Deep Joint Transmission-Recognition for Multi-View Cameras

Ezgi Özyılkan, Mikolaj Jankowski

August 27, 2020

Table of Contents

Introduction and Methods

Introduction

We mainly analyze two different paradigms for the task of person classification at the wireless edge carried out by multi-view cameras:

- Digital (separate) transmission scheme
- JSCC approaches
  - Single User Schemes
  - Joint Decoding *only* – Multiple-Access Channel Schemes
  - Joint Encoding + Joint Decoding Schemes

Specifically, **overlapping** multi-view cameras.

Stringent bandwidth values will be considered.

Introduction

Although any differentiable channel model can be employed to train the JSCC approaches, we will consider an additive white Gaussian noise (AWGN) channel:

$$\mathbf{y} = \mathbf{x} + \mathbf{z} \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^B$ and $\mathbf{y} \in \mathbb{R}^B$ be the channel input, channel output vectors. $\mathbf{z} \in \mathbb{R}^B$ is the noise vector such that $z_i \sim \mathcal{N}(0, \sigma^2_{noise})$.

The channel SNR is then defined as:

$$\mathrm{SNR} = 10 \log_{10} \left(\frac{P}{\sigma^2_{noise}} \right) \quad (\mathrm{dB}). \tag{2}$$

Introduction

For every channel input vector, we impose an average power constraint of $P = 1$:

$$\frac{1}{B}\sum_{i=1}^{B} x_i^2 \leq P \tag{3}$$

To compare JSCC approaches with the digital method, Shannon capacity formula is used, that is:

$$C = \frac{1}{2}\log_2\left(1 + \frac{P}{\sigma_{noise}^2}\right), \tag{4}$$

Classification Baseline

Person classification is going to be achieved by a DNN that outputs a multi-hot encoded vector.

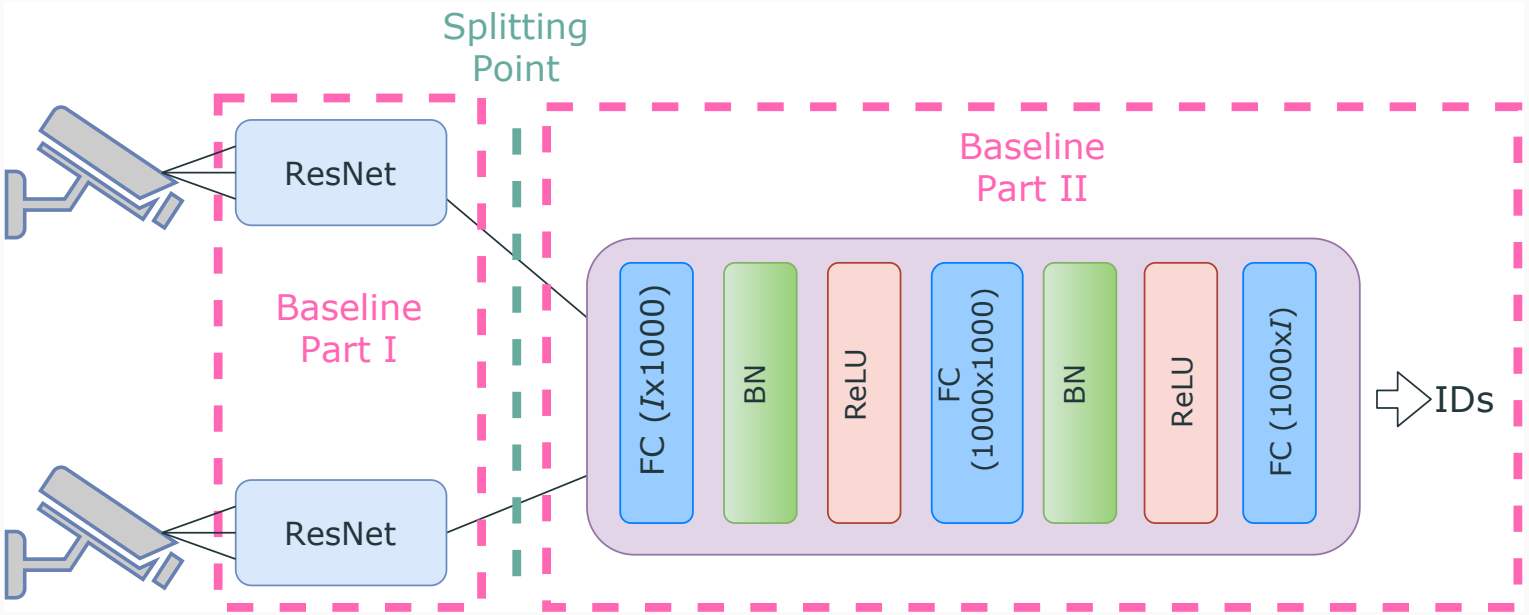


Figure 1: Proposed architecture for classification baseline. Fully-connected layer parameters are denoted as: input size $\times$ output size. For two-camera setting, I in the figure corresponds to $I = I_1 + I_2$, where $I_{\mathcal{K}}$ is equal to the number of the unique person IDs at Camera $\mathcal{K}$.

