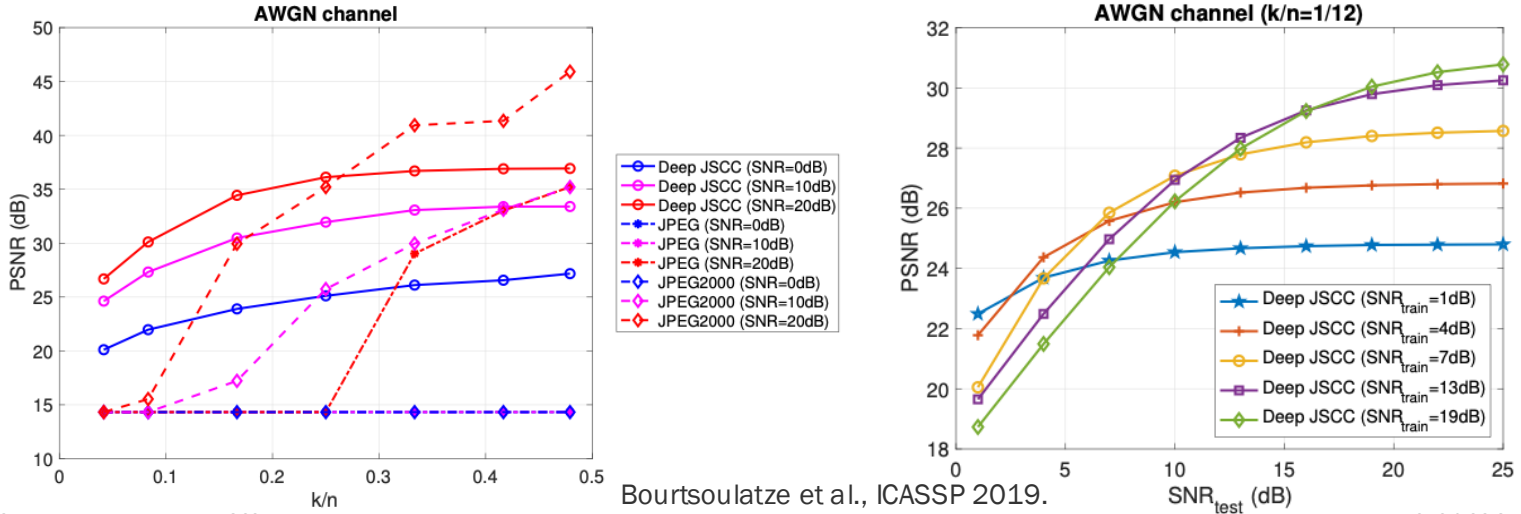
Bourtsoulatze et al., ICASSP 2019.

Demo: Deep JSCC over a noisy channel

- https://github.com/samerlahoud/deep-jscc-demo

Deep JSCC under constrained AWGN links

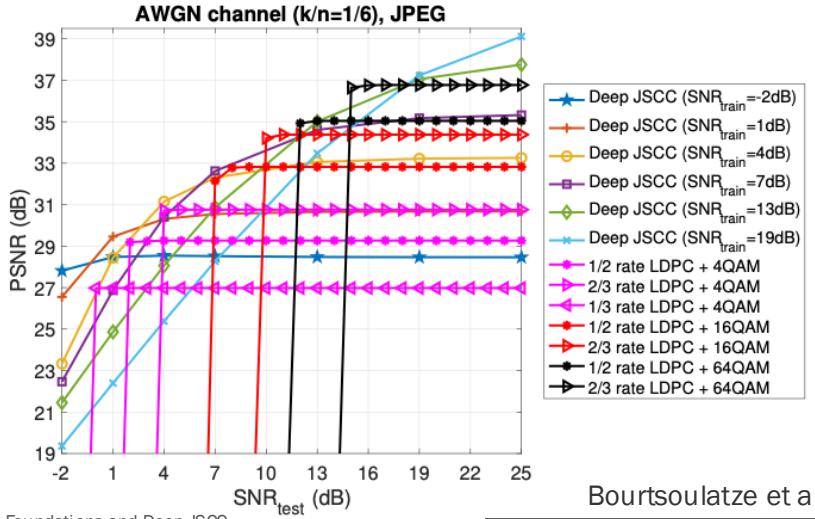
- At low SNR and low k/n, digital baselines fail to reconstruct images.
- Deep JSCC can preserve usable visual quality.
- Under SNR mismatch, quality degrades gradually.

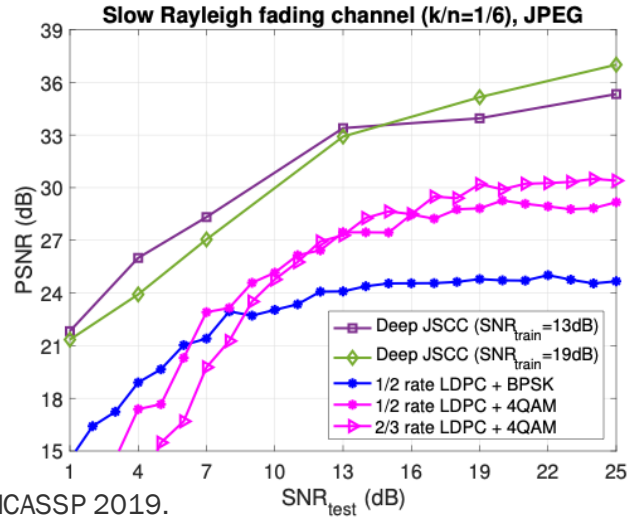


Bourtsoulatze et al., ICASSP 2019.

Deep JSCC against practical digital baselines

- Practical digital schemes show threshold behavior.
- Deep JSCC avoids hard decoding failure.
- The robustness advantage is stronger under fading.





Bourtsoulatze et al., ICASSP 2019.

Deep JSCC is an entry point for SemCom

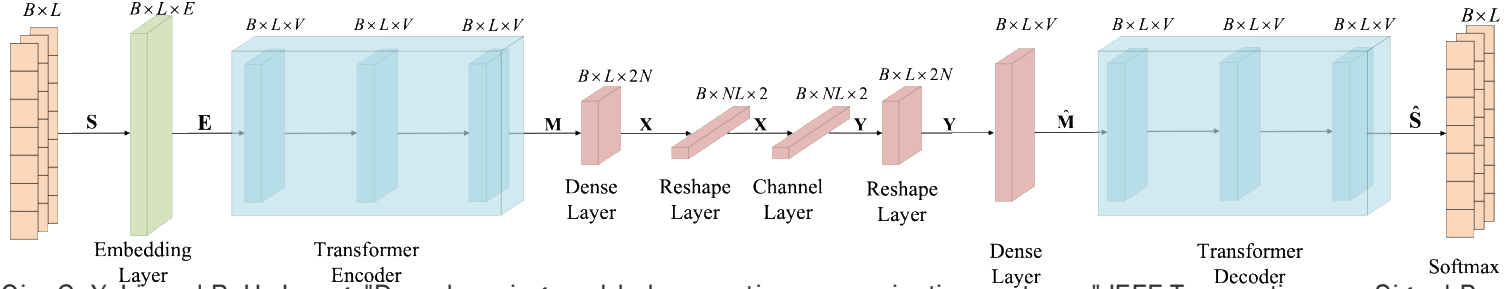
- Basic image Deep JSCC preserves visual structure.
- It becomes semantic when the objective is meaning, task utility, or receiver action.
- This distinction matters when choosing metrics and claims.

From images to other sources

- The Deep JSCC template generalizes across modalities:
  - Text: transformer encoder, sentence-level loss.
  - Speech: waveform or spectrogram encoder, perceptual/intelligibility loss.
  - Video: temporal encoder, perceptual and motion-aware loss.
  - Sensor streams: 1-D or recurrent encoder, task-aligned loss.
  - CSI feedback: representation targeting precoder or beamforming gain.
- Architecture and loss change with the modality; the end-to-end principle does not.

Text SemCom: DeepSC

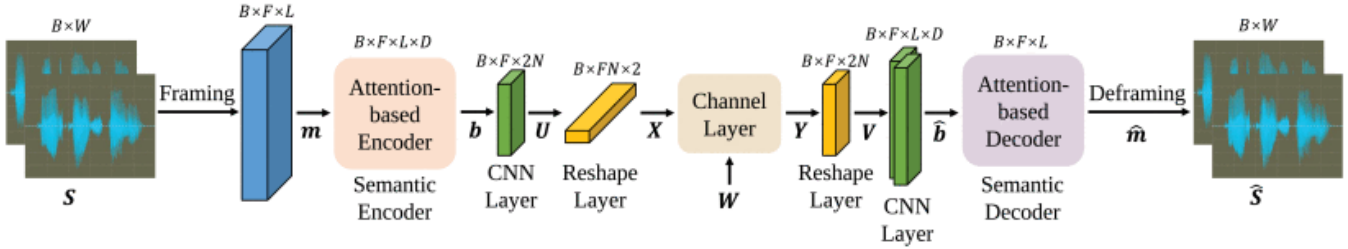
- End-to-end semantic transmission of natural-language sentences
  - Encoder: Transformer with self-attention over tokens
  - Loss: cross-entropy on tokens and mutual-information regularizer
  - Metric: BLEU score and sentence similarity in BERT embedding space
- Graceful degradation under low SNR
  - Received sentences remain semantically coherent



H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, "Deep learning enabled semantic communication systems," IEEE Transactions on Signal Processing, vol. 69, pp. 2663–2675, 2021.

