Semantic Joint Source Channel Coding for Distributed Subsurface Imaging in Multi-Agent Systems

Maximilian H. V. Tillmann, *Graduate Student Member, IEEE,* Ban-Sok Shin, *Member, IEEE,* Dmitriy Shutin, *Senior Member, IEEE,* Armin Dekorsy, *Senior Member, IEEE*

Abstract*—Multi-agent systems (MASs) are a promising solution for autonomous exploration tasks in hazardous or remote environments. In such settings, communication among agents is essential to ensure collaborative task execution, yet conventional approaches treat exploration and communication as decoupled subsystems. This work presents an approach that tightly integrates semantic communication into the MAS exploration process, adapting the communication system to the exploration methodology to improve overall task performance. Specifically, we investigate the application of semantic joint source-channel coding (JSCC) with over-the-air computation (AirComp) for distributed function computation for the application of cooperative subsurface imaging using the adapt-then-combine full waveform inversion (ATC-FWI) algorithm. Our results demonstrate that semantic JSCC significantly outperforms classical digital communication and conventional JSCC approaches, especially in high-connectivity networks. Furthermore, incorporating side information at the receiving agent enhances communication efficiency and imaging accuracy, a feature previously unexplored in MAS-based exploration. We validate our approach through a use case inspired by subsurface anomaly detection, showing measurable improvements in imaging performance per agent. This work underscores the potential of semantic communication in distributed multi-agent exploration to improve overall exploration accuracy and efficiency.

***Index Terms*—Semantic communication, multi-agent systems, autonomous exploration, subsurface imaging**

I. Introduction

MULTI-AGENT systems have emerged as a crucial tool to address various tasks that cannot be handled reliably by humans such as inspection, surveying, monitoring, and exploration of hazardous or remote environments [1], [2]. For instance, multi-agent systems (MASs) have been proposed and investigated for the use in planetary exploration, gas and subsurface exploration [3]–[5]. A MAS inherently requires communication among its agents to achieve the respective exploration goal such as reconstruction of a physical process or estimation of certain parameters. Usually, the information exchange among agents is realized by a wireless communication system and therefore, most works use standard communication schemes to achieve this, cf. [6], [7]. In this respect, exploration and communication are viewed as two systems that are designed separately although exploration and communication are often strongly entangled in a MAS. To achieve enhanced exploration performance both communication and exploration need to be considered jointly. Therefore, in this work, we propose to integrate semantic communication into a distributed exploration algorithm such that the overall system's exploration performance is enhanced. The specific exploration task considered here is subsurface exploration by means of seismic imaging which is illustrated in Fig. 1.

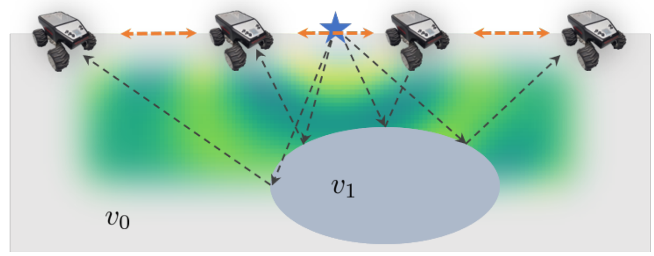


Fig. 1: Illustration of the exploration scenario: A MAS consisting of ground rovers reconstructs the velocity structure (v_0, v_1) of a subsurface using a seismic source (blue) and seismic waves reflecting from an anomaly. The rovers are connected over wireless links (orange) that enable data exchange within the MAS. The communication among rovers is realized via semantic communication to adapt the communication system to the exploration algorithm.

Semantic communication is a promising technology that enables task-specific communication beyond the Shannon limit of classical communication systems [8]–[11]. Compared to classical communication systems semantic communication aims to transmit only the required "meaning" of the message for accurate task execution to save bandwidth [8]. Recently, modern machine learning techniques have been used to develop such semantic communication systems. Such a system becomes especially advantageous in distributed exploration scenarios, where a task needs to be solved by multiple agents. In this case, semantic communication makes it possible to only exchange information relevant for the exploration task between the agents instead of transmitting the complete data.