Digital (Separate) Transmission Scheme

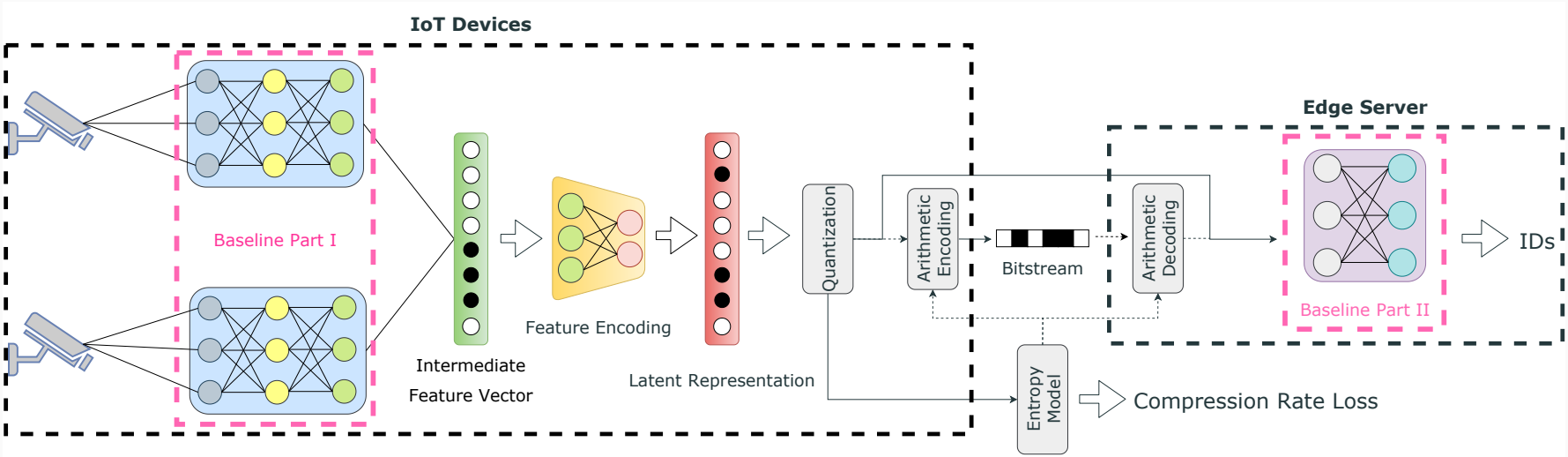


Figure 2: An overview of the digital (separate) transmission scheme. Arithmetic coding is only carried out during testing, as indicated with dashed arrows in the figure. For the sake of obtaining an upper bound on the performance, ideal channel coding scheme is assumed.

Digital (Separate) Transmission Scheme

To ensure that the quantization step is differentiable, we adopt the quantization noise from[1] during the training phase. We add the uniform noise to each element in the latent representation such as:

$$Q(\mathbf{r}) = \mathbf{r} + \mathcal{U}\left(-\frac{1}{2}, \frac{1}{2}\right), \tag{5}$$

where $Q(\cdot)$ is the approximated quantization, $\mathbf{r}$ is the latent representation, and $\mathcal{U}(\cdot,\cdot)$ is the uniform noise vector.

[1] R. M. Gray and D. L. Neuhoff. "Quantization". In: *IEEE Transactions on Information Theory* 44.6 (1998), pp. 2325–2383.

Digital (Separate) Transmission Scheme

The arithmetic coding we implement is based on estimating the distribution of the quantized latents. Assuming that the vector elements $q_i \in \mathbf{q} = Q(\mathbf{r})$ are i.i.d with some probability mass function (PMF) of $p(q)$, we first-order approximate $p(q)$ as a continuous-valued probability density function $p_c(q)$ as:

$$p_c(q) = \sum_{k=1}^{K} \alpha_k \frac{1}{\sigma_k \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{q-\mu_k}{\sigma_k}\right)^2}, \tag{6}$$

where K is the number of Gaussian mixtures, σ_k are mixture scales, μ_k are mean values, and α_k are the corresponding mixture weights.

Digital (Separate) Transmission Scheme

Finally, we evaluate PMF $p(q)$ at discrete values $q \in \mathbb{Z}$ by integrating $p_c(q)$ over $\left[q - \frac{1}{2}, q + \frac{1}{2}\right]$ in order to obtain:

$$p(q) = \int_{q-\frac{1}{2}}^{q+\frac{1}{2}} p_c(x)dx = F_c\left(q + \frac{1}{2}\right) - F_c\left(q - \frac{1}{2}\right), \tag{7}$$

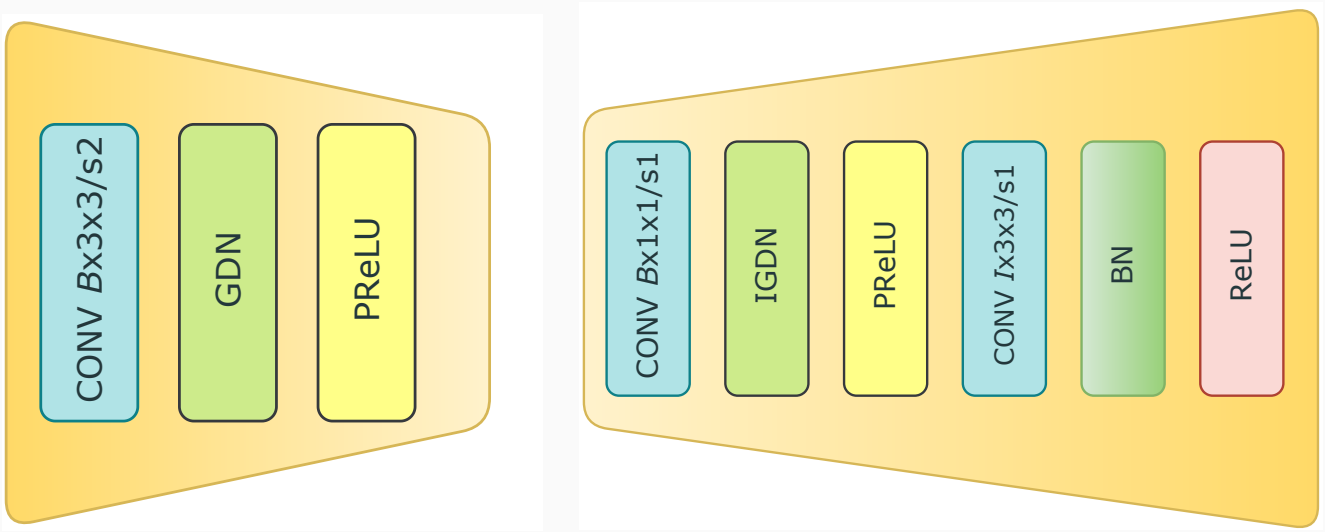
where F_c is the cumulative density function of the distribution $p_c(q)$.