Speech and audio

- End-to-end semantic transmission of speech.
  - Encoder over time-frequency features.
  - Loss combines waveform quality and speech-relevant weighting.
  - Metrics: SDR and PESQ.
- Preserves intelligibility under low-SNR conditions.



Z. Weng and Z. Qin, "Semantic communication systems for speech transmission," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 8, pp. 2434–2444, Aug. 2021.

What remains open after Deep JSCC?

- Generalization across sources, tasks, and channels.
- Adaptation without full retraining.
- Lightweight deployment on constrained devices.
- Compatibility with legacy network architectures.
- Evaluation metrics that reflect meaning and downstream action.

Part 2

Distributed and Adaptive SemCom for IoT

Why IoT changes SemCom deployment

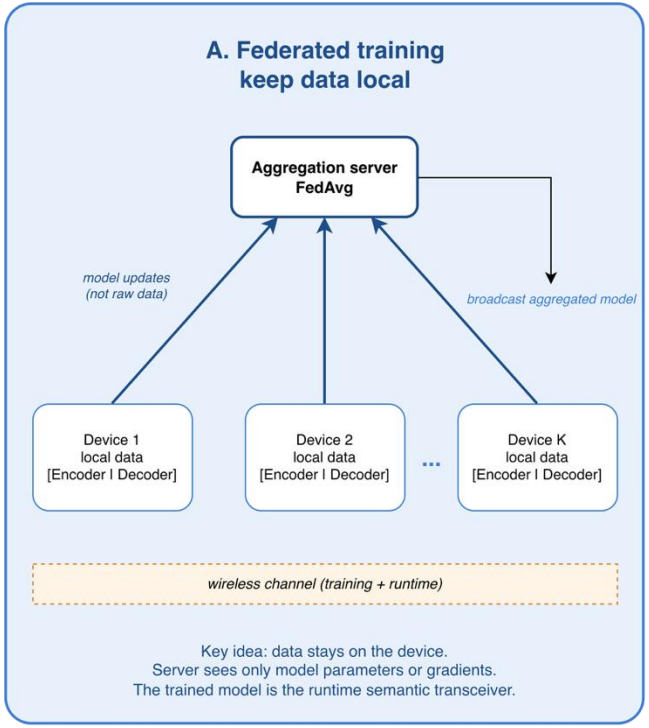
- Data constraints: locality, privacy, statistical heterogeneity.
- System constraints: compute, memory, energy.
- Network constraints: intermittent connectivity, scale, time-varying channels.

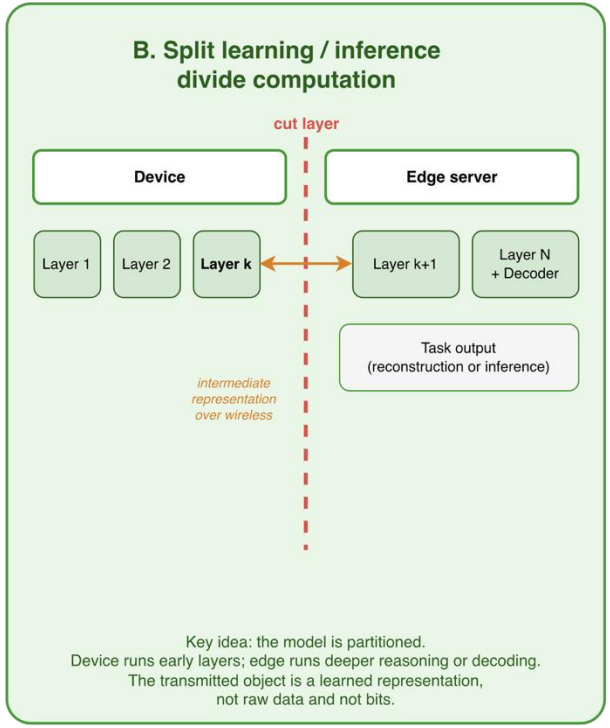
	Basic SemCom	IoT SemCom
Data	Global	Distributed, non-transferable
Compute	High-capacity GPU	Heterogeneous, energy-constrained
Channel	Idealized, stationary	Time-varying, intermittent

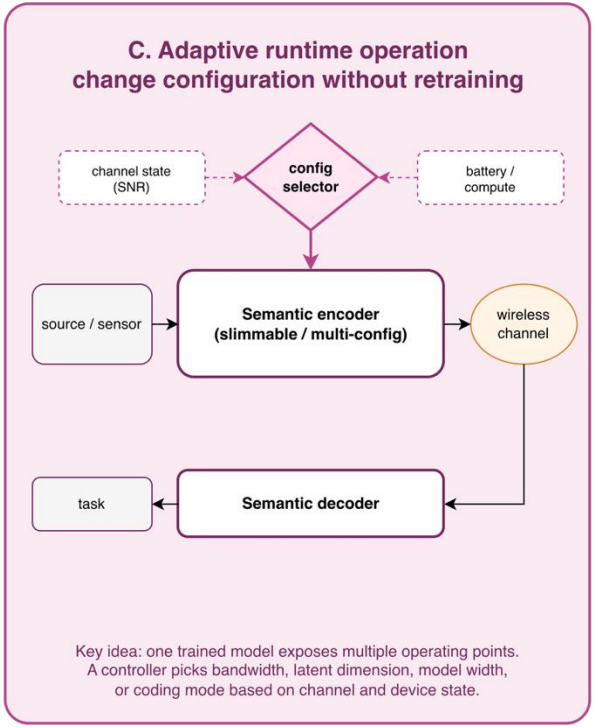
Three deployment questions

Mechanism	Question addressed	Design variable
Federated learning	Where are the parameters learned?	Data placement; aggregation
Split learning / inference	Where is the model executed?	Cut-layer; layer assignment
Runtime adaptation	How does the model operate over time?	Configuration selection (compute, SNR, rate)

