Semantic Communication for Task Execution and Data Reconstruction in Multi-View Scenarios

Maximilian H. V. Tillmann, *Graduate Student Member, IEEE,* Avinash Kankari, Carsten Bockelmann, *Member, IEEE,* Armin Dekorsy, *Senior Member, IEEE*

***Abstract*—Semantic communication has gained significant attention with the advances in machine learning. Most semantic communication works focus on either task execution or data reconstruction, with some recent works combining the two. In this work, we propose a semantic communication system for concurrent task execution and data reconstruction for a multi-view scenario, which we formulate as the maximization of mutual information. To investigate the trade-off between the two objectives, we formulate a joint objective as a convex combination of task execution and data reconstruction. We show that under specific assumptions, the Structural Similarity Index Measure (SSIM) loss can be obtained from the mutual information maximization objective for data reconstruction, which takes human visual perception into account. Furthermore, for constant resource use, we show that by increasing the weight of the reconstruction objective up to a certain point, the task execution performance can be kept nearly constant, while the data reconstruction can be significantly improved.**

***Index Terms*—Semantic communication, muti-view, infomax, deep learning, task execution and data reconstruction.**

I. Introduction

With the recent advances of deep learning methods, semantic communication is expected to play a key role in the development of future communication systems of 6G and beyond [1]–[3]. Unlike Shannon's goal of accurately transmitting every bit of message, semantic communication has the goal of accurately conveying the intended meaning behind a message [4].

Most works on semantic communication can be grouped into one of the two objectives: (i) Task execution, where a single or multiple agents observe some data, with the goal to execute a task at a remote location, meaning that all observation data is not required to be transmitted [1], [5], [6]. (ii) Data reconstruction, where the goal is to fully reconstruct all data, e.g., text, speech, or image data, according to a semantic accuracy measure [1], [7].

However, in many scenarios, it might be advantageous if both task execution and data reconstruction are possible. One such example is shown in Fig. 1 with error detection in a production line. If an error is missed and only identified at the end of the production line, data reconstruction is required so that a human expert can review the sensor data to localize the fault. Furthermore, if the task changes or new data processing methods are to be applied to the original data in the future, data reconstruction alongside task execution is required if retransmission is not feasible or not desired.

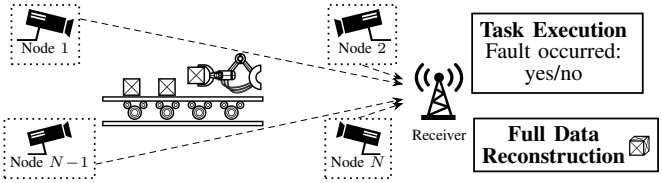

Fig. 1: Multi-view semantic communication with N sensing nodes transmitting observations to a receiver for joint task execution (e.g., fault detection) and data reconstruction.

Such scenarios of joint task execution and data reconstruction were only recently investigated in [8] and [9], where it was shown that if the communication system is only optimized for task execution, reconstruction is only possible in a limited way, and vice versa.

In [8] the authors focused on the scenario of image classification and image reconstruction under training data label corruption. They formulated a coding rate reduction maximization problem to achieve robustness to training data label corruption, and used the Mean Square Error (MSE) loss for the image reconstruction. Other loss functions for image reconstruction, like Structural Similarity Index Measure (SSIM), which is a measure of image similarity aligned for human visual perception [10], were not considered in their proposed approach. A system with SSIM as the loss function served only as a baseline, and the proposed methods were not evaluated for SSIM. In [9], the authors investigated joint image reconstruction and image segmentation. They used the cross entropy loss for image segmentation and the MSE loss for image reconstruction, again without considering other loss functions for image reconstruction like SSIM.

In contrast to these existing works, we focus on extending the image reconstruction problem beyond the MSE loss. To this end, we propose a unified framework for semantic communication system design, in which both image reconstruction and image classification are formulated as a mutual information maximization problem. We show how different image reconstruction losses (MSE, SSIM, or a combination of both) can be derived from the mutual information maximization problem under different assumptions on the decoder Probability Density Function (PDF). Furthermore, we evaluate the trade-off between image reconstruction quality, and task execution error rate for our proposed method.

This work was partially supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC-3036 The Martian Mindset, project number: 533607631, by the DFG under project number 500260669 (SCIL), and by the German Ministry of Research, Technology and Space (BMFTR) under project number 16KISK016 (Open6GGub). The authors are with the Department of Communications Engineering, University of Bremen, 28359 Bremen, Germany. E-mails: {tillmann, bockelmann, dekorsy}@ant.uni-bremen.de

DOI: 10.1109/LWC.2026.3712239

Finally, we mention that in contrast to the existing methods of joint task execution and data reconstruction, we consider the distributed multi-view scenario for classification. Each sensing node observes only part of the data, requiring extraction of task-relevant semantics for classification at the decoder. Unlike the centralized case, optimal classification cannot be achieved by any single sensing node alone as information from other views may be required.

Based on the above discussions, the main contributions are summarized as follows:

- We propose a semantic communication system for joint task execution and data reconstruction in multi-view scenarios, formulated as a mutual information maximization problem.
- We show the derivation of SSIM loss from the mutual information maximization problem.
- We analyze the trade-off between task execution and data reconstruction. Increasing the reconstruction weight improves reconstruction with little impact on task performance, but overly weighting on reconstruction degrades task execution.