Loss Function is defined as weighted sum of the two objectives, which we aim to minimize:

$$L = l_{\mathrm{cross-entropy}} - \lambda \cdot \log_2 p(\mathbf{q}), \tag{8}$$

where $l_{\mathrm{cross-entropy}}$ and $p(\mathbf{q})$ refer to cross-entropy loss between predicted IDs and ground truth for person classification task, and the PMF of the quantized vector $\mathbf{q}$ respectively.

JSCC Approaches – Autoencoder Architecture



(a) Encoder (b) Decoder

Figure 3: Proposed autoencoder architecture for the JSCC schemes. Convolutional layer parameters are denoted as: number of output channels × kernel height × kernel width / sampling stride (e.g. $/\mathrm{s}2$ means sampling with stride two). *B*, *I* and PReLU in the figure correspond to channel bandwidth, input dimension and parametric rectified linear unit (PReLU), respectively.

JSCC Approaches – Single User Schemes

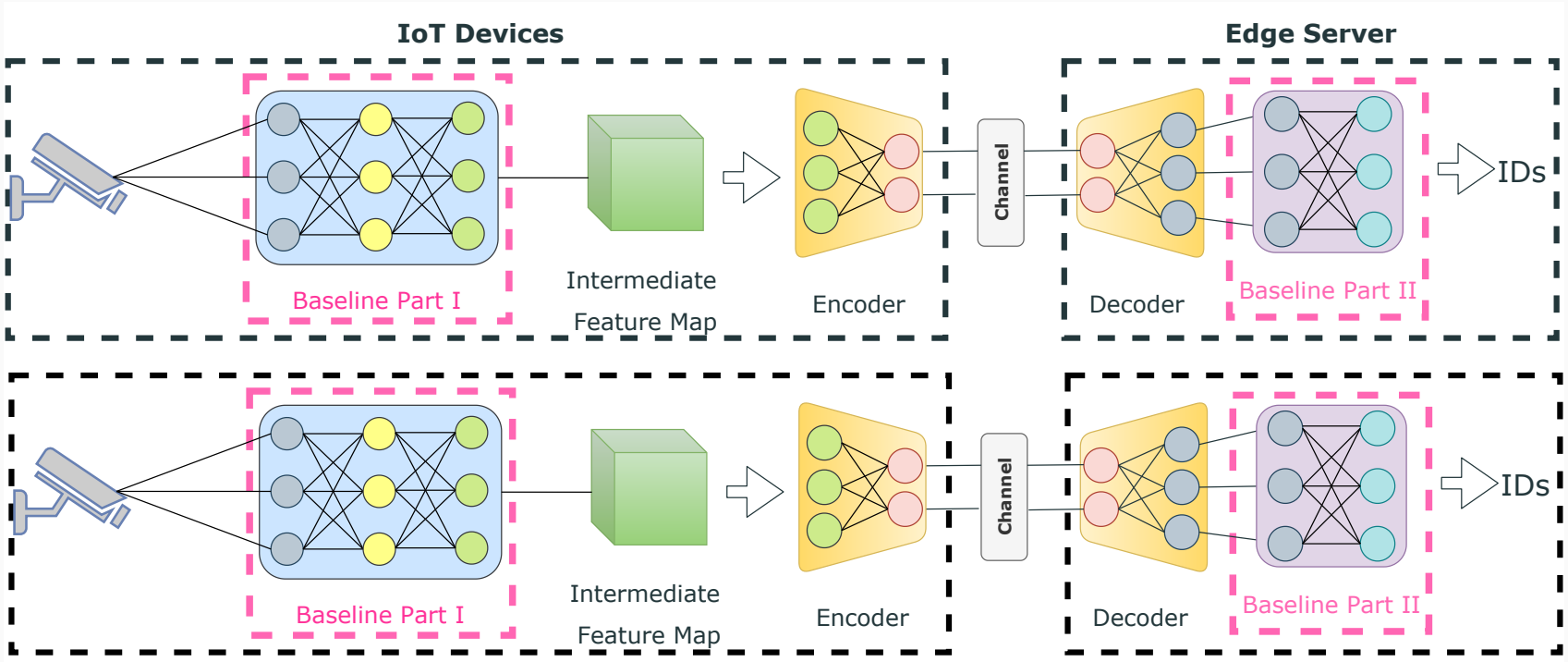


Figure 4: An overview of the single user schemes for both cameras. The total channel bandwidth for the single user schemes is calculated as $B = B_1 + B_2$, where $B_{\mathcal{K}}$ is the bandwidth allocated for Camera $\mathcal{K}$.