This work was supported in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under grant 500260669 (SCIL), by the DFG under Germany's Excellence Strategy – EXC-3036 The Martian Mindset, project number: 533607631, and by the German Ministry of Research, Technology and Space (BMFTR) under project number 16KISK016 (Open6GGub).

Maximilian H. V. Tillmann and Armin Dekorsy are with the Department of Communications Engineering, University of Bremen, 28359 Bremen, Germany (e-mail: tillmann@ant.uni-bremen.de; dekorsy@ant.uni-bremen.de).

Ban-Sok Shin and Dmitriy Shutin are with the Institute of Communications and Navigation, German Aerospace Center, 82234 Wessling, Germany (e-mail: ban-sok.shin@dlr.de; dmitriy.shutin@dlr.de).

DOI: 10.1109/TSIPN.2026.3719310 https://ieeexplore.ieee.org/document/11636215

One envisioned application scenario of MASs is the autonomous quantification of lunar caves in future exploration missions using seismic techniques [5], [12], [13]. Seismic methods typically require spatially distributed, simultaneous measurements. A MAS with agents carrying seismic sensors thus perfectly exemplifies the use case of interest. In this regard, the Adapt-Then-Combine Full Waveform Inversion (ATC-FWI) has been proposed, which enables subsurface imaging in a distributed fashion within a MAS [5]. In essence, ATC-FWI obtains subsurface images at each agent via data exchange among neighboring agents using distributed data processing. One fundamental operation in such distributed processing is averaging of exchanged data, which can be found in e.g., consensus averaging and diffusion-based schemes [14], [15]. From a semantic communication perspective the distributed averaging process of several input data at the receiver belongs to the problem of *distributed function computation* [16], [17]. It is at this point where we apply semantic communication techniques to appropriately encode the data of multiple transmitters (transmitting agents) to perform Distributed Function Computation (DFC) for the averaging step at the decoder (receiving agent). We employ a probabilistic model for the encoders and the decoder of the proposed semantic communication system, where we assume that the decoder is modeled with independent Gaussian distributions for every vector element.

In this work, we propose the use of semantic communication in a distributed subsurface imaging method for a MAS to obtain reliable imaging results for noisy inter-agent communication links. To the best of our knowledge, this is the first work to employ semantic communication for distributed exploration in a MAS. We summarize the main contributions as follows:

- **Semantic DFC:** Since the agents reconstruct the same subsurface from different positions, a receiving agent has correlated side information available. We exploit this fact by formulating the semantic DFC problem as a maximization of the mutual information given the side information, i.e., the received symbols should be maximally informative about the semantic variable given the side information. Since maximizing the mutual information directly is infeasible (cf. [18], [19]), we derive a lower bound of the conditional mutual information, which can be maximized using end-to-end training. In previous works on semantic communication, either no side information was considered [19], or an evidence lower bound was used when side information was present [20].
- **Integration to MAS:** We apply our proposed semantic DFC to the data exchange procedure within the ATC-FWI. We employ Over-the-Air Computation (AirComp) for semantic DFC to use channel resources between agents more efficiently. To this end, we train the encoders and the decoder for AirComp in an end-to-end manner using Deep Neural Networks (DNNs). This allows us to express the decoder as a nomographic function, which is necessary to enable addition of the encoded symbols over the communication channel.
- **Results:** In simulations, we show that the proposed method can improve bandwidth efficiency for inter-agent connections corrupted by white Gaussian noise in ATC-FWI compared to state-of-the-art Joint Source and Channel Coding (JSCC) methods. Our proposed approach of semantic DFC outperforms state-of-the-art JSCC methods, especially in scenarios, where many neighboring agents exchange data with each other. In particular, compared to a classical digital transmission our semantic DFC shows significantly improved communication performance and thus, better imaging performance for ATC-FWI.