II. System Model and Problem Formulation

A. Semantic Communication System Model

Our distributed semantic communication system model is shown in Fig. 2. We assume a semantic source as the joint probability distribution $p(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N, \boldsymbol{z})$ of the observations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N$ and the semantic variable $\boldsymbol{z}$. The goal at the receiver is to reconstruct both the semantic variable, as well as all observations, which are denoted by $\hat{\boldsymbol{z}}$ and $\hat{\boldsymbol{s}} = \left[\hat{\boldsymbol{s}}_1^\top, \ldots, \hat{\boldsymbol{s}}_N^\top\right]^\top$, respectively. We use one encoder per sensing node to optimally use the channel resources for the joint objective of task execution and data reconstruction. On the receiver side, we use separate decoders for the two objectives, as a design where the decoder first reconstructs the data and then uses this data for task execution cannot improve performance due to the data processing inequality [11].

B. Problem Formulation for Joint Task Execution and Data Reconstruction

We formulate our optimization problem as the convex combination of two mutual information terms that should be maximized to balance the trade-off between the two objectives. We design the encoders $p_{\boldsymbol{\theta}_i}(\boldsymbol{c}_i|\boldsymbol{s}_i)$ with Neural Network (NN) parameters $\boldsymbol{\theta}_i$, such that the received symbols $\boldsymbol{y}$ are most informative about the semantic variable $\boldsymbol{z}$ for task execution with weight $1-\alpha$, and about the observations $\boldsymbol{s}$ for data reconstruction with weight $\alpha \in [0,1]$. As shown in Fig. 2, $\boldsymbol{y}$ depends on $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$ via $\boldsymbol{y} \sim p(\boldsymbol{y}|\boldsymbol{c}_1, \ldots, \boldsymbol{c}_N)$ and $\boldsymbol{c}_i \sim p_{\boldsymbol{\theta}_i}(\boldsymbol{c}_i|\boldsymbol{s}_i)$. For simplicity, we do not make this dependency explicit in our notation. We optimize the NN parameters $\boldsymbol{\theta}_i$ of the encoders distributions $p_{\boldsymbol{\theta}_i}(\boldsymbol{c}_i|\boldsymbol{s}_i)$ as

$$\begin{aligned} \underset{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_N}{\operatorname{argmax}} \quad & \alpha I(\boldsymbol{s};\boldsymbol{y}) + (1-\alpha) I(\boldsymbol{z};\boldsymbol{y}) \\ \text{s.t.} \quad & \boldsymbol{c}_i \in \mathbb{R}^{N_{\text{Tx}}/N},\, P_i \leq 1, \text{ for } i = 1, \ldots, N, \end{aligned} \tag{1}$$

where $\boldsymbol{s} = \left[\boldsymbol{s}_1^\top, \ldots, \boldsymbol{s}_N^\top\right]^\top$, N_{Tx} is the number of total channel uses with $\boldsymbol{y} \in \mathbb{R}^{N_{\text{Tx}}}$, and $P_i = 1/(N_{\text{Tx}}/N) \sum_{n=1}^{N_{\text{Tx}}/N} c_{i_n}^2$ is the power constraint for node i per channel use, where c_{i_n} is the n-th element of $\boldsymbol{c}_i$. The mutual information between $\boldsymbol{z}$ and $\boldsymbol{y}$ [11] is given as $I(\boldsymbol{z};\boldsymbol{y}) = \mathrm{E}_{p(\boldsymbol{z},\boldsymbol{y})}\left[\log\left(\frac{p(\boldsymbol{z}|\boldsymbol{y})}{p(\boldsymbol{z})}\right)\right]$, where $\mathrm{E}_{p(x)}[\cdot]$ is the expectation with respect to $x \sim p(x)$, and $\log(\cdot)$ is the natural logarithm. However, the posterior $p(\boldsymbol{z}|\boldsymbol{y})$ is typically not available, making it infeasible to optimize the encoder for a $\boldsymbol{y}$ that maximizes the mutual information terms [6]. A common approach is to use variational approximations of the posterior, $q_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{y})$ with parameters $\boldsymbol{\psi}$ and $q_{\boldsymbol{\phi}}(\boldsymbol{s}|\boldsymbol{y})$ with parameters $\boldsymbol{\phi}$ for the semantic decoder and the reconstruction decoder, respectively. This way, a lower bound of the mutual information can be derived that can be optimized instead. The following derivations are only shown for $I(\boldsymbol{z};\boldsymbol{y})$, but can be analogously obtained for $I(\boldsymbol{s};\boldsymbol{y})$ by replacing $\boldsymbol{z}$ by $\boldsymbol{s}$.

$$\begin{aligned} I(\boldsymbol{z};\boldsymbol{y}) &= \mathrm{E}_{p(\boldsymbol{z},\boldsymbol{y})}\left[\log\left(\frac{p(\boldsymbol{z}|\boldsymbol{y})}{p(\boldsymbol{z})}\frac{q_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{y})}{q_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{y})}\right)\right] && (2)\\ &= \mathrm{E}_{p(\boldsymbol{y})}\left[\mathrm{E}_{p(\boldsymbol{z}|\boldsymbol{y})}\left[\log\left(q_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{y})\right)\right]\right] \\ &\quad - \mathrm{E}_{p(\boldsymbol{z})}[\log\left(p(\boldsymbol{z})\right)] + \mathrm{E}_{p(\boldsymbol{z},\boldsymbol{y})}\left[\log\left(\frac{p(\boldsymbol{z}|\boldsymbol{y})}{q_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{y})}\right)\right] && (3)\\ &= -\mathrm{E}_{p(\boldsymbol{y})}\left[\mathrm{H}\left(p(\boldsymbol{z}|\boldsymbol{y}), q_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{y})\right)\right] \\ &\quad + \mathrm{H}(\boldsymbol{z}) + \mathrm{E}_{p(\boldsymbol{y})}\left[D_{\mathrm{KL}}\left(p(\boldsymbol{z}|\boldsymbol{y})||q_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{y})\right)\right] && (4)\\ &\geq -\mathrm{E}_{p(\boldsymbol{y})}\left[\mathrm{H}\left(p(\boldsymbol{z}|\boldsymbol{y}), q_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{y})\right)\right] + \mathrm{H}(\boldsymbol{z}), && (5) \end{aligned}$$