JSCC Approaches – Joint Decoding *only*

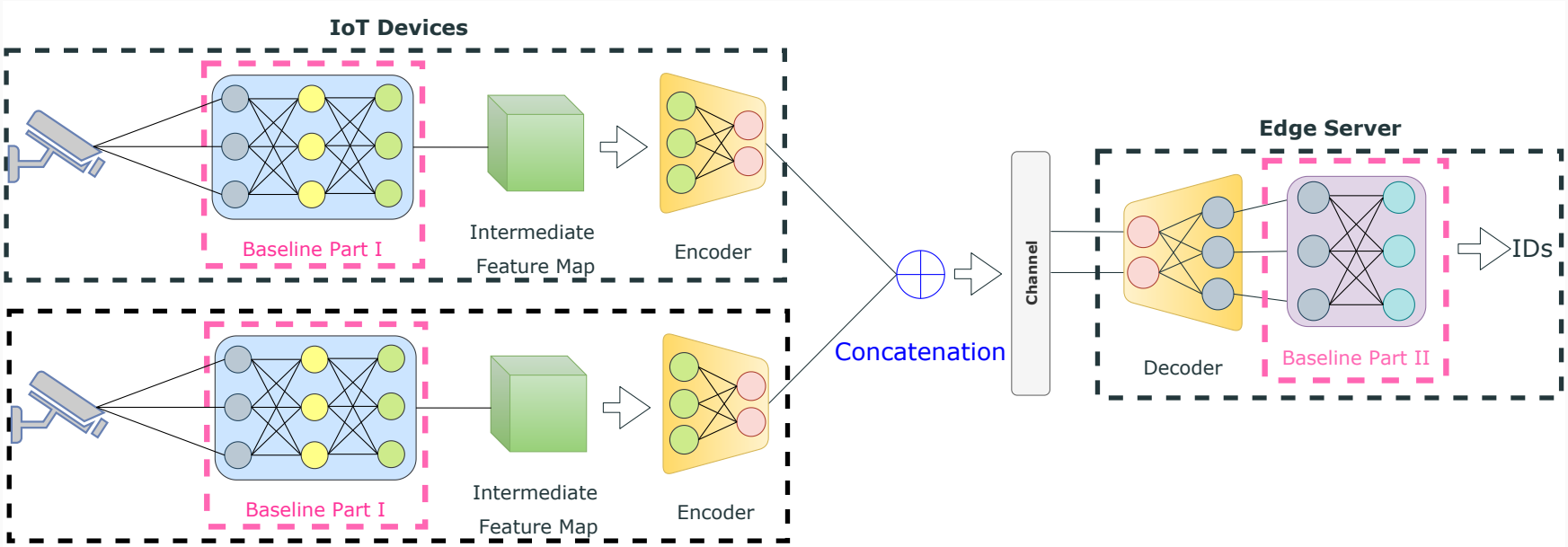
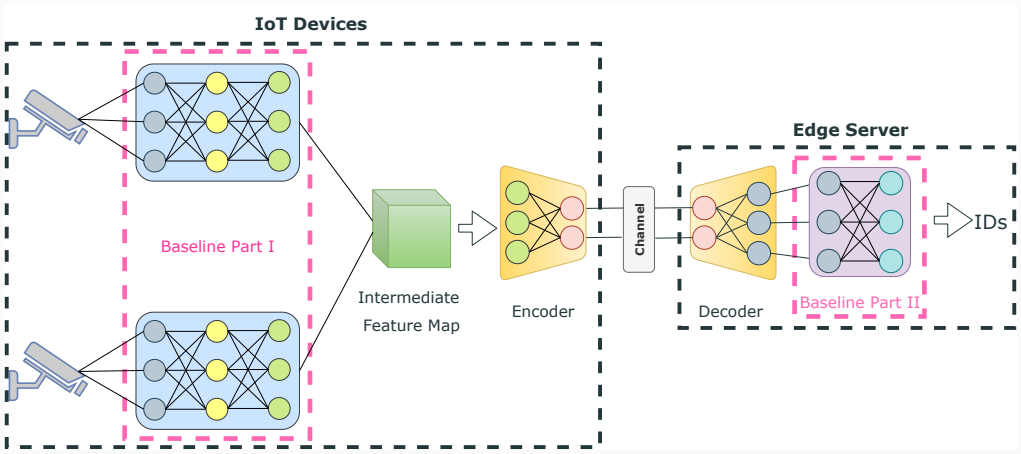
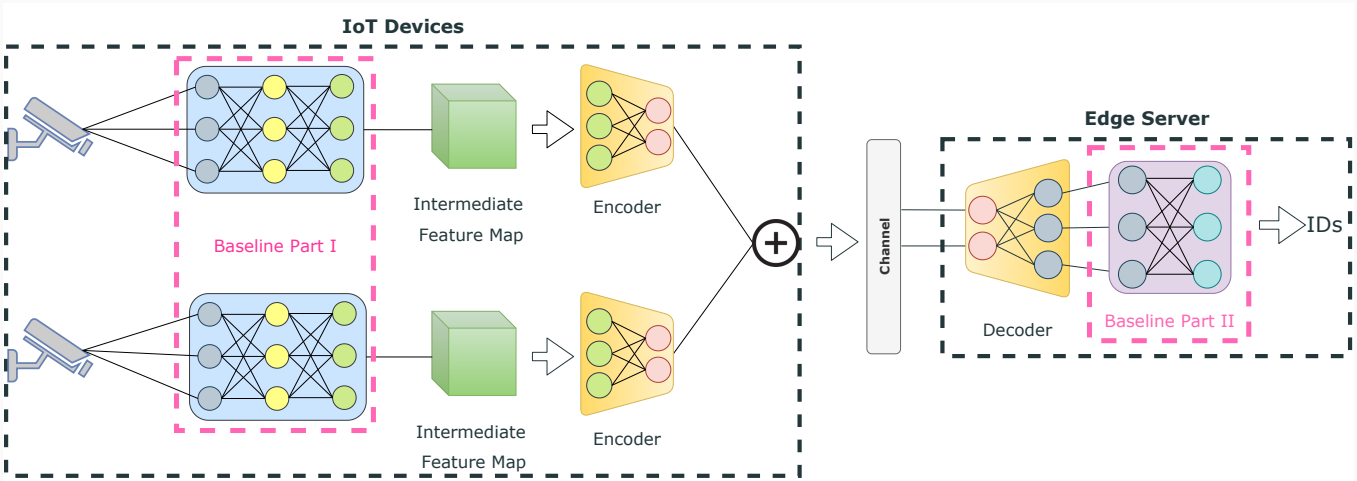


Figure 5: An overview of the evaluated joint decoding schemes. Note that for the *joint decoding + ortogonal transmission* scheme, the bandwidth allocation is set to be equal for both users. For the implementation of the *joint decoding + non-orthogonal tranmission* scheme, we replace the concatenation operation in the figure by the element-wise summation of the output of 2 encoders.

