Task-aware Distributed Source Coding under Dynamic Bandwidth

Po-han Li*[1], Sravan Kumar Ankireddy*[1], Ruihan Zhao[1], Hossein Nourkhiz Mahjoub[2], Ehsan Moradi-Pari[2], Ufuk Topcu[1], Sandeep Chinchali[1], Hyeji Kim[1]

*Equal contribution, [1]University of Texas at Austin, [2]Honda Research Institute

TEXAS The University of Texas at Austin

Honda Research Institute US

NEURAL INFORMATION PROCESSING SYSTEMS

Background and Motivation

- Data compression eases communication overload in data-rich multi-sensor networks.
- Often, data collected by sensors is correlated and also analyzed by AI algorithms, not humans.
- Our goal is to (1) eliminate redundant information transmission from correlated data, (2) focus on transmitting task-relevant features, and (3) dynamically allocate bandwidth to sensors based on the importance of the task.

Contributions

- Identifying and measuring the **importance** of task-relevant features
- Elevating task performance by transmitting **task-relevant** features under bandwidth constraint.
- Theoretical analysis and optimal solution for the case of a **linear** compressor and task.
- A task-aware distributed source coding framework that performs **variable-rate** compression using a single model.

Notation

- X: correlated data
- Z: representations of data
- E: encoder
- D: decoder
- Φ: task function
- Y: task output
- $\mathcal{L}$: loss function
- $\hat{X}$: reconstructed data

Problem Formulation

$$\underset{E_1,\ldots,E_k,D}{\operatorname{argmin}} \quad \mathcal{L}_{\text{task}}(Y,\hat{Y}) + \lambda \mathcal{L}_{\text{rec}}(x,\hat{x})$$

$$\text{s.t.} \quad \hat{x}_i = D(E_1(x_1)), \text{for } i = 1,\ldots,k$$

$$Y = \Phi(x_1,\ldots,x_k)$$

$$\hat{Y} = \Phi(\hat{x}_1,\ldots,\hat{x}_k))$$

Neural Distributed Principal Component Analysis (NDPCA) Framework

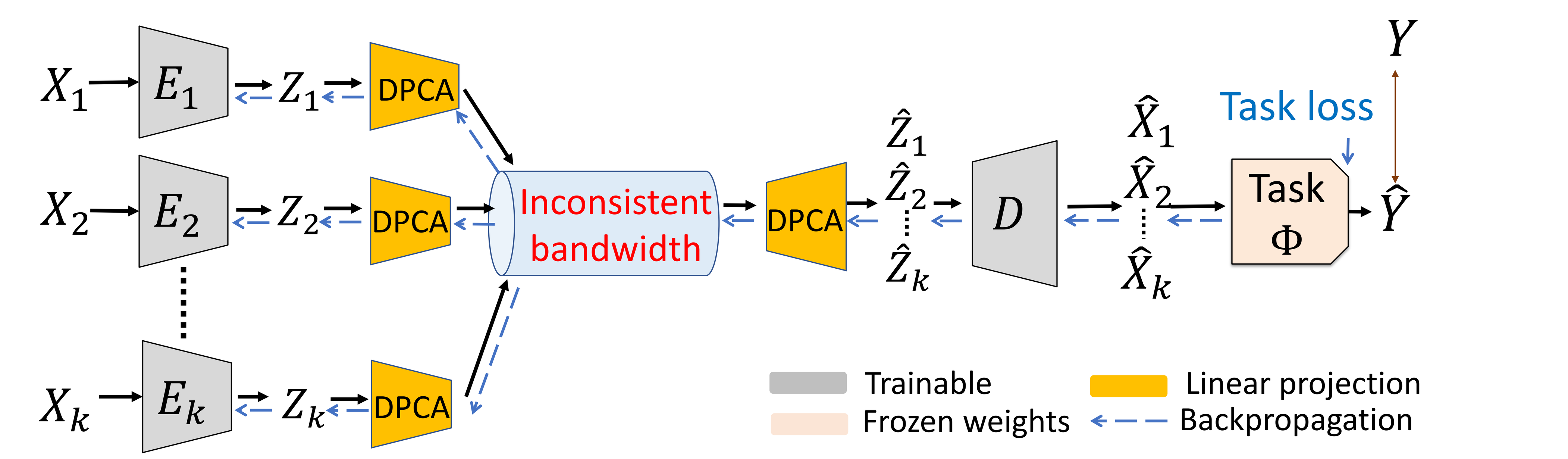

Figure: **Task-aware distributed source coding with NDPCA:** $X_1,\ldots,X_k$ are correlated data sources. Neural encoders $E_1,\ldots,E_k$ independently compress data to latent representations $Z_1,\ldots,Z_k$. The proposed DPCA module, which is a linear matrix, allocates the bandwidth of sources based on the importance of the task Φ.