where $\mathrm{H}(p,q)$ is the cross-entropy between p and q, $\mathrm{H}(\cdot)$ is the entropy, and $D_{\mathrm{KL}}(p,q)$ is the Kullback-Leibler (KL) divergence from q to p. The last step uses that the KL divergence is non-negative [11]. From (4) and (5), it can be seen that the lower bound on the mutual information is tight if the KL divergence is zero, meaning that the true posterior $p(\boldsymbol{z}|\boldsymbol{y})$ and its variational approximation $q_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{y})$ are identical almost everywhere [11].

The objective in (1) now becomes the minimization of the mutual information lower bound (5). With the respective weights of α and $1-\alpha$ we have

$$\begin{aligned} &-\alpha\mathrm{E}_{p(\boldsymbol{y})}\left[\mathrm{H}\left(p(\boldsymbol{s}|\boldsymbol{y}), q_{\boldsymbol{\phi}}(\boldsymbol{s}|\boldsymbol{y})\right)\right] + \alpha\mathrm{H}(\boldsymbol{s}) \qquad (6)\\ &-(1-\alpha)\mathrm{E}_{p(\boldsymbol{y})}\left[\mathrm{H}\left(p(\boldsymbol{z}|\boldsymbol{y}), q_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{y})\right)\right] + (1-\alpha)\mathrm{H}(\boldsymbol{z}), \end{aligned}$$

which, by applying (4), can be rewritten as

$$\begin{aligned} &\underbrace{\alpha I(\boldsymbol{s};\boldsymbol{y}) + (1-\alpha) I(\boldsymbol{z};\boldsymbol{y})}_{\text{weighted encoder objective}} - \alpha \underbrace{\mathrm{E}_{p(\boldsymbol{y})}\left[D_{\mathrm{KL}}\left(p(\boldsymbol{s}|\boldsymbol{y})||q_{\boldsymbol{\phi}}(\boldsymbol{s}|\boldsymbol{y})\right)\right]}_{\text{Reconstruction decoder objective}} \\ &- (1-\alpha)\underbrace{\mathrm{E}_{p(\boldsymbol{y})}\left[D_{\mathrm{KL}}\left(p(\boldsymbol{z}|\boldsymbol{y})||q_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{y})\right)\right]}_{\text{Semantic decoder objective}}, \end{aligned} \tag{7}$$

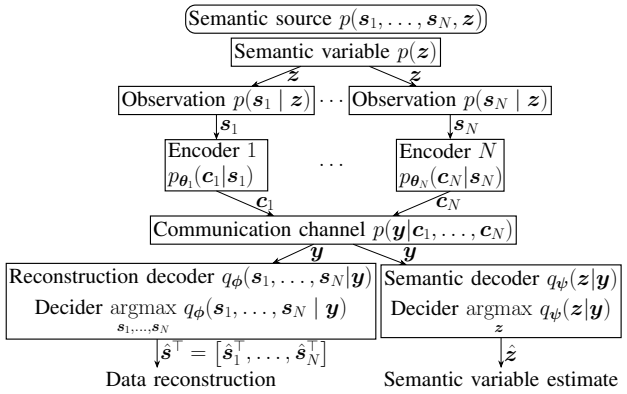

Fig. 2: Multi-view semantic communication system model for joint task execution and data reconstruction.

which shows that minimizing the weighted mutual information lower bound implicitly optimizes encoder and decoder objectives. The objective for the encoders is to maximize the weighted mutual information for the reconstruction and the task execution. The objectives for the two decoders are minimizing the KL divergence between the true decoder and the variational approximation, which are independent of α, as $q_{\boldsymbol{\phi}}(\boldsymbol{s}|\boldsymbol{y})$ and $q_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{y})$ are independent. Finally, as the entropies $\mathrm{H}(\boldsymbol{z})$ and $\mathrm{H}(\boldsymbol{s})$ do not depend on any encoder or decoder parameters, our optimization problem is the minimization of the weighted cross-entropy terms with respect to the NN parameters $\{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_N,\boldsymbol{\phi},\boldsymbol{\psi}\}$ given as

$$\begin{aligned}\underset{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_N,\boldsymbol{\phi},\boldsymbol{\psi}}{\operatorname{argmin}} \quad & \alpha \mathrm{E}_{p(\boldsymbol{y})}\left[\mathrm{H}\left(p(\boldsymbol{s}|\boldsymbol{y}),q_{\boldsymbol{\phi}}(\boldsymbol{s}|\boldsymbol{y})\right)\right] \\ & +(1-\alpha)\mathrm{E}_{p(\boldsymbol{y})}\left[\mathrm{H}\left(p(\boldsymbol{z}|\boldsymbol{y}),q_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{y})\right)\right] \\ \text{s.t.} \quad & \boldsymbol{c}_i \in \mathbb{R}^{N_{\mathrm{Tx}}/N},\ P_i \leq 1,\ \text{for } i=1,\ldots,N.\end{aligned} \tag{8}$$

III. Solving the Optimization Problem

For our goal of solving (8), we iteratively update the NN parameters $\{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_N,\boldsymbol{\phi},\boldsymbol{\psi}\}$ using stochastic gradient descent based optimization with the reparametrization trick [6], [12]. The gradients of the cross-entropy terms of the optimization objective are obtained from the training data samples using Monte-Carlo approximation.