JSCC Approaches – Joint Decoding + Joint Encoding



(a)



(b)

Figure 6: An overview of the evaluated joint encoding schemes.

Training Strategy

Training Strategy for the JSCCs

1. Pretrain the classification baseline
2. Extract all possible intermediate feauture maps at the splitting point and pretrain the autoencoder
3. Train the entire network end-to-end

Training Strategy for the Digital (Separate) Scheme

1. Train the entire network end-to-end

Results

Experimental Setup

(a) Camera 1. **(b)** Camera 2. **(c)** Camera 3. **(d)** Camera 4.

(e) Camera 5. **(f)** Camera 6. **(g)** Camera 7.

Figure 7: Synchronized corresponding frames from 7 static cameras of the 'WILDTRACK' dataset[2].

[2] T. Chavdarova et al. "WILDTRACK: A Multi-camera HD Dataset for Dense Unscripted Pedestrian Detection". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, pp. 5030–5039.

Experimental Setup

Due to unbalanced nature of labels in the 'WILDTRACK' dataset, the loss function for the classification baseline and for the entire device-edge models is chosen to be weighted cross-entropy[3], which aims to optimize the metric:

$$\text{Balanced Accuracy} = \frac{\text{TPR} + \text{TNR}}{2}, \tag{9}$$

for all class predictions, where TPR and TNR stand for True Positive Rate and True Negative Rate, respectively[4].

---

[3] We used the PyTorch implementation of *`BCEWithLogitsLoss`*, which is documented at: *`https://pytorch.org/docs/stable/nn.html`*

[4] Note that

$$\text{TPR} = \frac{\text{TP}}{\text{P}} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \ \text{TNR} = \frac{\text{TN}}{\text{N}} = \frac{\text{TN}}{\text{TN} + \text{FP}}\ .$$

Experimental Setup

In order to make a fair comparison among the JSCC approaches, we define an evaluation metric for the single user scheme, $\mathrm{acc}_{\mathrm{sep}}$, for a given bandwidth budget of B as the following:

$$\mathrm{acc}_{\mathrm{sep}}(B) = \frac{1}{|E|}\sum_{e\in E}\max_{\substack{b_1,b_2\in\mathcal{B}\\ s.t. b_1+b_2=B}}(w_4\cdot\mathrm{acc}_{4,b_1}+w_5\cdot\mathrm{acc}_{5,b_2}), \quad (10)$$

$$\text{where } w_4=\frac{|ID_4|}{|ID_4|+|ID_5|}, w_5=\frac{|ID_5|}{|ID_4|+|ID_5|},$$

E : set of independent experiments,

$\mathcal{B}$: set of evaluated bandwidths,

$\mathrm{acc}_{k,b}$: mean balanced accuracy across all classes for Camera k at bandwidth b,

ID_k : set of all people appearing at Camera k.

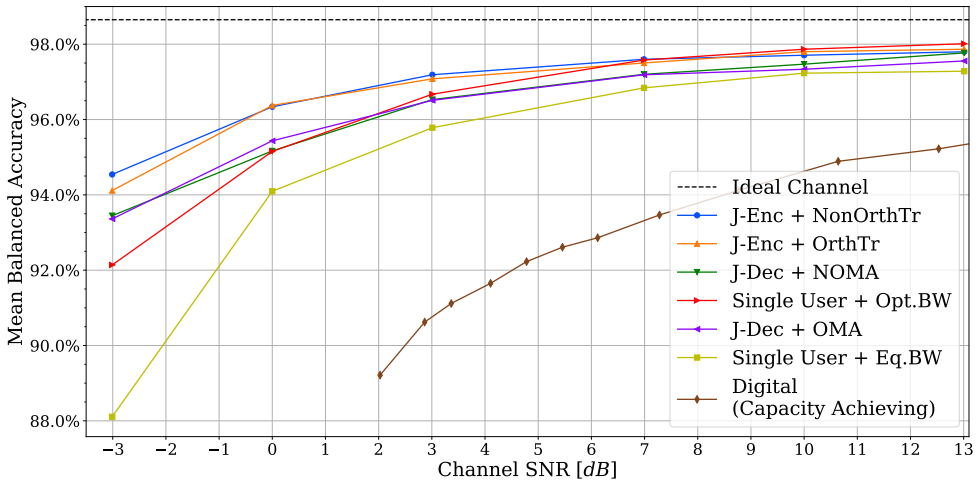
We refer to this metric as *Optimal Combination*.