Three mechanisms for SemCom in IoT

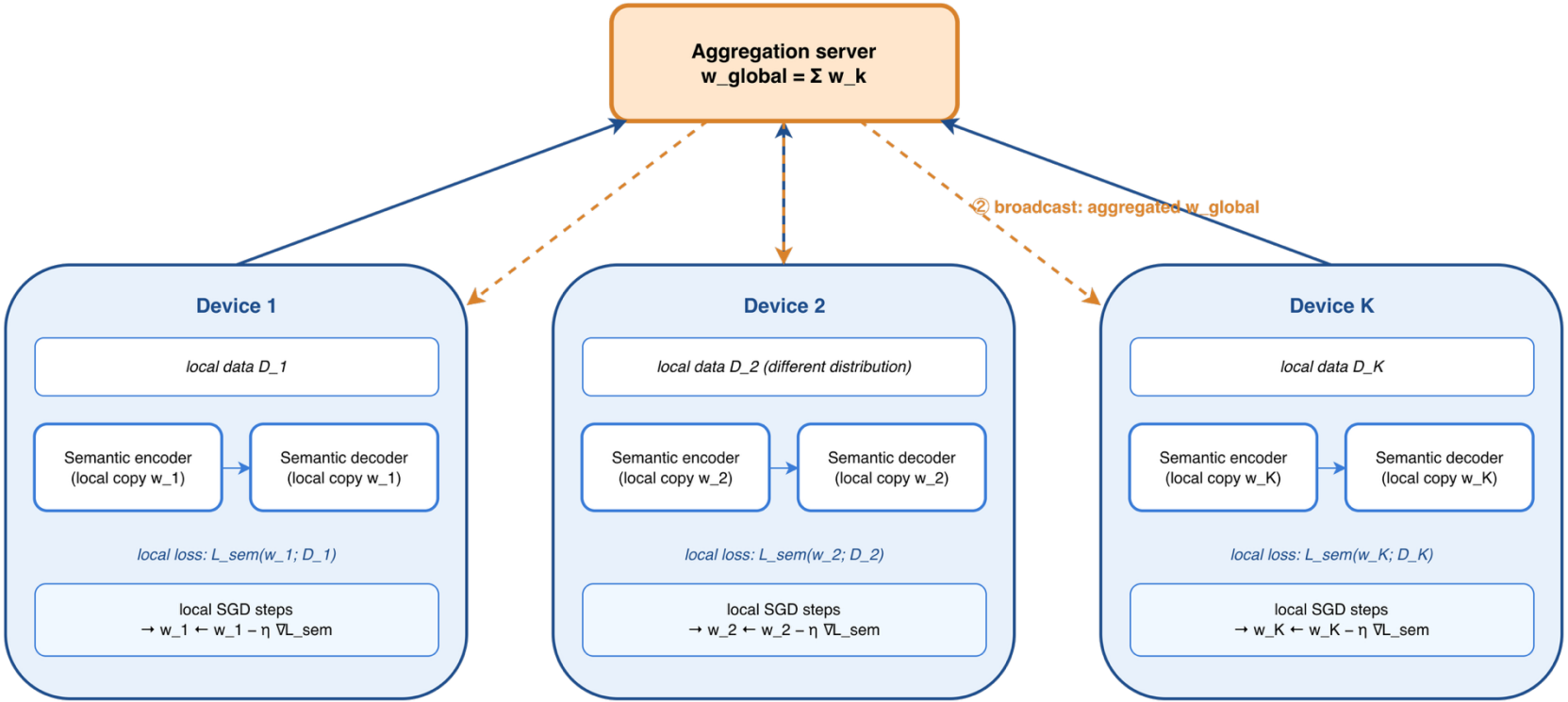






The same semantic encoder–decoder pair, viewed three ways: where to train (A), where to execute (B), and how to operate (C).

Federated learning: training without centralizing data



Federated SemCom trains the communication model

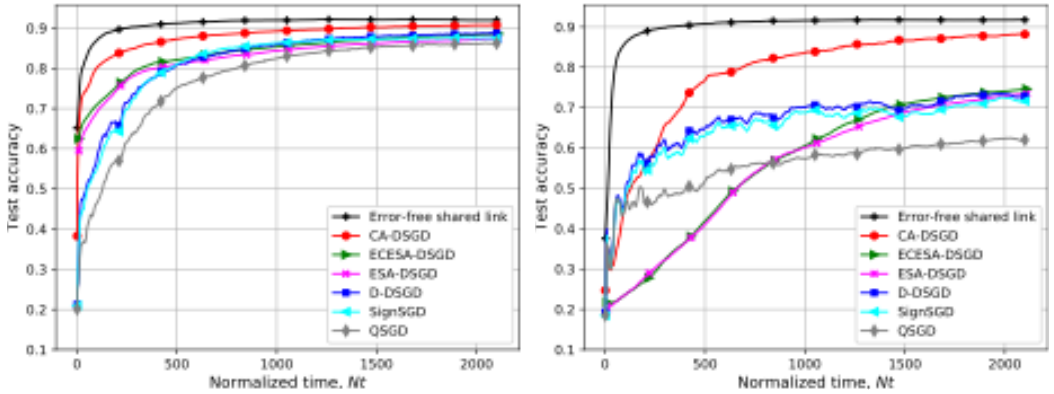
- Devices train local copies of a shared semantic encoder–decoder.
- A server aggregates model parameters across devices.
- Training quality determines the learned channel-side representation.
- Wireless impairments during aggregation affect the deployed encoder.
- Training communication cost competes with the eventual semantic gain.

Federated SemCom design variables

Mechanism	Decision
Wireless-aware aggregation	How are fading/noisy updates handled?
Semantic objective	Reconstruction, task loss, or hybrid?
Client selection	Who participates under channel/energy limits?
Model sharing	Global, clustered, or personalized?

Wireless-aware aggregation

- Uplink model updates are bandwidth-limited.
- Fading and noise distort aggregated parameters.
- Wireless aggregation affects convergence and final model quality.

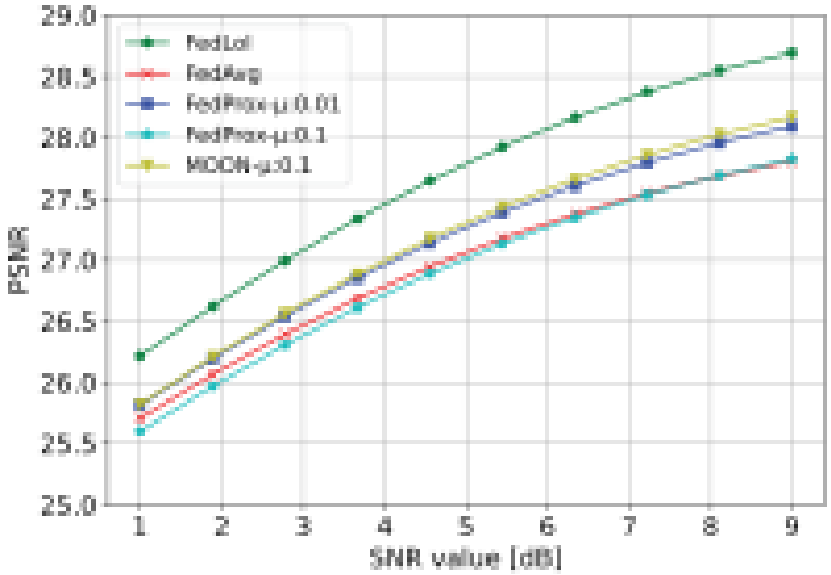


(a) IID data distribution (b) Non-IID data distribution

M. M. Amiri and D. Gündüz, "Federated Learning Over Wireless Fading Channels," in *IEEE Transactions on Wireless Communications*, vol. 19, no. 5, pp. 3546-3557, May 2020,

Loss-aware aggregation for SemCom FL

- Weight clients using local reconstruction loss.
- Avoid updating all semantic/channel modules every round.
- Reduce training overhead while preserving semantic fidelity.
- Aggregation affects semantic fidelity, not only convergence.



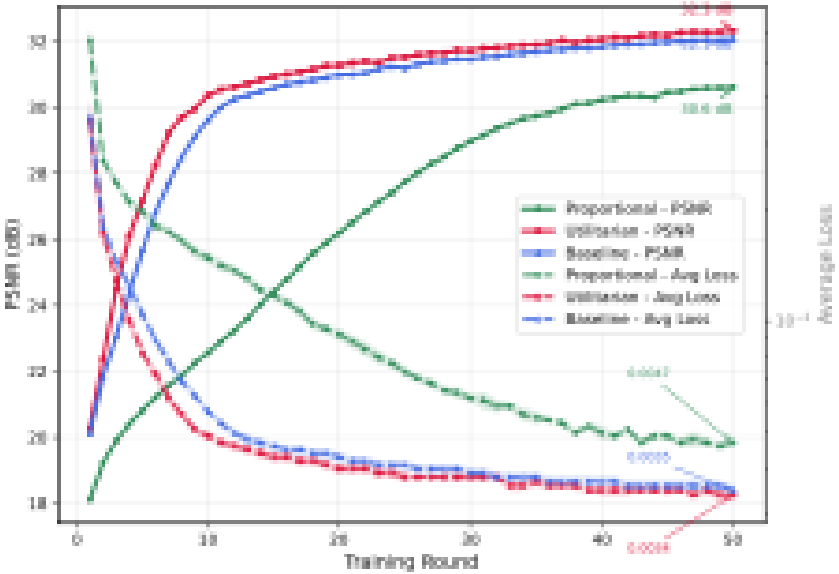
L. X. Nguyen et al., "An Efficient Federated Learning Framework for Training Semantic Communication Systems," in IEEE Transactions on Vehicular Technology, vol. 73, no. 10, pp. 15872-15877, Oct. 2024

Client selection and fairness

- Average semantic fidelity is not enough.
- Some clients may be repeatedly selected because they have larger datasets, cleaner data, or better channels.
- In SemCom FL, unfair selection creates unequal semantic fidelity after deployment.
- Design question: who participates when data quality, channel quality, and energy cost differ?

Client selection: quality–fairness trade-off

- Heterogeneous IoT clients differ in data size, distribution, and participation cost.
- Three strategies: baseline, utilitarian, proportional fairness.
- Utilitarian maximizes PSNR, while proportional fairness keeps competitive quality with lower inequality.

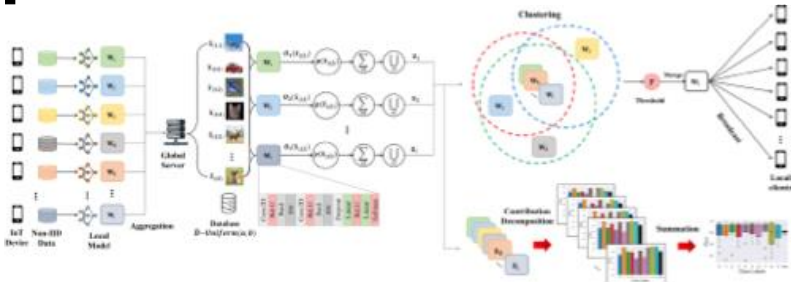


S. Lahoud and K. Khawam, "Client Selection Strategies for Federated Semantic Communications in Heterogeneous IoT Networks," 2025 IEEE 102nd Vehicular Technology Conference (VTC2025-Fall), Chengdu, China, 2025

Model sharing under heterogeneity

- One global encoder is simple but may be biased.
- Heterogeneous IoT devices may need different semantic representations.
- Research directions:
  - Global FL: one shared encoder–decoder.
  - Clustered FL: one model per group.
  - Personalized FL: local encoders with shared alignment.
  - Hierarchical FL: local aggregation, then global coordination.