II. Distributed Subsurface Imaging for Multi-Agent Systems

In the following section, we briefly describe the Full Waveform Inversion (FWI) as the classical, centralized imaging scheme from geophysics. Afterwards, we explain how FWI can be adapted to a MAS for distributed imaging, when each agent obtains a global subsurface image locally. This leads to the ATC-FWI algorithm which requires explicit data exchange among neighboring agents and thus, suitable communication.

A. Background

To address the exploration problem in terms of subsurface imaging we consider a technique called FWI. In particular, FWI is a geophysical imaging technique that enables reconstruction of subsurface structures by exploiting wave physics [21], [22]. Despite being rooted in seismic subsurface reconstruction, FWI is applied to various other modalities such as ultrasound, ground penetrating radar and electrical resistivity tomography. In its traditional form, FWI is a centralized scheme that requires all measurement data to be allocated at a central entity that performs imaging. Hence, for application in a MAS it needs to be adapted to enable imaging at each agent via exchange of data within the network. In our previous work, we proposed the ATC-FWI which enables distributed imaging within a MAS [5]. However, for simplicity ATC-FWI has been considered with error-free communication among the agents in the previous work, which cannot be achieved with noisy communication channels. Other distributed or decentralized schemes for subsurface imaging have been proposed in the literature, cf. [23], [24]. However, in these works communication aspects are considered separately from the imaging problem. That is, the communication is realized by using standard wireless communication systems and an adapted design of the communication system to the subsurface imaging is not considered.

B. Centralized Full Waveform Inversion

In its ordinary formulation, FWI aims at minimizing a least-squares cost functional of a data residual with respect to a subsurface parameter $m(\boldsymbol{x})$ that varies over space coordinate $\boldsymbol{x}$. The residual is evaluated between observed measurements $d^{\mathrm{obs}}(t)$ and predicted or synthesized measurements $d^{\mathrm{syn}}(t, m)$ over time t. The synthesized measurements $d^{\mathrm{syn}}(t, m)$ are generated by solving the acoustic wave equation and sampling the obtained wavefield $u_s(\boldsymbol{x},t)$ at the receiver positions via a sampling operator $S_{s,r}$ for a receiver r and the respective source s. The fact that the minimization is coupled to the acoustic wave equation can be formulated as a constraint in the resulting optimization problem:

$$\underset{m}{\operatorname{argmin}}\, \mathcal{L}(m) = \frac{1}{2}\sum_{s,r}\int_0^{\tau} [\underbrace{S_{s,r}u_s(\boldsymbol{x},t)}_{d_{s,r}^{\mathrm{syn}}(t,m)} - d_{s,r}^{\mathrm{obs}}(t)]^2\,\mathrm{d}t \tag{1a}$$

$$\text{s.t.}\quad \frac{1}{m^2(\boldsymbol{x})}\frac{\partial^2 u_s(\boldsymbol{x},t)}{\partial t^2} - \frac{\partial^2 u_s(\boldsymbol{x},t)}{\partial \boldsymbol{x}^2} = f_s(\boldsymbol{x},t). \tag{1b}$$

where $\sum_{s,r} := \sum_{s=1}^{N_{\mathrm{S}}}\sum_{r=1}^{N_{\mathrm{R}}}$ and $N_{\mathrm{S}}, N_{\mathrm{R}}$ is the number of seismic sources and receivers, respectively. In our case, the subsurface parameter $m(\boldsymbol{x})$ is a function over space and represents the spatial distribution of the P-wave velocity. As can be seen, FWI essentially aims at solving an optimization problem that is constrained by a Partial Differential Equation (PDE). In the wave equation (1b), $f_s(\boldsymbol{x},t)$ is the seismic source specific to source s. Solving (1b) for a subsurface model $m(\boldsymbol{x})$ results in the wavefield $u_s(\boldsymbol{x},t)$ which is then sampled at the receiver positions $\{\boldsymbol{x}_r\}_{r=1}^{N_{\mathrm{R}}}$ to give $d_{s,r}^{\mathrm{syn}}(t,m)$.