Table of Definitions for the JSCCs

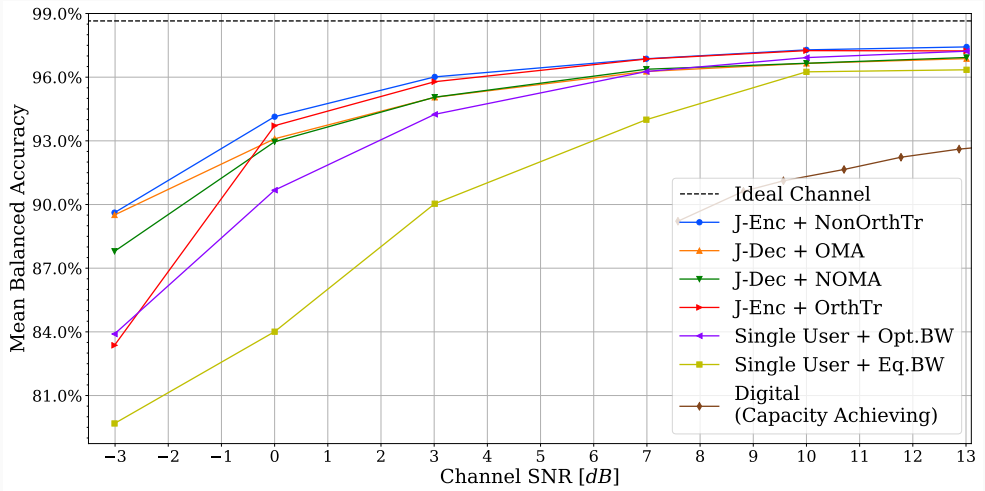
Short Name	JSCC Scheme
Single Users + Eq.BW	single-user schemes for both cameras
Single Users + Opt.BW	single-user schemes evaluated with *Optimal Combination*
J-Dec + OMA	joint decoding + orthogonal multiple access w\ equal power allocation for both users
J-Dec + NOMA	joint decoding + non-orthogonal multiple access w\ equal power allocation for both users
J-Enc + OrthTr	joint encoding + orthogonal transmission
J-Enc + NonOrthTr	joint encoding + non-orthogonal transmission

Table 1: Table of definitions for the evaluated JSCC approaches.

Performance for Different Methods



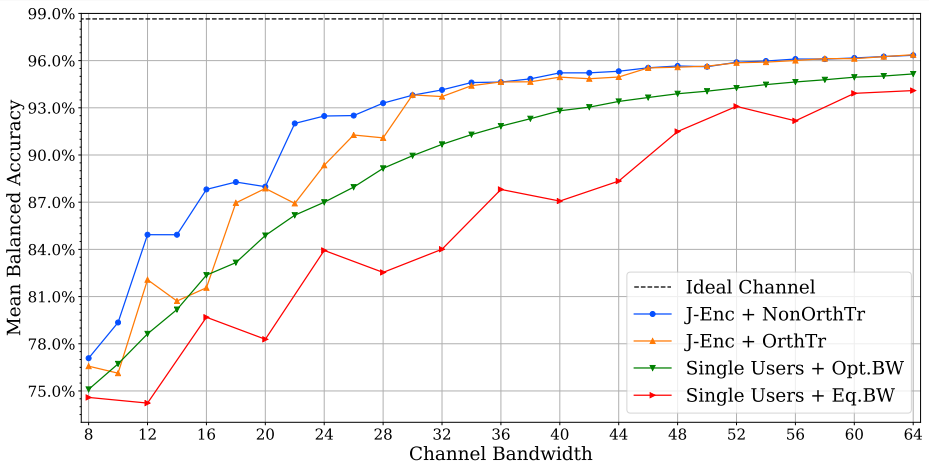
(a) $B = 64$



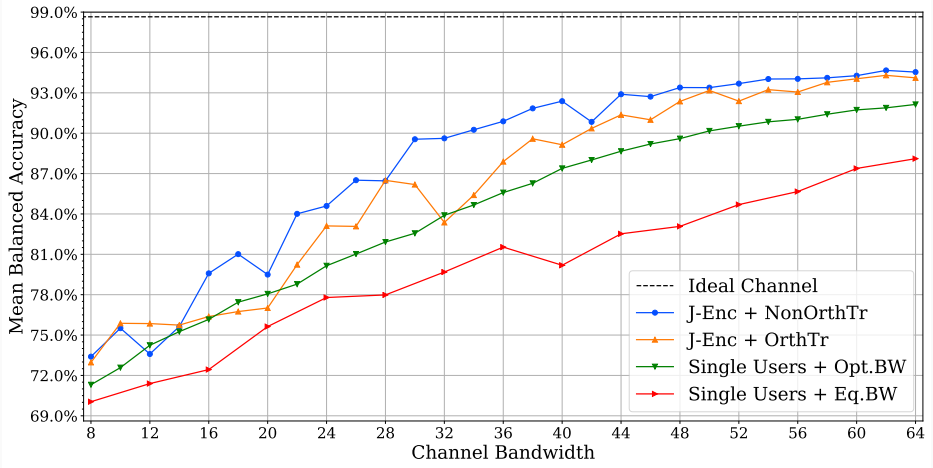
(b) $B = 32$

Figure 8: Comparison of performance for different approaches for $B \in \{32, 64\}$.

Performance for Different Channel SNRs – Joint Encoding Schemes



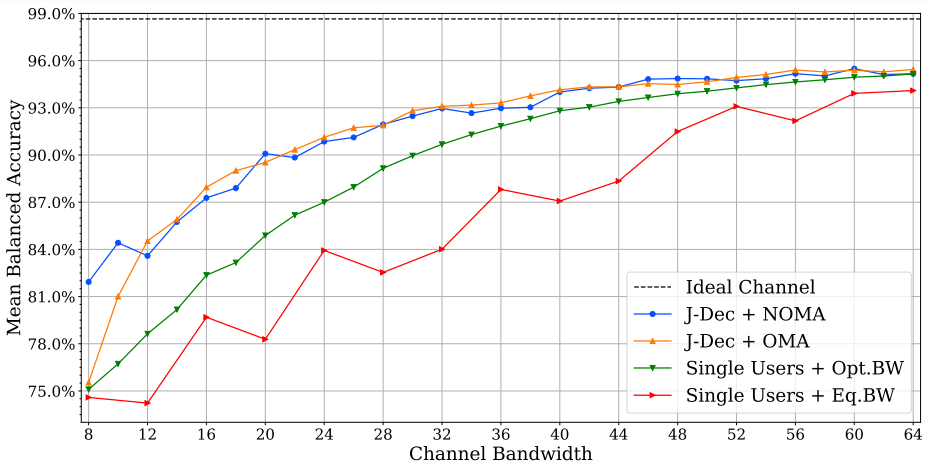
(a) $\mathrm{SNR}_{\mathrm{train}} = \mathrm{SNR}_{\mathrm{test}} = 0$ dB