Clustered FL: one model per semantic context

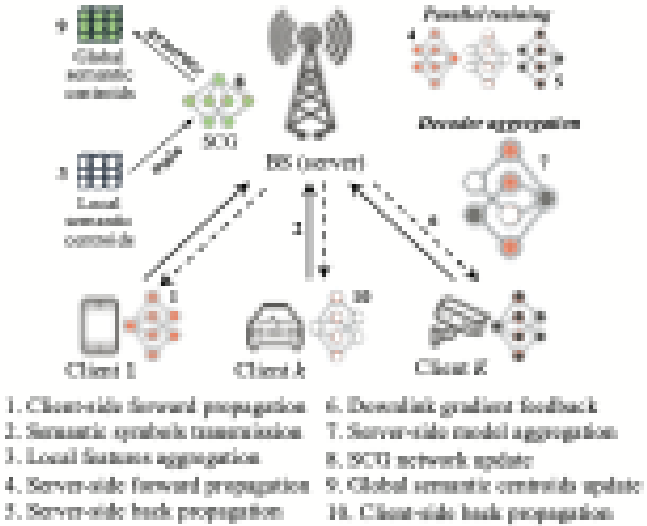


- Group devices with similar data/tasks.
- Train one SemCom model per cluster.
- Useful when devices observe different semantic contexts.
- Convergence quality within each cluster depends on cosine similarity of gradients and per-device participation probability.
- Risk: cluster assignment and participation become design problems.

B. Xu et al., "Clustered Federated Learning in Internet of Things: Convergence Analysis and Resource Optimization," IEEE Internet of Things Journal, vol. 11, no. 2, pp. 3217-3232, Jan. 2024.
H. Lee and D. Seo, "CLSM-FL: Clustering-Based Semantic Federated Learning in Non-IID IoT Environment," in *IEEE Internet of Things Journal*, vol. 11, no. 18, pp. 29838-29851, 15 Sept.15, 2024

Personalized FL: local encoders, shared semantics

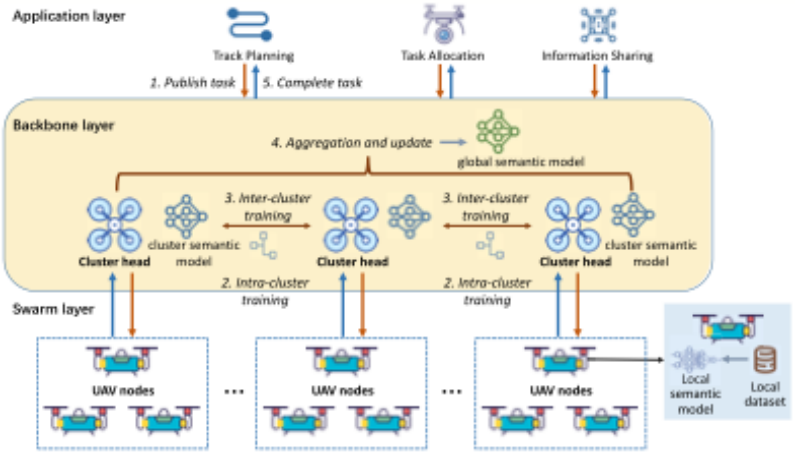
- Each device adapts its encoder to local data/channel.
- A shared decoder or shared semantic space keeps models compatible.
- Better for strong non-IID data and device-specific channels.
- Risk: maintaining semantic alignment across clients.



Y. Wang, et al., "Federated Contrastive Learning for Personalized Semantic Communication," IEEE Wireless Communications Letters, 2024.

Hierarchical FL: two-tier aggregation

- Shared backbone, aggregated in two tiers
  - Cluster heads collect intra-cluster updates
  - then consensus across cluster heads
- Natural fit for dynamic topologies and intermittent links



J. Xu, H. Yao, R. Zhang, T. Mai, S. Huang, S. Guo, "Federated Learning Powered Semantic Communication for UAV Swarm Cooperation," IEEE Wireless Communications, vol. 31, no. 4, pp. 140-146, Aug. 2024.

Split learning and split inference

- Partition the semantic model across device and edge.
  - Device executes the early layers of the encoder.
  - Edge executes the deeper layers and/or the decoder.
- The transmitted object is an intermediate representation.
- Two transmissions per training step:
  - Smashed data on the uplink.
  - Gradients on the downlink.
- Inference only requires the forward pass.

Why split SemCom?

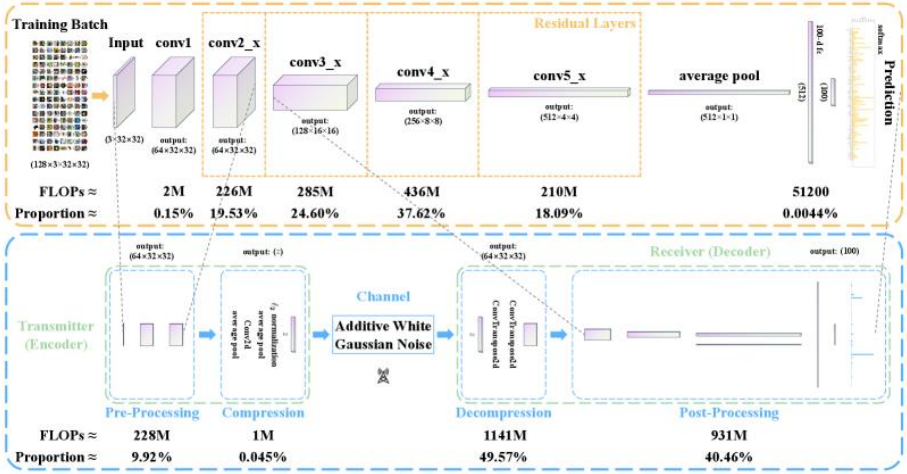
- Problem
  - Semantic encoders can be too heavy for end devices
  - IoT data are local, scarce, and privacy-sensitive
- Split principle: partition the model at a cut layer
  - Device runs the lightweight front-end; edge runs the heavier back-end
- SemCom interpretation
  - The cut layer controls the compute/uplink trade-off.
  - The transmitted object is a learned semantic representation.

Split SemCom design variables

Mechanism	Question addressed	Design variable
Cut placement	Where is the model partitioned?	Layer index k
Cut representation	What crosses the link?	Feature dimension, quantization, compression
Multi-device coordination	How does training scale?	Sequential, parallel, or hybrid with FL

Empirical sweep: where does computation happen?

- ResNet-34 split across device and edge
- Candidate cuts SP-1 to SP-5, after successive residual blocks



C. Wang, et al. "Deep Semantic Inference over the Air: An Efficient Task-Oriented Communication System." arXiv preprint arXiv:2508.12748 (2025).

Empirical sweep: what happens to accuracy?

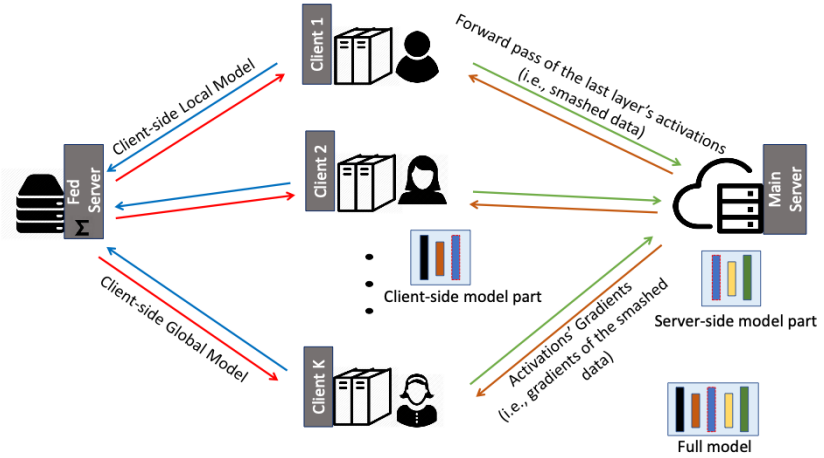
- Accuracy versus location
  - SP-1 collapses (≈0.43 top-1, even with large feature size)
  - SP-2 already reaches ≈0.68 against a 0.78 baseline
  - Over 85% of baseline accuracy retained with appropriate feature compression

Split Point	Transmitter (Edge Device)				Receiver (Cloud Device)				Top-1 Accuracy
	FLOPs	Prop. (%)	Param.	Prop. (%)	FLOPs	Prop. (%)	Param.	Prop. (%)	
SP-1 (conv1-conv2_x)	2.82	0.12	0.067	0.25	2,298.53	99.88	26.55	99.75	0.4290
SP-2 (conv2_x-conv3_x)	229.31	9.96	0.29	1.08	2,072.04	90.04	26.33	98.92	0.6781
SP-3 (conv3_x-conv4_x)	515.57	37.83	1.47	5.61	847.30	62.17	24.69	94.39	0.7602
SP-4 (conv4_x-conv5_x)	953.88	79.12	8.41	35.35	251.71	20.88	15.39	64.65	0.7715
SP-5 (conv5_x-avgpool, fc)	1,167.79	98.93	21.78	93.06	12.63	1.07	1.62	6.94	**0.7781**

C. Wang, et al. "Deep Semantic Inference over the Air: An Efficient Task-Oriented Communication System." arXiv preprint arXiv:2508.12748 (2025).