Problem (1) is nonlinear in $m(\boldsymbol{x})$ and therefore requires gradient-based optimization to converge to a suitable subsurface model. Commonly used methods include gradient-descent, nonlinear conjugate gradient or L-BFGS. To this end, the gradient of $\mathcal{L}(m)$ needs to be calculated. Note that $\mathcal{L}(m)$ is constrained by a PDE which complicates direct calculation of the gradient. Typically, the so-called *adjoint state method* is used to compute the gradient of cost functionals that are constrained by a PDE, cf. [25]. Applying the adjoint state method to (1a) together with (1b) results in the following gradient computation [5]:

$$\frac{\mathrm{d}\mathcal{L}(m)}{\mathrm{d}m} = -\frac{2}{m(\boldsymbol{x})^3}\sum_{s=1}^{N_{\mathrm{S}}}\int_0^{\tau} q_s(\boldsymbol{x},\tau-t)\frac{\partial^2 u_s(\boldsymbol{x},t)}{\partial t^2}\mathrm{d}t. \tag{2}$$

Here, the function $q_s(\cdot,\cdot)$ is a so-called adjoint wavefield, which is obtained by solving the following PDE which is an adjoint wave equation:

$$\left(\frac{1}{m(\boldsymbol{x})^2}\frac{\partial^2}{\partial t^2} - \frac{\partial^2}{\partial \boldsymbol{x}^2}\right) q_s(\boldsymbol{x},t) = \sum_{r=1}^{N_{\mathrm{R}}} S_{s,r}(d_{s,r}^{\mathrm{syn}}(\tau-t) - d_{s,r}^{\mathrm{obs}}(\tau-t)). \tag{3}$$

Note that to solve (3), the data residuals $(d_{s,r}^{\mathrm{syn}}(\tau-t) - d_{s,r}^{\mathrm{obs}}(\tau-t))$ are now used as a source signal that is reversed in time and injected at the receiver positions by a sampling operator $S_{s,r}$. After obtaining the adjoint wavefield, $q_s(\boldsymbol{x},t)$ is again reversed in time and correlated with the second time derivative of the wavefield $u_s(\boldsymbol{x},t)$ following (2).

To solve PDEs in (1b) and (3) we require a numerical solver on a discrete space and time grid. In our work, we use the Devito package that provides highly efficient finite difference solvers for PDEs [26]. By discretizing the considered 2D computational domain into a $N_{\mathrm{x}} \times N_{\mathrm{z}}$ grid we obtain vector quantities for the subsurface model $m(\boldsymbol{x})$ and the gradient function $\mathrm{d}\mathcal{L}(m)/\mathrm{d}m$. Therefore, we replace these functions by their discrete vector description, i.e.,

$$m(\boldsymbol{x}) \to \boldsymbol{m}(\boldsymbol{x}) \in \mathbb{R}^{N_{\mathrm{x}}N_{\mathrm{z}}}, \quad \mathrm{d}\mathcal{L}(m)/\mathrm{d}m \to \delta\boldsymbol{m} \in \mathbb{R}^{N_{\mathrm{x}}N_{\mathrm{z}}}. \tag{4}$$

Given a gradient by (2), we can now minimize $\mathcal{L}(m)$ iteratively using gradient descent. This results in the following update procedure for the estimated subsurface model:

$$\boldsymbol{m}^{[k+1]} = \boldsymbol{m}^{[k]} - \alpha^{[k]}\delta\boldsymbol{m}^{[k]} \tag{5}$$

The parameter $\alpha^{[k]} > 0$ is a step size that can be adapted over the iterations e.g. by a simple exponential decay or a line search method.

C. Adapt-Then-Combine Full Waveform Inversion (ATC-FWI)

FWI in its classical setting is centralized algorithm, i.e., it requires all measurement data $d_{s,r}^{\mathrm{obs}}$ to be available at a central entity to perform the inversion procedure. However, in the MAS no central entity is available that collects all measurement data from the agents. Instead, we demand a computation within the MAS by exploiting agent-to-agent communication.