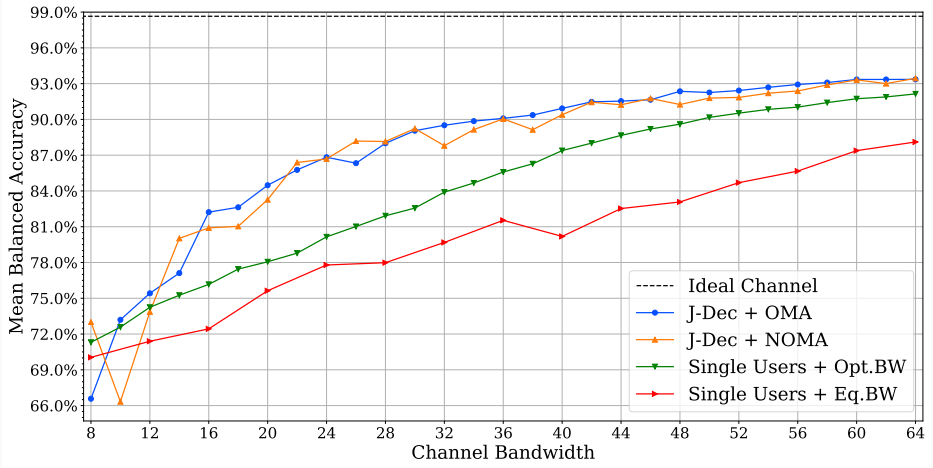
(b) $\mathrm{SNR}_{\mathrm{train}} = \mathrm{SNR}_{\mathrm{test}} = -3$ dB

Figure 9: Accuracy as a function of the channel bandwidth for $\mathrm{SNR} \in \{-3, 0\}$ dB.

Performance for Different Channel SNRs – Joint Decoding *Only* Schemes



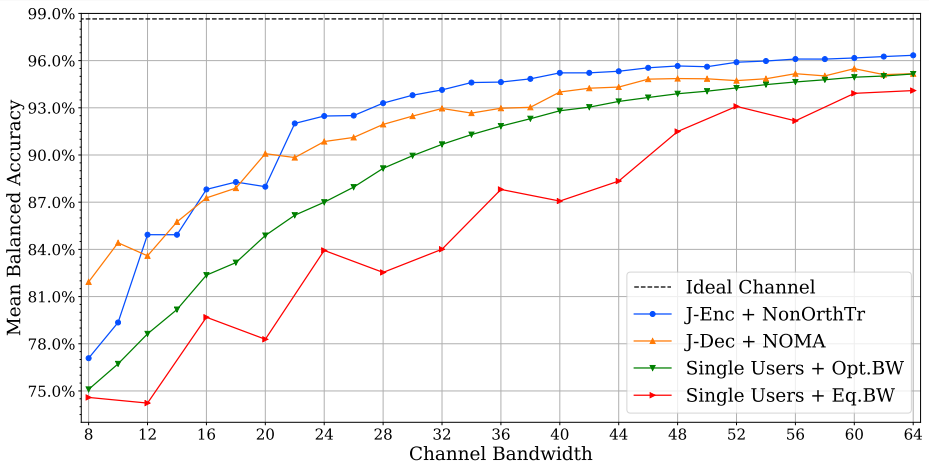
(a) $\mathrm{SNR}_{\mathrm{train}} = \mathrm{SNR}_{\mathrm{test}} = 0$ dB



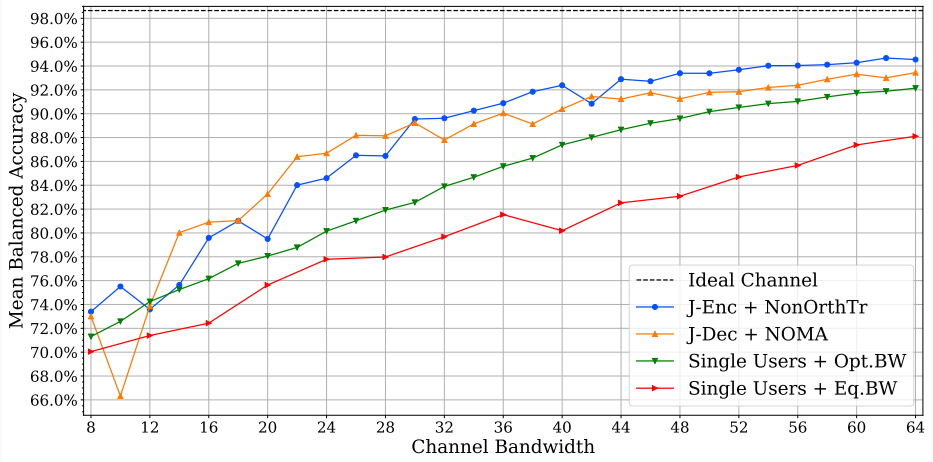
(b) $\mathrm{SNR}_{\mathrm{train}} = \mathrm{SNR}_{\mathrm{test}} = -3$ dB

Figure 10: Accuracy as a function of the channel bandwidth for $\mathrm{SNR} \in \{-3, 0\}$ dB.