Scaling split learning to many devices

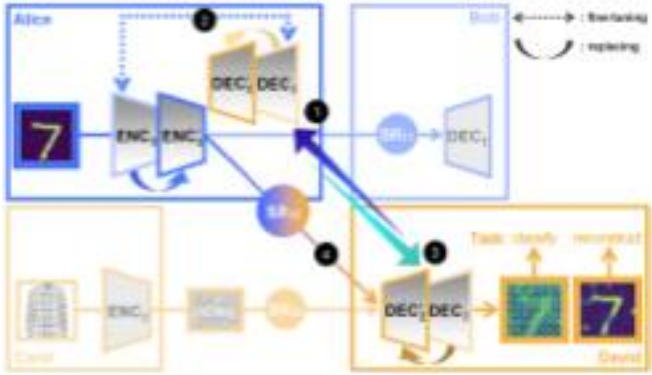
- Scalability problem in split learning
  - Devices train sequentially, one at a time
  - End-to-end training latency scales with the number of devices
- SplitFed: split learning meets federated averaging
  - Each device runs its split half in parallel
  - A main server hosts the shared edge-side network
  - A Fed server periodically averages device-side weights



C. Thapa, P. C. M. Arachchige, S. Camtepe, and L. Sun, "SplitFed: When federated learning meets split learning," Proc. AAAI, vol. 36, no. 8, pp. 8485–8493, 2022.

Split learning for semantic alignment

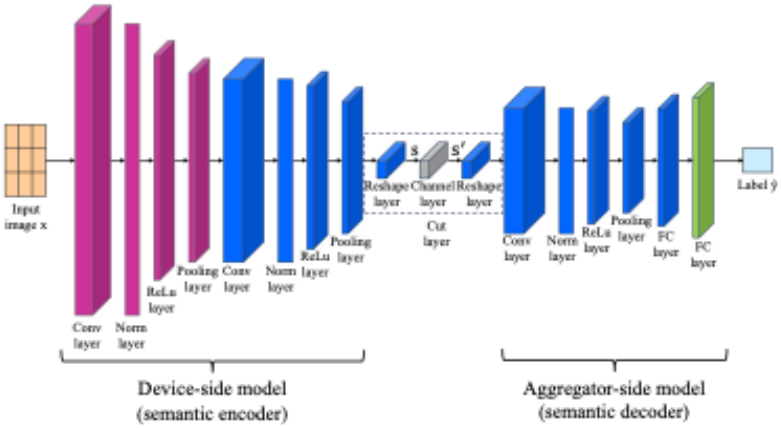
- The multi-user problem
  - Independent neural transceivers may learn incompatible representations.
- Split learning as an alignment mechanism
  - Partial fine-tuning aligns transceivers across users.
  - Layer freezing reduces the training burden.
  - No need to retrain end-to-end from scratch.



J. Choi et al., “Semantics Alignment via Split Learning for Resilient Multi-User Semantic Communication,” IEEE Transactions on Vehicular Technology, 2024.

Few-shot semantic split learning

- IoT data challenge
  - End devices hold few labeled samples
  - Distributions differ across sites
- Meta-split learning
  - Device-side encoder is meta-trained across related tasks
  - Edge-side head adapts to a new task from a small number of samples
- TinyML-compatible by construction
  - Device footprint sized for microcontroller-class hardware
  - Wireless channel included in the meta-training loop



E. Eldeeb, M. Shehab, H. Alves, and M.-S. Alouini, "Semantic Meta-Split Learning: A TinyML Scheme for Few-Shot Wireless Image Classification," IEEE Transactions on Machine Learning in Communications and Networking, vol. 3, pp. 491-501, 2025.

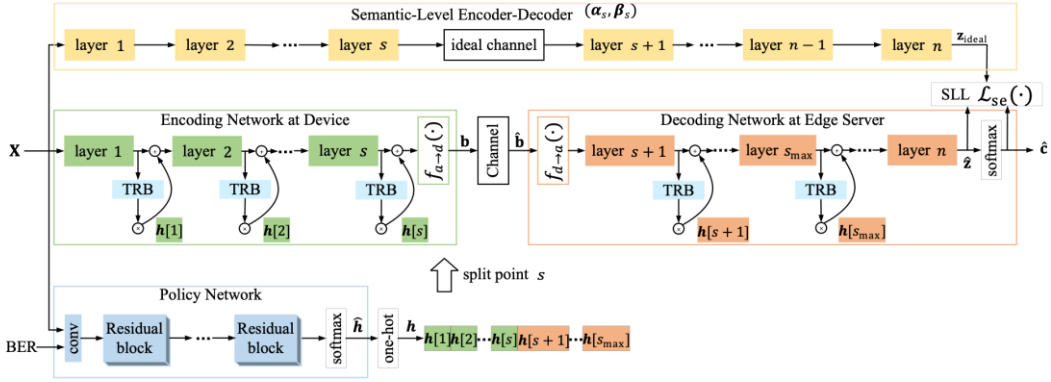
Static versus adaptive cut selection

- Static cut
  - Simple, predictable, easy to deploy
  - Insensitive to battery, channel, or load
- Adaptive cut at training time
  - Joint optimization of cut layer, transmit power, participation
  - Constrained by per-device energy budget
- Adaptive cut at inference time
  - Multiple split configurations trained jointly offline
  - Runtime selects the configuration that fits the current channel SNR
  - Switching cost negligible when configurations share weights

Z. Li, W. Wu, S. Wu, and W. Wang, "Adaptive Split Learning over Energy-Constrained Wireless Edge Networks," IEEE INFOCOM 2024 Workshops, pp. 1–6.
J. Lee, H. Lee, and W. Choi, "Wireless Channel Adaptive DNN Split Inference for Resource-Constrained Edge Devices," IEEE Commun. Lett., vol. 27, no. 6, pp. 1520–1524, 2023.

Adaptive digital SemCom with learned quantization

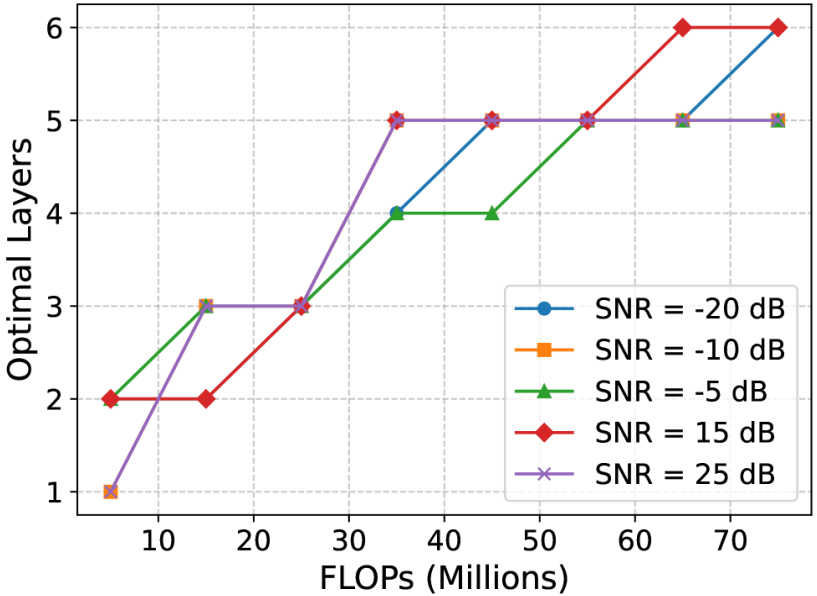
- Device hosts encoding network and learned non-linear quantizer
- Edge hosts decoding network
- Policy network selects, per input and per channel state
  - Split-layer index
  - Transmitted feature dimension
  - Quantization levels



L. Guo, W. Chen, Y. Sun and B. Ai, "Digital-SC: Digital Semantic Communication With Adaptive Network Split and Learned Non-Linear Quantization," in IEEE Transactions on Cognitive Communications and Networking, vol. 11, no. 4, pp. 2499-2511, Aug. 2025