To optimize the cross-entropy of the reconstruction decoder $q_{\boldsymbol{\phi}}(\boldsymbol{s}|\boldsymbol{y})$ for image data, we have the problem of high dimensionality. Even for small color images of size $32\times32\times3$ with 8 bit resolution, e.g., as in the CIFAR-10 dataset [13], this means that there are $2^{8\cdot32\cdot32\cdot3}$ discrete probability events of $q_{\boldsymbol{\phi}}(\boldsymbol{s}|\boldsymbol{y})$, which are infeasible to estimate. Therefore, $q_{\boldsymbol{\phi}}(\boldsymbol{s}|\boldsymbol{y})$ needs to be simplified. A simple way to reduce the complexity is to approximate the discrete $q_{\boldsymbol{\phi}}(\boldsymbol{s}|\boldsymbol{y})$ by a parameterized continuous distribution with a tractable number of parameters.

In the following, we assume two different parameterized continuous distributions. We show how the MSE and SSIM losses are derived from the cross-entropy under the assumed distributions $q'_{\boldsymbol{\phi}}(\boldsymbol{s}|\boldsymbol{y})$ and $q''_{\boldsymbol{\phi}}(\boldsymbol{s}|\boldsymbol{y})$, respectively.

A. Deriving the MSE Loss

We show that the MSE loss is derived from the cross-entropy under the assumption that all $s_1,\ldots,s_L$ are independently Gaussian distributed with fixed and equal variances, where s_n is the n-th element of $\boldsymbol{s}\in\mathbb{R}^L$ [12].

We call this distribution $q'_{\boldsymbol{\phi}}(\boldsymbol{s}|\boldsymbol{y})$, which we model with a NN with NN parameters $\boldsymbol{\phi}$. The input to the NN is $\boldsymbol{y}$ and the output is $\boldsymbol{\mu}(\boldsymbol{y})$, which depends on $\boldsymbol{y}$, and where $\mu_n(\boldsymbol{y})$ is the n-th element of $\boldsymbol{\mu}(\boldsymbol{y})\in\mathbb{R}^L$. In our notation, we omit the dependence of $\boldsymbol{\mu}(\boldsymbol{y})$ from the NN parameters $\boldsymbol{\phi}$ for simplicity.

With the above assumptions and a given variance σ^2, $q'_{\boldsymbol{\phi}}(\boldsymbol{s}|\boldsymbol{y})$ is fully described by the mean $\boldsymbol{\mu}(\boldsymbol{y})$ of the Gaussian distribution. The cross-entropy then simplifies to

$$\begin{aligned}&-\mathrm{E}_{p(\boldsymbol{y})}\left[\mathrm{E}_{p(\boldsymbol{s}|\boldsymbol{y})}\left[\log\left(q'_{\boldsymbol{\phi}}(\boldsymbol{s}|\boldsymbol{y})\right)\right]\right] \\ &=-\mathrm{E}_{p(\boldsymbol{y})}\left[\mathrm{E}_{p(\boldsymbol{s}|\boldsymbol{y})}\left[\log\left(\prod_{n=1}^{L}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{\left(s_n-\mu_n(\boldsymbol{y})\right)^2}{2\sigma^2}\right)\right)\right]\right] \qquad (9)\\ &=\mathrm{E}_{p(\boldsymbol{y})}\left[\mathrm{E}_{p(\boldsymbol{s}|\boldsymbol{y})}\left[\sum_{n=1}^{L}\left(\frac{\left(s_n-\mu_n(\boldsymbol{y})\right)^2}{2\sigma^2}\right)\right]\right]+\frac{L}{2}\log\left(2\pi\sigma^2\right). \qquad (10)\end{aligned}$$

To minimize (10) with fixed variance σ^2, we have to minimize $\mathrm{E}_{p(\boldsymbol{y})}\left[\mathrm{E}_{p(\boldsymbol{s}|\boldsymbol{y})}\left[\mathrm{MSE}(\boldsymbol{s},\boldsymbol{\mu}(\boldsymbol{y}))\right]\right]$, with $\mathrm{MSE}(\boldsymbol{s},\boldsymbol{\mu}(\boldsymbol{y}))=\frac{1}{L}\sum_{n=1}^{L}\left(s_n-\mu_n(\boldsymbol{y})\right)^2$, which we approximate using the training data samples [12]. Finally, using the maximum likelihood criterion to get the estimate $\hat{\boldsymbol{s}}$, we have $\hat{\boldsymbol{s}}=\boldsymbol{\mu}(\boldsymbol{y})$, as we assumed $q'_{\boldsymbol{\phi}}(\boldsymbol{s}|\boldsymbol{y})$ to be Gaussian, whose maximum occurs at $\boldsymbol{\mu}(\boldsymbol{y})$.