- The framework uses **neural encoders** and their corresponding **neural decoder** to minimize the task loss, by measuring the task-performance on the reconstructed data, $\phi(\hat{X})$, and on uncompressed data, $\phi(X)$.
- The neural autoencoders are encouraged to generate **low-rank representations** Z, which helps achieve a systematic trade-off between latent dimension and performance with a **single model**.

Experimental Results

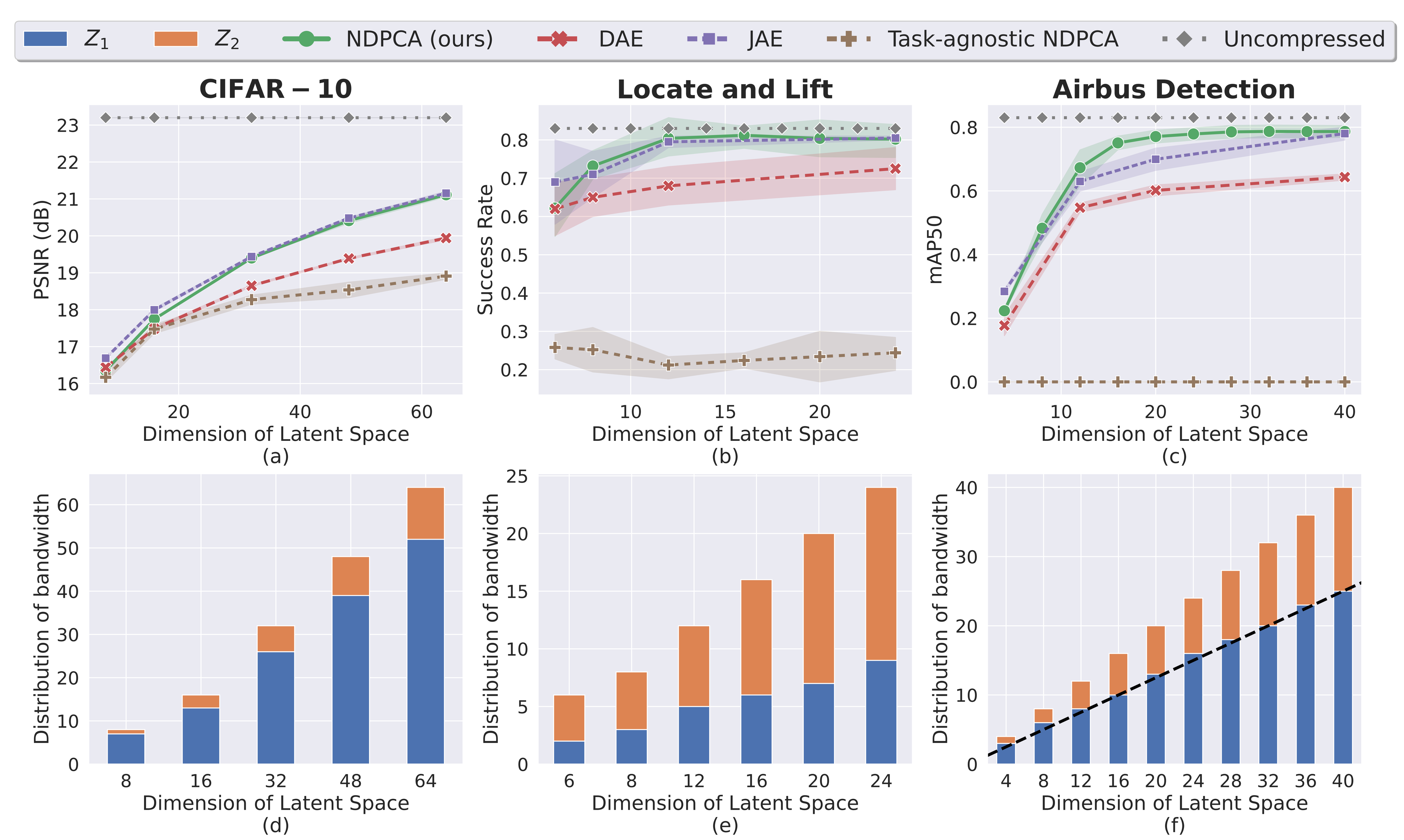

Figure: **Top:** Our method achieves equal or higher performance than other methods while reaching the upper bound of performance without data compression. **Bottom:** Distribution of total available bandwidth (latent space) among the two sources for NDPCA. The unequal allocation highlights the difference in the importance of the sources for a given task.