Joint search over split design variables

- SNR- and FLOPs-adaptive Deep JSCC
  - Distributed CNN across IoT device and edge server
  - Adaptation to both channel SNR and device compute budget
- Design variables searched jointly
  - Split index
  - Filter count and kernel size
  - Latent feature dimension
- Learning-assisted genetic algorithm
  - Discards infeasible configurations early under FLOPs constraint
  - Random-forest regressors guide the search with offline-trained accuracy estimates



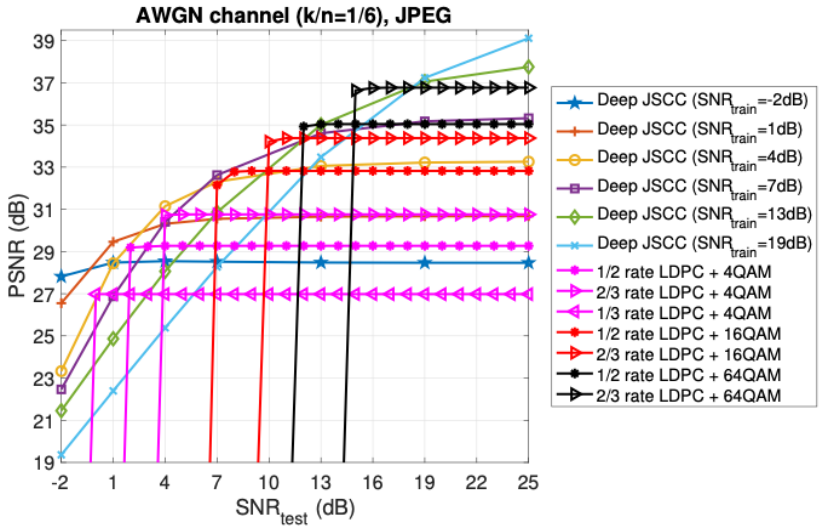
A. Waqas and S. Coleri, "SNR and Resource Adaptive Deep JSCC for Distributed IoT Image Classification," 2025 IEEE 36th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), Istanbul, Turkiye, 2025

Runtime adaptation

- Training once is not enough
  - Channel state varies after training.
  - Device energy and compute budgets change over time.
  - Traffic patterns and task requirements drift.
- Adaptation mechanisms
  - Bandwidth, latent dimension, model width, cut layer, coding mode
  - Configurable at runtime without full retraining.

Why adaptation starts from Deep JSCC

- Deep JSCC degrades smoothly under SNR mismatch.
- But graceful degradation is not the same as optimal operation.
- Runtime control is still needed to meet task, energy, and latency targets.

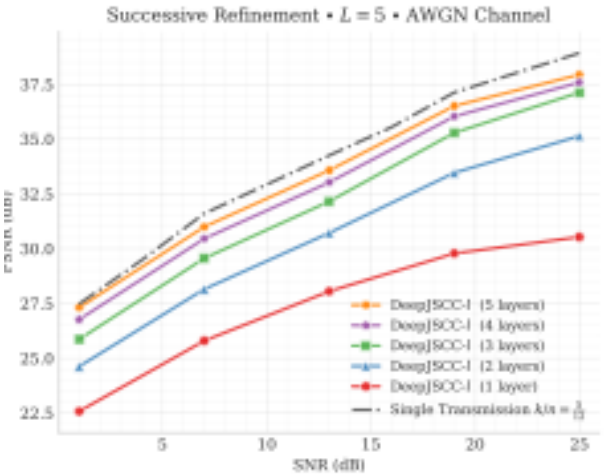


Channel-adaptive and device-adaptive models

- Channel-adaptive:
  - SNR-conditioned: encoder reads the current SNR and adjusts its rate
  - Attention-based: features are rescaled by channel quality
  - Channel-blind: model self-adapts from the received signal alone, without explicit SNR feedback
- Device-adaptive: one model, multiple operating points
  - Configurations vary in width, depth, latent dimension, or coding rate
  - Knowledge distillation keeps the smaller configurations accurate by transferring knowledge from the full model during training

Bandwidth-agile Deep JSCC

- One trained encoder, many channel uses
  - Output is an ordered stream of channel symbols
  - Decoder can stop reading at any rate above a minimum
- Successive refinement of the representation
  - Earlier symbols carry coarse semantic content
  - Later symbols refine progressively
- Implications for IoT
  - Rate adaptation without reloading a new model
  - One trained artifact spans the operating range



D. B. Kurka and D. Gündüz, "Bandwidth-Agile Image Transmission With Deep Joint Source-Channel Coding," in IEEE Transactions on Wireless Communications, vol. 20, no. 12, pp. 8081-8095, Dec. 2021

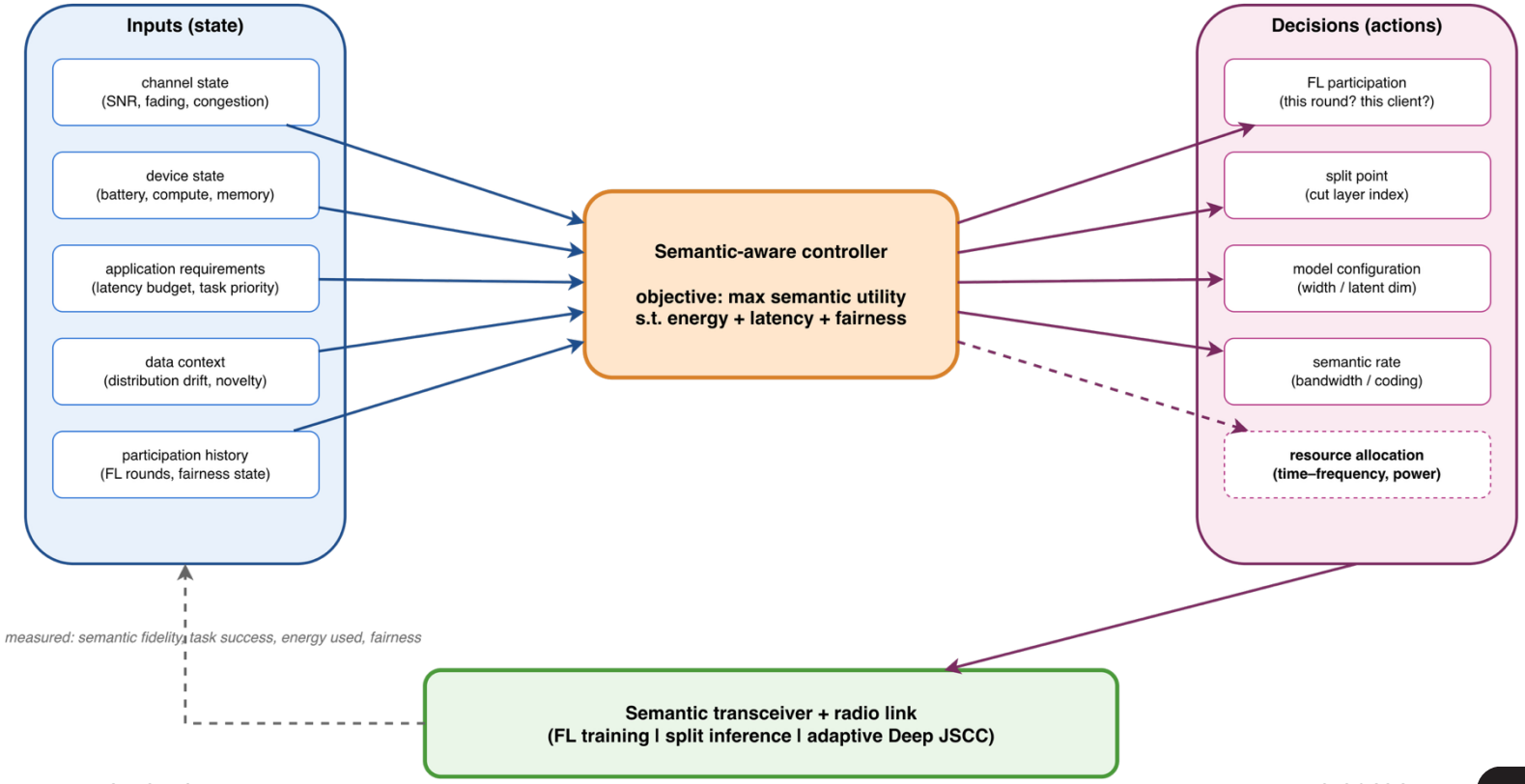
Train once, reconfigure at runtime

- The per-regime alternative scales poorly
  - One model per channel regime or per device class
  - Storage, retraining cost, and version management grow with diversity
- A single dynamic encoder covers a family of operating points
  - One trained model exposes multiple configurations: layer depth, width, latent dimension
  - Hierarchical layer activation: shallower modes are subgraphs of deeper ones
  - Sequential randomized training keeps all sub-models accurate
- Configuration selection is itself a control problem
  - Inputs: current SNR, available compute budget, task accuracy target
  - Output: which split point, which width, which latent dimension