B. Deriving the SSIM Loss

We show that the SSIM loss is derived from the cross-entropy by assuming a certain parameterized distribution for the decoder. The SSIM between two images is defined as a weighted average over M windows, where the weight associated with the n-th pixel of window i is denoted by w_{i_n} with $\sum_{n=1}^{L}w_{i_n}=1$ for $i=1,\ldots,M$. Typically, these weights are specified by two dimensional finite-extent Gaussian or rectangular shaped windows [10]. The SSIM between images $\boldsymbol{s}$ and $\boldsymbol{\gamma}$ is then given as

$$\mathrm{SSIM}(\boldsymbol{s},\boldsymbol{\gamma})=\frac{1}{M}\sum_{i=1}^{M}l_i(\boldsymbol{s},\boldsymbol{\gamma})\,g_i(\boldsymbol{s},\boldsymbol{\gamma}), \tag{11}$$

$$l_i(\boldsymbol{s},\boldsymbol{\gamma})=\frac{2\left(\sum_{n=1}^{L}w_{n_i}\gamma_n\right)\left(\sum_{n=1}^{L}w_{n_i}s_n\right)+c_1}{\left(\sum_{n=1}^{L}w_{n_i}\gamma_n\right)^2+\left(\sum_{n=1}^{L}w_{n_i}s_n\right)^2+c_1}, \tag{12}$$

$$g_i(\boldsymbol{s},\boldsymbol{\gamma})=\frac{2\sum_{n=1}^{L}w_{n_i}\gamma_n s_n-\left(\sum_{n=1}^{L}w_{i_n}\gamma_n\right)\left(\sum_{n=1}^{L}w_{i_n}s_n\right)+c_2}{\sum_{n=1}^{L}w_{i_n}\left(\gamma_n^2+s_n^2\right)-\left(\sum_{n=1}^{L}w_{i_n}\gamma_n\right)^2-\left(\sum_{n=1}^{L}w_{i_n}s_n\right)^2+c_2}, \tag{13}$$

with constants $c_1,c_2>0$. We have that $\mathrm{SSIM}(\boldsymbol{s},\boldsymbol{\gamma})=1$ if and only if the images are identical [10]. Furthermore, $\mathrm{SSIM}(\boldsymbol{s},\boldsymbol{\gamma})$ of 0 indicates no structural similarity, and of -1 indicates negative correlation between the images [10].

We define $q''_{\boldsymbol{\phi}}(\boldsymbol{s}|\boldsymbol{y})$, parameterized by $\boldsymbol{\gamma}(\boldsymbol{y})$, such that we get the SSIM loss to minimize the cross-entropy. We model $q''_{\boldsymbol{\phi}}(\boldsymbol{s}|\boldsymbol{y})$ with a NN with NN parameters $\boldsymbol{\phi}$. The input to the NN is $\boldsymbol{y}$ and the output is $\boldsymbol{\gamma}(\boldsymbol{y})$, which are the parameters fully describing $q''_{\boldsymbol{\phi}}(\boldsymbol{s}|\boldsymbol{y})$. Again, in our notation, we omit the dependence of $\boldsymbol{\gamma}(\boldsymbol{y})$ from the NN parameters $\boldsymbol{\phi}$ for simplicity. We normalized all pixel values to be in the range of $[0,1]$, and we define

$$q''_{\boldsymbol{\phi}}(\boldsymbol{s}|\boldsymbol{y})=\begin{cases}\exp\left(\mathrm{SSIM}\left(\boldsymbol{s},\boldsymbol{\gamma}(\boldsymbol{y})\right)-1\right), & \text{if } \boldsymbol{s}\in[0,d]^L,\\ 0, & \text{otherwise,}\end{cases} \tag{14}$$

and d such that $\int_0^d\ldots\int_0^d\exp\left(\mathrm{SSIM}\left(\boldsymbol{s},\boldsymbol{\gamma}(\boldsymbol{y})\right)-1\right)ds_1\ldots ds_L=1$.

As we have $|\mathrm{SSIM}\left(\boldsymbol{s},\boldsymbol{\gamma}(\boldsymbol{y})\right)|\leq 1$ [10], we have $e^{-2}\leq\exp\left(\mathrm{SSIM}\left(\boldsymbol{s},\boldsymbol{\gamma}(\boldsymbol{y})\right)-1\right)\leq 1$. Therefore, we have $d\in[1,e^{\frac{2}{L}}]$, such that $q''_{\boldsymbol{\phi}}(\boldsymbol{s}|\boldsymbol{y})$ is not zero for $\boldsymbol{s}\in[0,1]$, which is the rage the image data is normalized to. We have shown now that $q''_{\boldsymbol{\phi}}(\boldsymbol{s}|\boldsymbol{y})$ is a PDF, since is non-negative and integrates to one over its domain.

Now, for $\boldsymbol{s}\in[0,d]^L$, the cross-entropy between $p(\boldsymbol{s}|\boldsymbol{y})$ and $q''_{\boldsymbol{\phi}}(\boldsymbol{s}|\boldsymbol{y})$ becomes

$$-\mathrm{E}_{p(\boldsymbol{y})}\left[\mathrm{E}_{p(\boldsymbol{s}|\boldsymbol{y})}\left[\log\left(q''_{\boldsymbol{\phi}}(\boldsymbol{s}|\boldsymbol{y})\right)\right]\right]=\mathrm{E}_{p(\boldsymbol{y})}\left[\mathrm{E}_{p(\boldsymbol{s}|\boldsymbol{y})}\left[1-\mathrm{SSIM}(\boldsymbol{s},\boldsymbol{\gamma}(\boldsymbol{y}))\right]\right], \tag{15}$$

which we can minimize over $\boldsymbol{\phi}$, and thus over $\boldsymbol{\gamma}(\boldsymbol{y})$, by approximating the expected value over $\boldsymbol{y}$ and $\boldsymbol{s}$ using train-

ing data samples. This way we obtain the loss function of $1 - \mathrm{SSIM}(\boldsymbol{s}|\boldsymbol{y})$ to train the encoders and the decoders.