Datasets

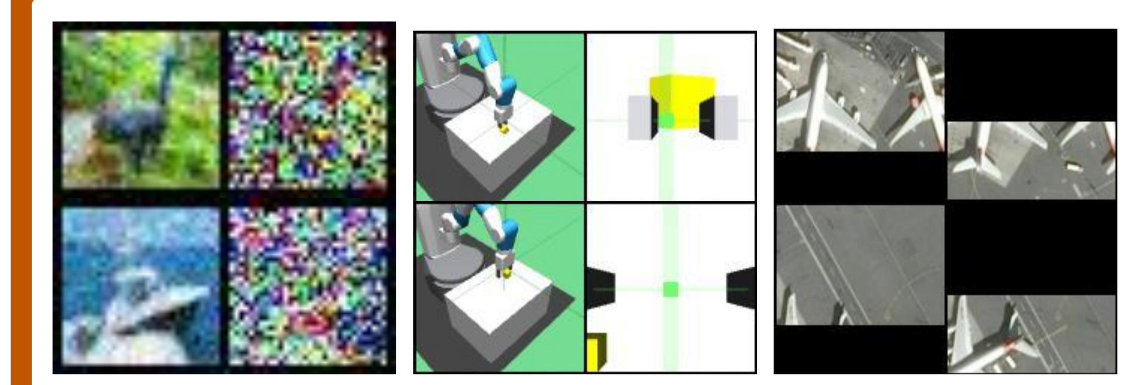

Figure: Two columns represent different sources of data. The two sources are **correlated**, but one is considered **more important** than the other because it is more relevant to the downstream task.

Task-aware vs. Task-agnostic

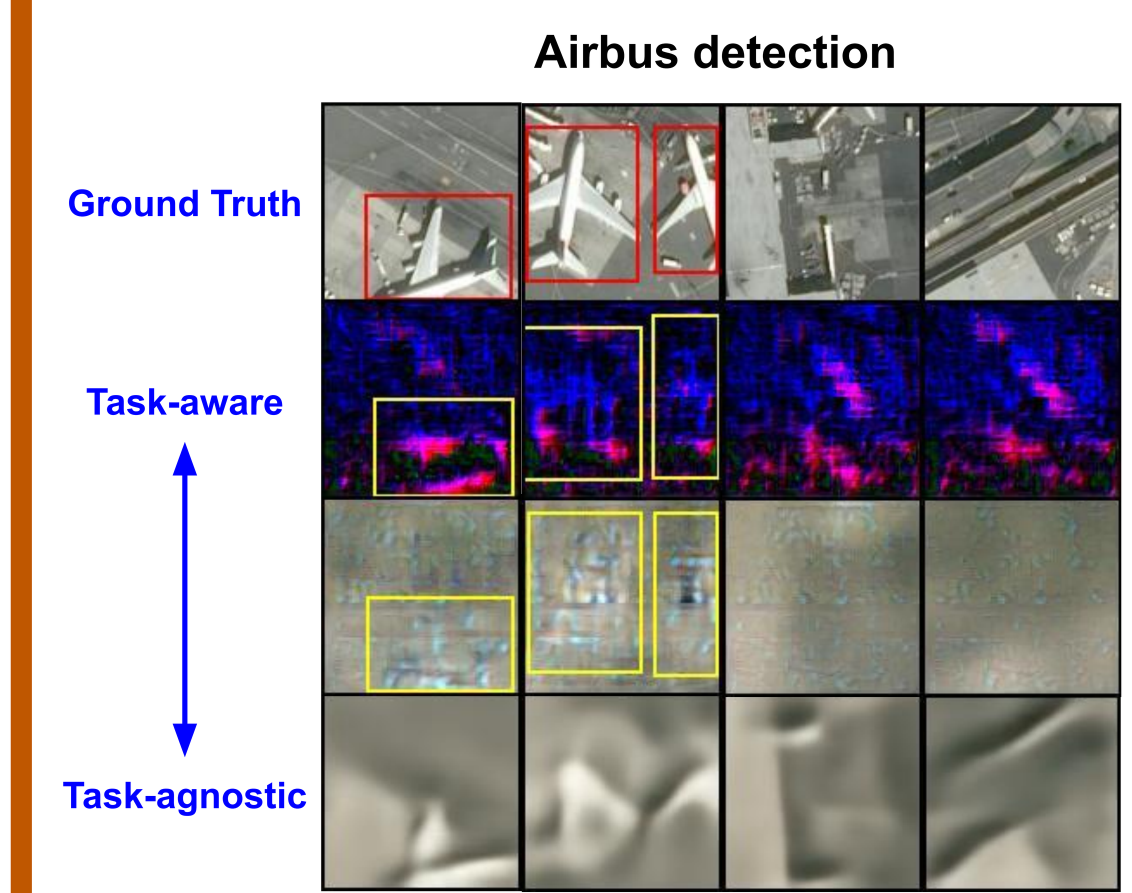

Figure: We can use a weighted reconstruction loss to trade off task-awareness and task-agnostic. Weighted task-aware images faintly reconstruct the original images while restoring task-relevant features with high-frequency noise.

Takeaways

- We design a data compression framework for distributed source coding of **correlated** sources in the presence of a task that adapts to any communication bottleneck with a **single model**, without the need for retraining.
- Using our method, we can measure the importance of the task and allocate bandwidth to sources correspondingly.