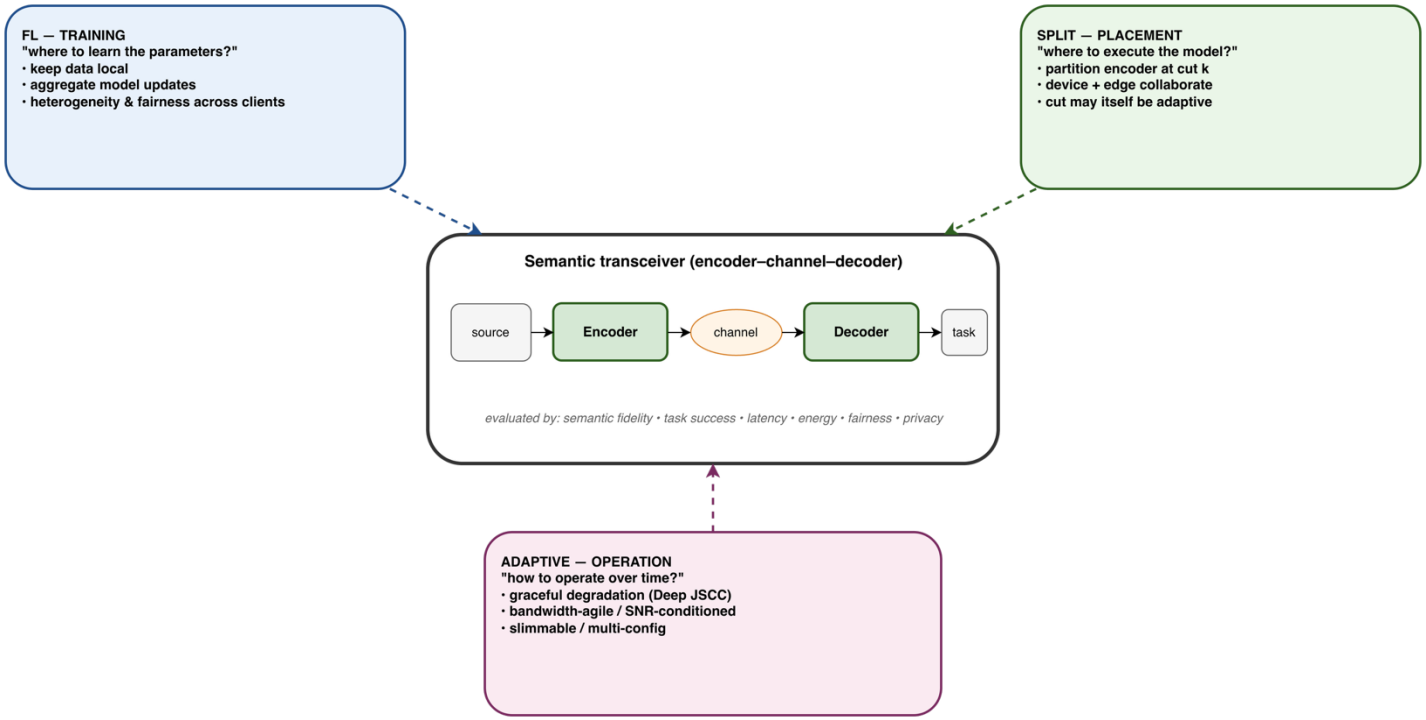
A. D. Raha, A. Adhikary, M. Gain, Y. Park, W. Saad and C. S. Hong, "DD-JSCC: Dynamic Deep Joint Source-Channel Coding for Semantic Communications," *ICC 2025 - IEEE International Conference on Communications*, Montreal, QC, Canada, 2025

A. Waqas and S. Coleri, "SNR and Resource Adaptive Deep JSCC for Distributed IoT Image Classification," 2025 IEEE 36th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), Istanbul, Turkiye, 2025

Semantic-aware control loop



Section takeaway



Part 3

Semantic-Aware Resource Allocation

From bit-centric to semantic-aware allocation

Design aspect	Bit-centric allocation	Semantic-aware allocation
Objective	Throughput, BER, outage	Task success, semantic fidelity
Unit of value	Bit, packet, flow	Feature, latent symbol, semantic update
Failure mode	Bit outage	Semantic outage

Semantic utility

- Semantic fidelity: how much meaning is preserved.
- Latency: whether the meaning arrives in time.
- Energy: how costly the semantic update is.
- Scheduling objective: maximize useful meaning per communication resource.

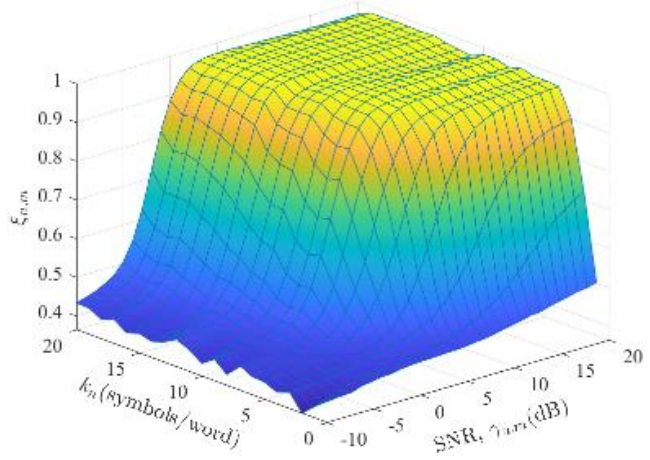
Semantic utility turns communication quality into a resource-allocation objective.

Semantic control variables

Control variable	Semantic interpretation
Feature selection	Which meaning-bearing components enter the channel
Semantic rate / compression	How much detail is preserved
Time / frequency / space	Where important features are transmitted
Power / beamforming	How strongly important features are protected
Multiple access	How semantic flows coexist
Computing placement	Where semantic processing is executed

Semantic spectral efficiency

- Build an empirical map between SNR, semantic-symbol budget, and recovered meaning.
- Define semantic rate from recovered similarity and communication cost.
- Increasing semantic symbols improves meaning recovery but consumes more channel resources
- The scheduler jointly selects semantic-symbol budget and channel assignment
  - Goal: maximize recovered meaning per radio resource.

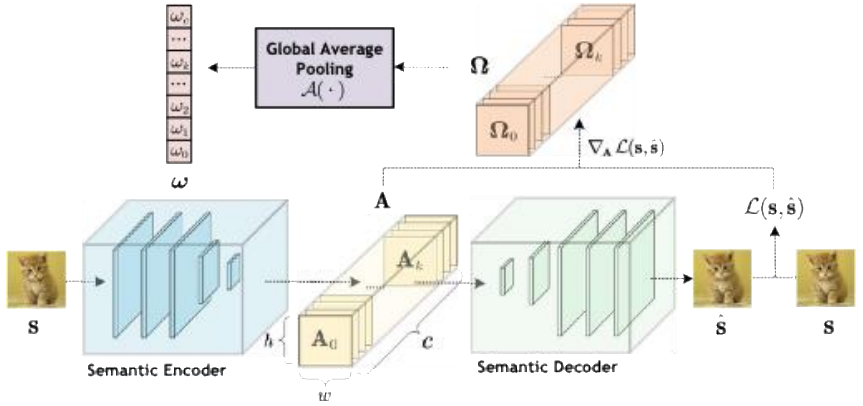


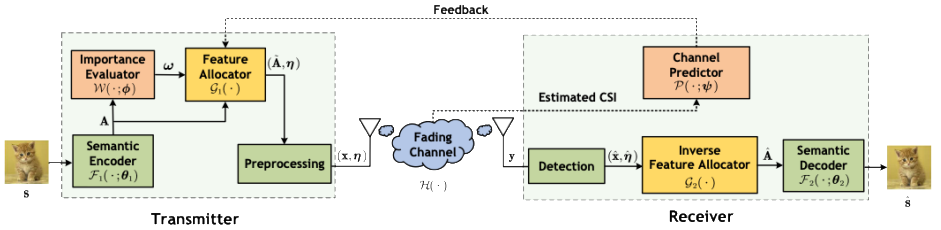
DeepSC semantic-similarity map: recovered meaning versus SNR and semantic-symbol budget.

L. Yan, Z. Qin, R. Zhang, Y. Li, and G. Y. Li, "Resource Allocation for Text Semantic Communications," IEEE Wireless Communications Letters, vol. 11, no. 7, pp. 1394–1398, Jul. 2022.

Feature-importance-aware allocation

- Estimate semantic feature importance from loss sensitivity.
  - Offline: gradients of reconstruction loss identify influential features.
  - Deployment: a lightweight evaluator predicts feature ranking.
- Channel prediction ranks future resources by expected quality.
- High-importance features are matched to high-quality resources.