As before, using the maximum likelihood criterion to get the estimate $\hat{\boldsymbol{s}}$, we need to find the maximizer of $q''_{\boldsymbol{\phi}}(\boldsymbol{s}|\boldsymbol{y})$ with respect to $\boldsymbol{s}$, where its clear from (14) that is has to be in $[0, d]^L$. Furthermore, the unique maximizer of $\mathrm{SSIM}(\boldsymbol{s}, \boldsymbol{\gamma}(\boldsymbol{y}))$ is $\boldsymbol{s} = \boldsymbol{\gamma}(\boldsymbol{y})$ [10]. As we have for $q''_{\boldsymbol{\phi}}(\boldsymbol{s}|\boldsymbol{y})$ a strictly monotonically increasing function of $\mathrm{SSIM}(\boldsymbol{s}, \boldsymbol{\gamma}(\boldsymbol{y}))$, its maximizer does not change. Therefore, we have $\boldsymbol{s} = \boldsymbol{\gamma}(\boldsymbol{y})$ as the unique maximizer of $q''_{\boldsymbol{\phi}}(\boldsymbol{s}|\boldsymbol{y})$, which means our reconstruction estimate is $\hat{\boldsymbol{s}} = \boldsymbol{\gamma}(\boldsymbol{y})$.

C. Combining MSE and SSIM Loss

To solve (8), we now assume the reconstruction decoder $q_{\boldsymbol{\phi}}(\boldsymbol{s}|\boldsymbol{y})$ to be a convex combination of the parameterized PDFs $q'_{\boldsymbol{\phi}}(\boldsymbol{s}|\boldsymbol{y})$and $q''_{\boldsymbol{\phi}}(\boldsymbol{s}|\boldsymbol{y})$ with parameters $\boldsymbol{\mu}(\boldsymbol{y})$ and $\boldsymbol{\gamma}(\boldsymbol{y})$, respectively. Since we want a convex combination of MSE and SSIM as loss function, we do parameter sharing with $\boldsymbol{v}_{\boldsymbol{\phi}}(\boldsymbol{y}) := \boldsymbol{\mu}(\boldsymbol{y}) = \boldsymbol{\gamma}(\boldsymbol{y}) \in \mathbb{R}^L$, which depends on the NN parameters $\boldsymbol{\phi}$. The combined PDF can then be written as

$$q_{\boldsymbol{\phi}}(\boldsymbol{s}|\boldsymbol{y}) = (1-\beta)\, q'_{\boldsymbol{\phi}}(\boldsymbol{s}|\boldsymbol{y}) + \beta\, q''_{\boldsymbol{\phi}}(\boldsymbol{s}|\boldsymbol{y}), \tag{16}$$

with $\beta \in [0, 1]$. The maximum likelihood estimate of the reconstruction is then given by $\hat{\boldsymbol{s}} = \boldsymbol{v}_{\boldsymbol{\phi}}(\boldsymbol{y})$, as this is the maximizer of both $q'_{\boldsymbol{\phi}}(\boldsymbol{s}|\boldsymbol{y})$ and $q''_{\boldsymbol{\phi}}(\boldsymbol{s}|\boldsymbol{y})$, as discussed above.

As we assumed the variance σ^2 of the Gaussian $q'_{\boldsymbol{\phi}}(\boldsymbol{s}|\boldsymbol{y})$ to be fixed, we choose $\sigma^2 = L/2$ here without restricting generality, as all cases of weighting MSE and SSIM are covered with $\beta \in [0, 1]$. Using (10) and (15), we can calculate the cross entropy between (16) and $p(\boldsymbol{s}|\boldsymbol{y})$, and by ignoring the constant terms, we get the solvable minimization problem

$$\begin{aligned} \underset{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_N,\,\boldsymbol{\phi},\,\boldsymbol{\psi}}{\operatorname{argmin}} \quad & \alpha\Big(\mathrm{E}_{p(\boldsymbol{y})}\left[\mathrm{E}_{p(\boldsymbol{s}|\boldsymbol{y})}\left[(1-\beta)\,\mathrm{MSE}\left(\boldsymbol{s}, \boldsymbol{v}_{\boldsymbol{\phi}}(\boldsymbol{y})\right)\right.\right. \\ & \left.\left. +\beta\left(1 - \mathrm{SSIM}\left(\boldsymbol{s}, \boldsymbol{v}_{\boldsymbol{\phi}}(\boldsymbol{y})\right)\right)\right]\right) \\ & +(1-\alpha)\,\mathrm{E}_{p(\boldsymbol{y})}\left[\mathrm{H}\left(p(\boldsymbol{z}|\boldsymbol{y}), q_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{y})\right)\right] \\ \text{s.t.} \quad & \boldsymbol{c}_i \in \mathbb{R}^{N_{\mathrm{Tx}}/N}, \quad P_i \leq 1, \quad \text{for } i = 1, \ldots, N. \end{aligned} \tag{17}$$

IV. Simulation Results

We evaluate the proposed semantic communication system for data reconstruction and task execution for an example setting of image reconstruction and image classification using the CIFAR-10 dataset consisting of color images of size $32 \times 32 \times 3$ [13]. We consider a multi-view scenario with $N = 4$ sensing nodes, where each node has access to a non-overlapping square quarter of the image, corresponding to an image of size $16 \times 16 \times 3$. Each sensing node independently encodes the data and transmits over independent channels to a central receiver for classification and image reconstruction.