The objective of a distributed imaging scheme is to obtain an image locally at each agent/receiver via cooperation with neighboring agents, i.e., those that are in the communication range of the respective agent. To this end, the ATC-FWI has been proposed in [5]. In the following, we give a short summary of the algorithm.

Let us first introduce assumptions about the network topology of the MAS. We assume a network of N_{R} agents, where each agent is equipped with a geophone or a seismic sensor. Each agent r can communicate with neighboring agents ℓ that are contained in a neighborhood set $\mathcal{N}_r \subseteq \{1,\ldots,N_{\mathrm{R}}\}$ and vice versa. We assume that the neighborhood set $\mathcal{N}_r$ contains the agent r itself. Furthermore, the network graph is undirected and connected, i.e., each agent can be reached by any other agent in the network over multiple hops.

The first step in deriving ATC-FWI is to represent the global cost in (1a) over all N_{R} agents in the network such that

$$\mathcal{L}(m) = \sum_{r=1}^{N_{\mathrm{R}}} \mathcal{L}_r(m), \tag{6}$$

with each agent's local cost given now as

$$\mathcal{L}_r(m) = \frac{1}{2}\sum_{s=1}^{N_{\mathrm{S}}}\int_0^{\tau}\left(d_{s,r}^{\mathrm{syn}}(t,m) - d_{s,r}^{\mathrm{obs}}(t)\right)^2 dt. \tag{7}$$

Note that each local cost $\mathcal{L}_r(m)$ depends only on locally observed measurement data $d_{s,r}^{\mathrm{obs}}(t)$. Based on its local cost each agent r is able to compute a gradient w.r.t. the subsurface model m in the same manner as in centralized FWI, however, with local measurement data $d_{s,r}^{\mathrm{obs}}(t)$. To ensure that global subsurface information is available to all agents in the network, we introduce N_{R} different (local) subsurface models which are then combined using diffusion strategies [14]. Specifically, we assign individual subsurface models $m_r(\boldsymbol{x})$ to each agent. Then each agent solves the wave equation (1b) and the adjoint equation (3) using its own subsurface model $m_r(\boldsymbol{x})$ to obtain the wavefield $u_{s,r}(\boldsymbol{x},t)$ and the adjoint wavefield $q_{s,r}(\boldsymbol{x},t)$ according to

$$\frac{1}{m_r(\boldsymbol{x})^2}\frac{\partial^2 u_{s,r}(\boldsymbol{x},t)}{\partial t^2} - \frac{\partial^2 u_{s,r}(\boldsymbol{x},t)}{\partial \boldsymbol{x}^2} = f_s(\boldsymbol{x},t) \tag{8}$$

$$\left(\frac{1}{m_r(\boldsymbol{x})^2}\frac{\partial^2}{\partial t^2} - \frac{\partial^2}{\partial \boldsymbol{x}^2}\right) q_s(\boldsymbol{x},t) = S_{s,r}^{\mathsf{T}}(d_{s,r}^{\mathrm{syn}}(\tau - t) - d_{s,r}^{\mathrm{obs}}(\tau - t)), \tag{9}$$

followed by a local gradient computation $\delta m_r(\boldsymbol{x})$ via correlation:

$$\delta m_r(\boldsymbol{x}) = -\frac{2}{m(\boldsymbol{x})^3}\sum_{s=1}^{N_{\mathrm{S}}}\int_0^\tau q_{s,r}(\boldsymbol{x},\tau - t)\frac{\partial^2 u_{s,r}(\boldsymbol{x},t)}{\partial t^2}\mathrm{d}t \tag{10}$$