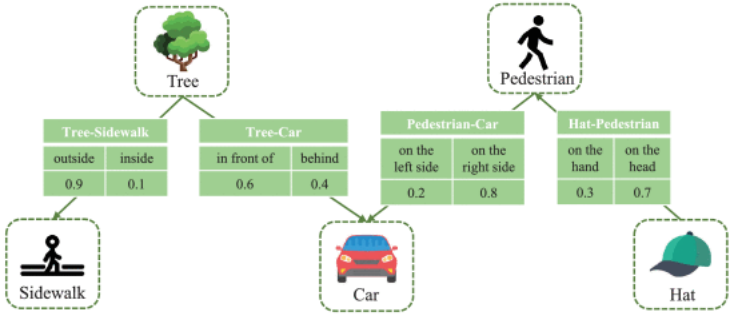




K. Zhou, G. Zhang, Y. Cai, Q. Hu, G. Yu, and A. L. Swindlehurst, "Feature Allocation for Semantic Communication with Space-Time Importance Awareness," IEEE Transactions on Wireless Communications, 2025.

Compression-aware resource allocation

- The semantic encoder can transmit coarse or detailed meaning.
  - Higher compression reduces payload and delay, but may reduce fidelity.
  - Power and RSMA protect the selected semantic representation.
- The scheduler jointly selects compression level and radio resources.
- Goal: maximize semantic utility under latency and power constraints.



Z. Zhao et al., "Compression Ratio Allocation for Probabilistic Semantic Communication With RSMA," in IEEE Transactions on Communications, vol. 73, no. 9, pp. 7304-7318, Sept. 2025

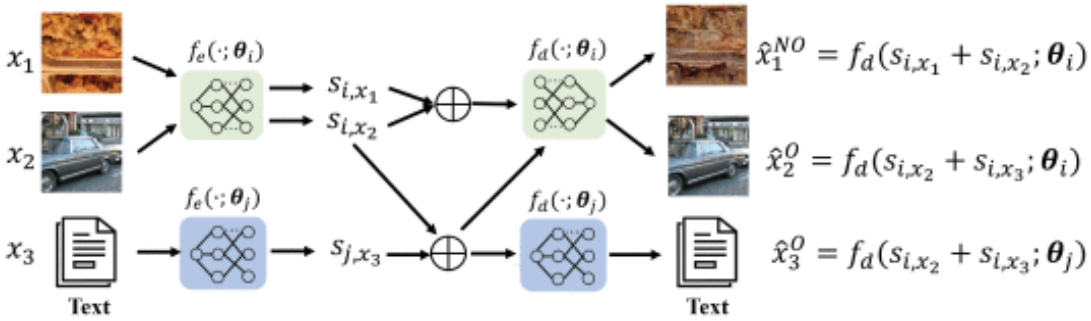
Semantic multiple access or model space as a resource

- Classical multiple access separates users by time, frequency, space, code, or power.
- SemCom introduces another dimension: semantic/model representation.
- Key question: can superposed semantic signals remain decodable?
- Two approaches:
  - model orthogonality,
  - semantic-interference-aware pairing.

Method	Approach
OMA / OFDMA	separate time-frequency blocks
NOMA	shared block, separated by power/SIC
Semantic MA	shared block, separated by semantic/model representation

Example: Orthogonal model division multiple access

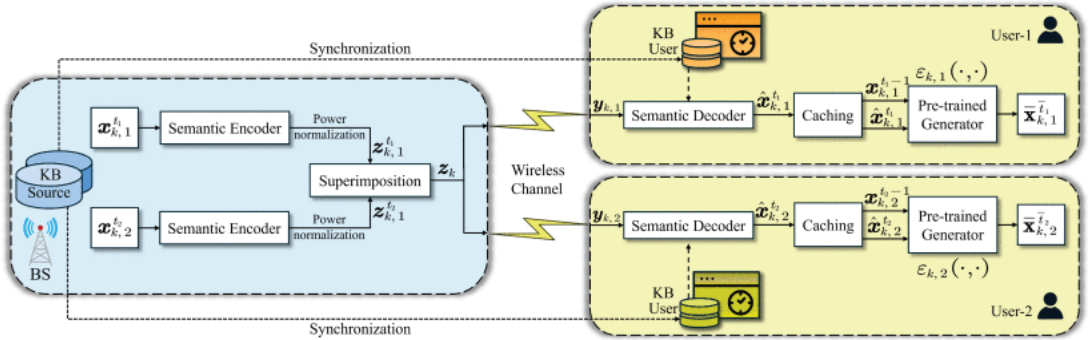
- Users share the same physical resource block.
- Each user uses a distinct semantic encoder-decoder pair.
- Other users' semantic signals behave as structured interference.
- Decoding is performed by the matched semantic model.



H. Liang et al., "Orthogonal Model Division Multiple Access," IEEE Transactions on Wireless Communications, vol. 23, no. 9, pp. 11693–11707, Sept. 2024.

Example: Semantic feature multiple access

- The base station superposes semantic features requested by paired users.
  - Each user decodes its own frame from the shared signal.
  - Generative interpolation favors pairs with small temporal gaps.
  - Semantic interference modifies the effective SINR.
- The scheduler jointly optimizes pairing and power allocation.



J. Wang et al., "Generative AI Empowered Semantic Feature Multiple Access (SFMA) Over Wireless Networks," IEEE Transactions on Cognitive Communications and Networking, vol. 11, no. 2, pp. 791–804, April 2025.

Conclusion

From Open Challenges to Deployable SemCom Systems

From research concepts to deployable SemCom systems

- A maturity path is emerging

Research stage	Main question
Semantic representation	What should be transmitted?
Semantic model	How is meaning encoded, decoded, and adapted?
Semantic system	Where do computation, storage, and radio resources interact?
Semantic prototype	What survives real channels, hardware, and latency?
Semantic standard	What can be exposed through interoperable interfaces?

Knowledge bases move SemCom beyond black-box features

- A semantic KB can support:
  - compact semantic representation
  - missing information recovery
  - semantic reasoning
  - task-dependent reconstruction
  - cross-modal and cross-context reuse
- But open problems remain:
  - source and destination KB mismatch
  - partially shared or evolving KBs
  - KB update and synchronization cost
  - KB poisoning and trust
  - quantitative modeling of semantic flow through KBs

Generative and LLM-based SemCom

- The shift
  - Foundation models (LLMs, diffusion, multimodal) serve as shared knowledge bases
  - Receiver generates content from compact semantic descriptors
  - Foundation models may enable cross-modal reuse across text, images, and sensor data
- New open problems
  - Knowledge-base drift between transmitter and receiver foundation models
  - Hallucination as a failure mode unique to generative SemCom
  - LLM inference energy is currently incompatible with mW-class IoT devices

Goal-oriented metrics

- A family of effectiveness metrics is converging, beyond accuracy.

Metric	What it captures
Age of Information (AoI)	Freshness — time since last update
Age of Incorrect Information (AoII)	Freshness penalized only when info is wrong
Value / Urgency of Information	Task-importance weighting of updates
Goal-oriented Tensor (GoT)	Unifies the above under one framework

A. Li, S. Wu, S. Meng, R. Lu, S. Sun and Q. Zhang, "Toward Goal-Oriented Semantic Communications: New Metrics, Framework, and Open Challenges," in IEEE Wireless Communications, vol. 31, no. 5, pp. 238-245, October 2024

Semantic security and robustness

- New attack surface
  - Semantic adversarial perturbations flip task output without changing bits much
  - Knowledge-base poisoning at training or update time
  - Channel-inversion attacks specific to learned encoders
  - Hallucination injection in generative SemCom
- Early-stage defenses
  - Adversarial training transfers
  - Masked Vector Quantized-Variational AutoEncoder codebooks show promise
  - No semantic-layer analogue of physical-layer encryption yet exists

V. -T. Hoang, et a.., "Adversarial Attacks Against Shared Knowledge Interpretation in Semantic Communications," in *IEEE Transactions on Cognitive Communications and Networking*, vol. 11, no. 2, pp. 1024-1040, April 2025
Q. Hu, G. Zhang, Z. Qin, Y. Cai, G. Yu and G. Y. Li, "Robust Semantic Communications With Masked VQ-VAE Enabled Codebook," in IEEE Transactions on Wireless Communications, vol. 22, no. 12, pp. 8707-8722, Dec. 2023

Standardization requires compatibility, interfaces, and validation

- Standardization is unlikely to start from abstract semantics.
- It will likely require:
  - representative use cases
  - unified evaluation metrics
  - modular encoder/decoder interfaces
  - compatibility with existing 5G/6G stacks
  - semantic-aware QoS mapping
  - reference models and datasets
  - simulation and experimental validation platforms
  - fallback to classical transmission when semantic processing fails