We use the ResNet14 architecture [14] for the encoders and reconstruction decoder. Each encoder consists of one convolution layer, six residual blocks, and a fully connected layer with power normalization. The reconstruction decoder includes two fully connected layers followed by residual blocks (one standard, two transposed for upsampling, and one standard) and a final transposed convolution layer. The classification decoder consists of three fully connected layers[1].

[1]The code is available at https://github.com/ant-uni-bremen/Semantic-Communication-for-Task-Execution-and-Data-Reconstruction

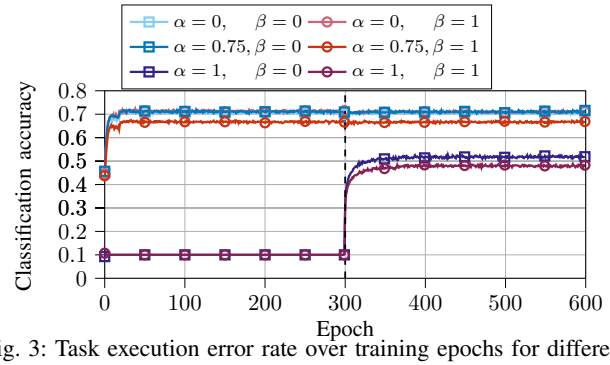

Fig. 3: Task execution error rate over training epochs for different α and β values.

We train the NN parameters of the proposed semantic communication system in an end-to-end manner according to (17) for $\alpha \in [0, 1]$ and $\beta \in [0, 1]$. We use a batch size of 32 and a learning rate of 10^{-4}. For the cases of $\alpha = 0$ and $\alpha = 1$, the respective objective of data reconstruction and task execution in (17) vanishes, meaning that the respective decoder is not trained at all. To mitigate this problem, we first train the whole system for 300 epochs. Then the encoder weights are frozen and the two decoders are trained for another 300 epochs, which are independent of α, as discussed before for (7). This training procedure is shown in Fig. 3 for the image classification accuracy over the training epochs for some combinations of α and β values. For $\alpha = 1$, the classification accuracy is 0.1 for the first 300 epochs, as the classification decoder is only trained in the second training phase. For $\alpha < 1$, the classification decoder is trained normally, and the training converges rather quickly after about 100 epochs.

For the following simulations, the number of channel uses per encoder is $N_{\mathrm{Tx}}/N = 50$, our channel model is an Additive White Gaussian Noise (AWGN) channel, and the training and evaluation Signal to Noise Ratio (SNR) per channel use is 3 dB. We evaluate the task execution accuracy by the classification accuracy, i.e., the ratio of correctly classified images. The data reconstruction performance is measured by Peak Signal to Noise Ratio (PSNR) and SSIM. The PSNR between reconstructed image $\hat{\boldsymbol{s}}$ and true image $\boldsymbol{s}$ with our normalization is given as $\mathrm{PSNR} = 10\log_{10}\left(\frac{1}{\frac{1}{L}\sum_{n=1}^{L}(s_n - \hat{s}_n)^2}\right)$. For $\mathrm{SSIM}(\boldsymbol{s}, \hat{\boldsymbol{s}})$ (11), we use the standard TensorFlow implementation using Gaussian weighted windows [10]. For both PSNR and SSIM, larger values correspond to higher similarity between true and estimated image.

For comparison, two baseline methods are considered. As a digital communication baseline, the images are subsampled, quantized and Huffman coded, followed by an LDPC channel code of rate $\frac{1}{2}$. The average number of channel uses per image is set to $N_{\mathrm{Tx}}/N = 50$. The subsampling factor, the number of quantization bits, and the modulation order are optimized via grid search for image reconstruction at $\mathrm{SNR} = 3\,\mathrm{dB}$. When evaluated with a classifier trained independently on the CIFAR-10 dataset, classification of the reconstructed images effectively fails, achieving an accuracy only slightly above the random baseline of 0.1. As an upper bound on the accuracy of the multi-view classification problem, a perfect channel baseline is included with uncompressed and noiseless transmission.

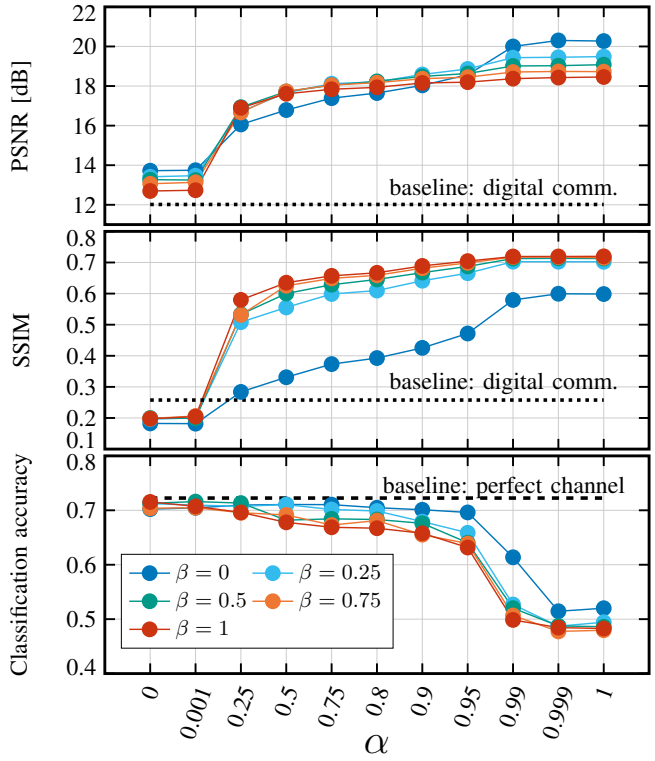

Fig. 4: Trade-off between PSNR, SSIM, and classification accuracy across α values for different β weighting MSE and SSIM loss, with $N = 4$ sensing nodes, $N_{\text{Tx}}/N = 50$ channel uses per sensing node, and SNR $= 3$ dB.