Now, each agent has an individual subsurface model $m_r(\boldsymbol{x})$ and a local gradient $\delta m_r(\boldsymbol{x})$ available. To enable distributed imaging such that each agent's model approaches the centralized FWI result, we apply the adapt-then-combine (ATC) technique from [14]. Assuming again discretization of the spatial domain using finite differences we represent gradients and subsurface models via vectors. Applying ATC results in the following computations per agent r:

$$\widetilde{\boldsymbol{m}}_r^{[k+1]} = \boldsymbol{m}_r^{[k]} + \alpha^{[k]}\sum_{\ell\in\mathcal{N}_r} b_{\ell r}\delta\boldsymbol{m}_\ell^{[k]} \tag{11a}$$

$$\boldsymbol{m}_r^{[k+1]} = \sum_{\ell\in\mathcal{N}_r} a_{\ell r}\widetilde{\boldsymbol{m}}_\ell^{[k+1]} \tag{11b}$$

The coefficients $a_{\ell r}, b_{\ell r}$ need to be chosen appropriately such that the centralized FWI result can be approached by all agents, cf. [14]. In the simplest case, one can select these coefficients to be $1/|\mathcal{N}_r|$, where $|\cdot|$ denotes cardinality of the neighborhood set. This results in a uniform averaging of neighboring data.

From (11) we see that in each iteration k each agent r needs to exchange its gradient $\delta\boldsymbol{m}_r$ and its intermediate subsurface model $\widetilde{\boldsymbol{m}}_r$ with its neighbors $\ell \in \mathcal{N}_r$. In particular, the fusion process of neighboring data in ATC is given by an averaging process. It is here, where we will integrate semantic communications to exploit the averaging structure in the data exchange among the agents.

III. Semantic Distributed Function Computation for ATC-FWI

First, a background of semantic communication for MASs is provided. In what follows, we discuss the semantic communication approach towards computation of ATC over the network. In particular, we will outline the design of corresponding encoder-decoder structures for computing the averaging operations in (11a) and (11b) over a wireless communication channel. We already note here that, in order to design the proposed semantic communication system with probabilistic encoder and decoder models, we simplify the decoder distribution by modeling it with independent Gaussian distributions to efficiently solve the formulated optimization problem. The investigation of alternative decoder distribution types is left for future work.

A. Background

For semantic communication we want to compress and channel code the data at the transmitter such that we can reconstruct the semantic information at the receiver. To this end, deep learning techniques are used to learn an end-to-end encoder decoder system for JSCC [8], [27]. Semantic communication systems become increasingly important in distributed settings with multiple transmitters and one or more receivers [28], quite typical for MAS applications. For example, in [19] and [29] semantic communication was investigated for the case when multiple transmitters with different observations need to cooperatively solve a classification task. In [30], the authors investigate the case where one receiver that obtains two correlated images from two transmitters has to reconstruct both images.

The application of semantic communication to ATC-FWI requires consideration of two key aspects: side information at the decoder, i.e., the receiving agent and DFC. Side information should be incorporated into the semantic communication system for ATC-FWI, as the data available at each agent are correlated and can therefore assist the decoding process, and DFC should be considered since ATC-FWI requires the computation of a function of the exchanged data, which is here the weighted average. In this respect, the use of side information at the receiver for communication of images, was investigated in [20] and [31]. In [20], it was shown for image transmission that the compression level for the case when side information is only available at the decoder can nearly achieve the same compression level when side information is available at both the encoder and the decoder. On the other hand, DFC was studied in [32], with its variation with AirComp investigated in [17], [33], and deep learning based AirComp in [34]–[36]. Furthermore, AirComp was used in [37] for distributed convex optimization. AirComp aims at using channel resources more efficiently by letting users transmit at the same time and frequency. In this way, a function can be evaluated directly on the wireless channel as a sum of the transmitted symbols by exploiting the superposition property of electromagnetic waves [38], [39]. It was shown that despite advantages, AirComp mainly has two drawbacks: (i) the limitation that, with encoders $\boldsymbol{\varphi}_i$ and decoder $\boldsymbol{\phi}$, the transmitted symbols $\boldsymbol{\varphi}_i(\boldsymbol{s}_i)$ are required to be decodable after being added over the wireless communication channel, i.e., that they form a so-called nomographic function $\boldsymbol{f}(\boldsymbol{s}_1,\ldots,\boldsymbol{s}_N) = \boldsymbol{\phi}(\sum_{i=1}^N \boldsymbol{\varphi}_i(\boldsymbol{s}_i))$ and (ii) the requirement of accurate synchronization [38], [39]. A nomographic function is required since AirComp allows simultaneous multi-device transmission, resulting in a received signal that is the superposition of the individual carrier signal amplitudes [38].