Performance for Different Channel SNRs – ”J-Dec + NOMA” vs. ”J-Enc + NonOrhTr”



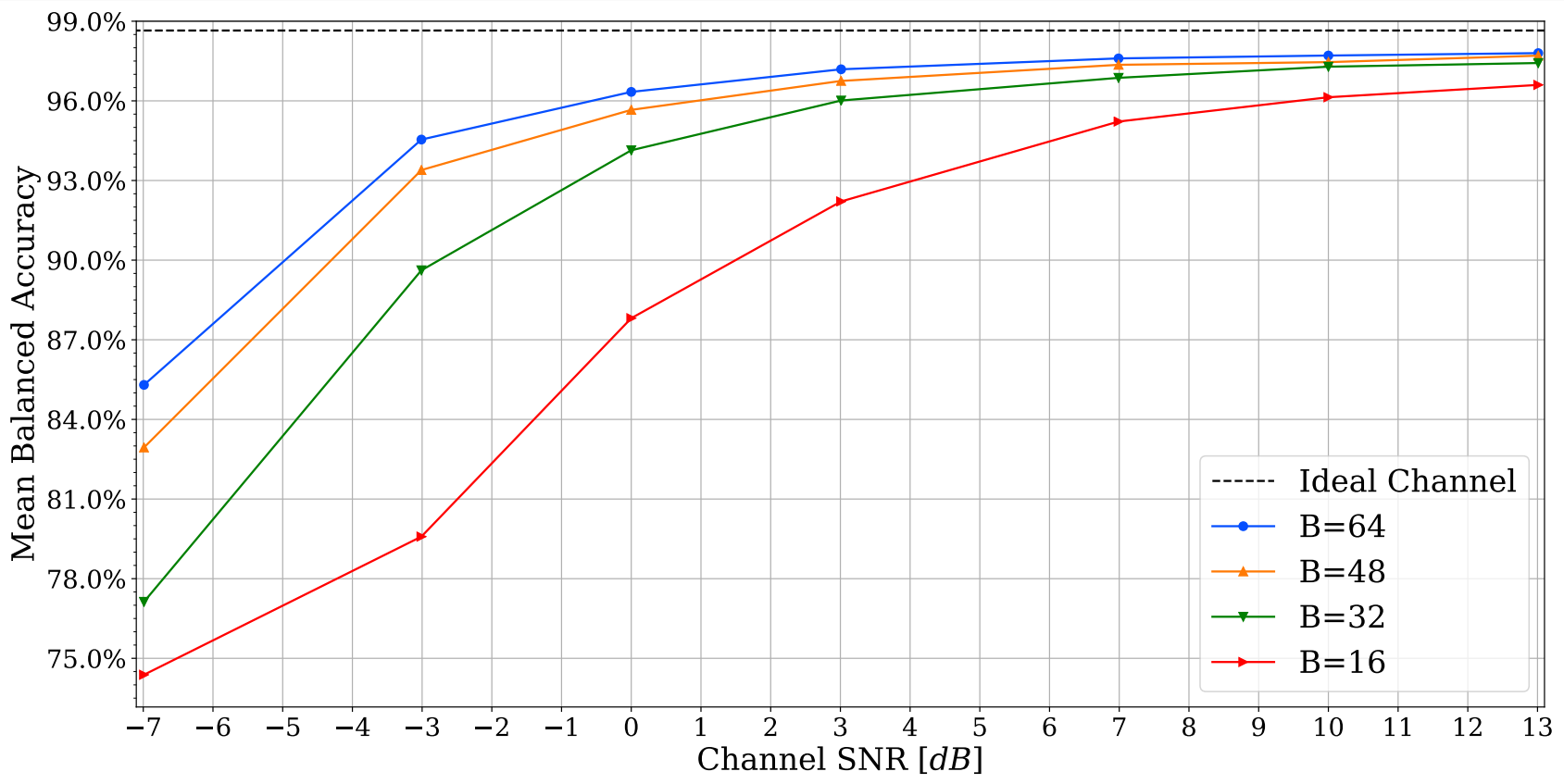
(a) $\mathrm{SNR}_{\mathrm{train}} = \mathrm{SNR}_{\mathrm{test}} = 0$ dB



(b) $\mathrm{SNR}_{\mathrm{train}} = \mathrm{SNR}_{\mathrm{test}} = -3$ dB

Figure 11: Accuracy as a function of the channel bandwidth for $\mathrm{SNR} \in \{-3, 0\}$ dB.

Performance for Different Channel Bandwidths for "J-Enc + NonOrthTr"



(a) Accuracy as a function of the channel SNR for $B \in \{16, 32, 48, 64\}$. The more bandwidth is allocated for the JSCC, the more robust it becomes against the channel noise.

Next Steps and Conclusion

Next Steps and Conclusion

Short Name	JSCC Scheme
Single Users + Eq.BW	single-user schemes for both cameras
Single Users + Opt.BW	single-user schemes evaluated with *Optimal Combination*
J-Dec + OMA1	joint decoding + orthogonal multiple access w\ equal power allocation for both users
J-Dec + NOMA1	joint decoding + non-orthogonal multiple access w\ equal power allocation for both users
J-Dec + OMA2	joint decoding + orthogonal multiple access w\ learned power allocation for both users
J-Dec + NOMA2	joint decoding + non-orthogonal multiple access w\ learned power allocation for both users
J-Enc + OrthTr	joint encoding + orthogonal transmission
J-Enc + NonOrthTr	joint encoding + non-orthogonal transmission

Table 2: Table of definitions for the JSCC approaches.

- Wrap up the results!

Questions?