We investigate the trade-off between data reconstruction and task execution over α for different weights β between MSE and SSIM, with the results shown in Fig. 4.

First, we analyze the case where only the MSE loss is used for image reconstruction ($\beta = 0$), over different α values. For small α values, the model favors classification accuracy and neglects transmitting features required for image transmission, achieving only a PSNR of about 14 dB and an SSIM of about 0.2. When α is increased up to about 0.9, the PSNR is increased to about 18 dB and SSIM to about 0.42, while the task accuracy remains basically stable. This indicates that for larger α up to $\alpha \approx 0.9$, features are selected for transmission that contain more information required for image reconstruction without having to sacrifice transmitting discriminative features required for classification. For larger α beyond this point, the classification accuracy declines while the reconstruction performance continues to improve. For $\alpha = 1$, the classification accuracy drops to about 0.5, as mostly features are transmitted for fine visual details, which do not help much for classification.

Furthermore, we investigate the trade-off between MSE and SSIM, where a larger β means a larger weight on SSIM. This is reflected in the results, where larger β increases SSIM, but decreases MSE for large values of α. It can be seen that $\beta > 0$ leads to larger PSNR in the range of $0.25 \leq \alpha \leq 0.9$, which can be explained by the fact that $1 - \text{SSIM}(\boldsymbol{s}, \hat{\boldsymbol{s}})$ is larger than $\text{MSE}(\boldsymbol{s}, \hat{\boldsymbol{s}})$, which leads to a larger weight for image reconstruction. Furthermore, MSE prioritizes pixel accuracy, preserving texture, but often causing blur, whereas SSIM prioritizes structure, enhancing shape preservation. A balanced trade-off between PSNR and SSIM and classification accuracy can be achieved for $\beta = 0.25$ and $\alpha = 0.75$, where the classification accuracy is 0.7, PSNR $= 18$ dB, and SSIM $= 0.6$. In comparison, for $\beta = 0$ and $\alpha = 0.9$, the classification accuracy remains 0.7 and PSNR $= 18$ dB, while the SSIM decreases significantly to SSIM $= 0.42$.

V. Conclusion

We investigated a semantic communication system designed for joint task execution and data reconstruction. As both objectives cannot be optimally fulfilled at the same time, a trade-off has to be made. We modeled this trade-off by formulating a joint objective as a convex combination of both objectives of task execution and data reconstruction.

Our findings show that data reconstruction quality can be significantly improved without limiting task execution significantly for moderate weighting of data reconstruction with $\alpha \leq 0.9$. This can be achieved by learning to transmit features that preserve fine visual detail for data reconstruction while still being as discriminative as possible for task execution. Finally, it was shown under which assumptions we get the SSIM loss from the mutual information maximization problem.

References

[1] D. Gündüz, Z. Qin, I. E. Aguerri, H. S. Dhillon, Z. Yang, A. Yener, K. K. Wong, and C.-B. Chae, "Beyond transmitting bits: Context, semantics, and task-oriented communications," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 1, pp. 5–41, 2022.

[2] T. M. Getu, G. Kaddoum, and M. Bennis, "Semantic communication: A survey on research landscape, challenges, and future directions," *Proceedings of the IEEE*, 2025.

[3] W. Tong and G. Y. Li, "Nine challenges in artificial intelligence and wireless communications for 6g," *IEEE Wireless Communications*, vol. 29, no. 4, pp. 140–145, 2022.

[4] M. Sana and E. C. Strinati, "Learning semantics: An opportunity for effective 6g communications," in *2022 IEEE 19th Annual Consumer Communications & Networking Conference (CCNC)*. IEEE, 2022.

[5] J. Shao, Y. Mao, and J. Zhang, "Learning task-oriented communication for edge inference: An information bottleneck approach," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 1, pp. 197–211, 2022.

[6] E. Beck, C. Bockelmann, and A. Dekorsy, "Semantic information recovery in wireless networks," *Sensors*, vol. 23, p. 6347, 7 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/14/6347

[7] 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.

[8] Z. Lyu, G. Zhu, J. Xu, B. Ai, and S. Cui, "Semantic communications for image recovery and classification via deep joint source and channel coding," *IEEE Transactions on Wireless Communications*, vol. 23, no. 8, pp. 8388–8404, 2024.

[9] W. Yuan, J. Ren, C. Wang, R. Zhang, J. Wei, D. I. Kim, and S. Cui, "Generative semantic communication for joint image transmission and segmentation," in *2025 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, 2025, pp. 1110–1115.

[10] A. K. Venkataramanan, C. Wu, A. C. Bovik, I. Katsavounidis, and Z. Shahid, "A hitchhiker's guide to structural similarity," *IEEE Access*, vol. 9, pp. 28 872–28 896, 2021.

[11] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, Ltd, 2005.

[12] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[13] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," 2009.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.