B. Notation

In the following, we denote the receiving agent by the index r. We define $N_r \triangleq |\mathcal{N}_r| - 1$ to count the number of neighbors of the receiving agent r, excluding the agent itself, as the neighborhood $\mathcal{N}_r$ includes (by definition) the receiving agent r as well. In order to be able to denote the neighbors of any agent r generally, we will use the following indices to address

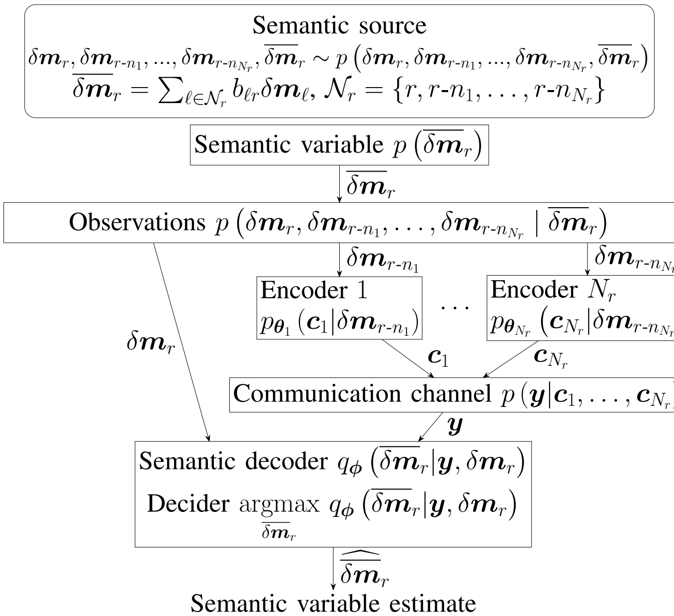


Fig. 2: Block diagram of the multi-agent semantic DFC system for agent r receiving the velocity gradients $\delta\boldsymbol{m}_\ell$ from its neighbors. The goal at the decoder is to directly reconstruct the semantic variable $\widehat{\overline{\delta\boldsymbol{m}}}_r$, which is here the weighted average of the velocity gradients from the neighboring agents.

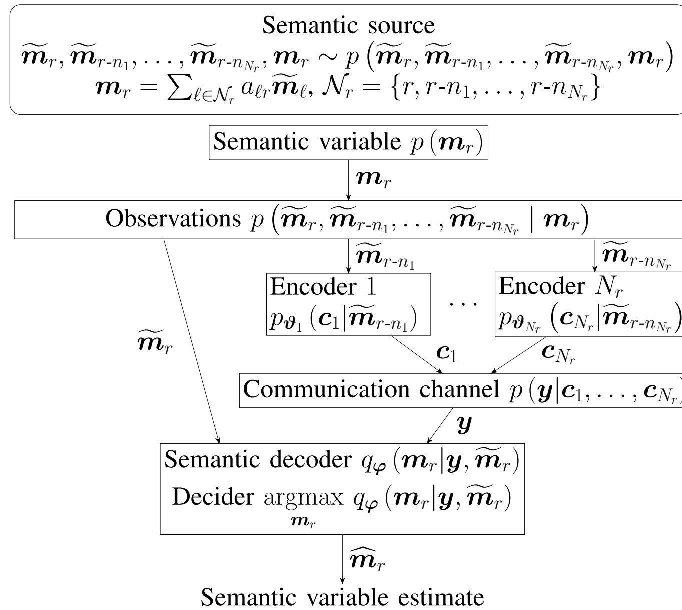


Fig. 3: Block diagram of the multi-agent semantic DFC system for agent r receiving the velocity models $\widetilde{\boldsymbol{m}}_\ell$ from its neighbors. The goal at the decoder is to directly reconstruct the semantic variable $\widehat{\boldsymbol{m}}_r$, which is here the weighted average of the velocity models from the neighboring agents.

the elements of the neighbor set $\mathcal{N}_r$: The agent r itself will be denoted by r, and all neighbors of agent r will be denoted by $r\text{-}n_1, r\text{-}n_2, \ldots, r\text{-}n_{N_r}$. Furthermore, we omit the algorithm iteration index k of the ATC-FWI to simplify the notation.

C. Semantic Communication System Model

For each agent to compute (11a), its neighbors have to share the velocity gradients $\delta\boldsymbol{m}_\ell$. Likewise, for each agent to compute (11b), its neighbors have to share the velocity models $\widetilde{\boldsymbol{m}}_\ell$. Our semantic communication system model for the exchange of the velocity gradients $\delta\boldsymbol{m}_\ell$ is shown in Fig. 2, and for the exchange of the velocity models $\widetilde{\boldsymbol{m}}_\ell$ it is shown in Fig. 3.

Although the velocity gradients $\delta\boldsymbol{m}_\ell$ and the velocity models $\widetilde{\boldsymbol{m}}_\ell$ are different quantities, the concept of semantic communication will be applied to each of them similarly. Therefore, to avoid redundancy, the semantic communication system model and optimization problem will only be shown for the exchange of the velocity models $\widetilde{\boldsymbol{m}}_\ell$ in the following.

We define a semantic source as the joint probability distribution $p(\widetilde{\boldsymbol{m}}_r, \widetilde{\boldsymbol{m}}_{r\text{-}n_1}, \ldots, \widetilde{\boldsymbol{m}}_{r\text{-}n_{N_r}}, \boldsymbol{m}_r)$ of the observations $\widetilde{\boldsymbol{m}}_r, \widetilde{\boldsymbol{m}}_{r\text{-}n_1}, \ldots, \widetilde{\boldsymbol{m}}_{r\text{-}n_{N_r}}$ and the semantic variable $\boldsymbol{m}_r$. The semantic variable in this case is defined as $\boldsymbol{m}_r = \sum_{\ell\in\mathcal{N}_r} a_{\ell r}\widetilde{\boldsymbol{m}}_\ell$. The goal at the receiver is to reconstruct the semantic variable $\boldsymbol{m}_r$, which together with the spatially distributed observations $\widetilde{\boldsymbol{m}}_r, \widetilde{\boldsymbol{m}}_{r\text{-}n_1}, \ldots, \widetilde{\boldsymbol{m}}_{r\text{-}n_{N_r}}$ follows the joint distribution of the semantic source $p(\widetilde{\boldsymbol{m}}_r, \widetilde{\boldsymbol{m}}_{r\text{-}n_1}, \ldots, \widetilde{\boldsymbol{m}}_{r\text{-}n_{N_r}}, \boldsymbol{m}_r)$.

The "local" observation $\widetilde{\boldsymbol{m}}_r$ should not be communicated since the agent r is already aware of it. Instead, we will treat this information differently and utilize it to properly "inform" the decoder at the receiver. We will therefore refer to the observation $\widetilde{\boldsymbol{m}}_r$ as *side information* for the decoding process. The encoders $p_{\boldsymbol{\theta}_\ell}(\boldsymbol{c}_\ell|\boldsymbol{s}_\ell)$ at the transmitting agents are modeled as Neural Networks (NNs) with parameters $\boldsymbol{\theta}_\ell$ to encode the observation data $\boldsymbol{s}_\ell$ into symbols $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_{N_r}$, which are transmitted over a wireless channel $p(\boldsymbol{y}|\boldsymbol{c}_1, \ldots, \boldsymbol{c}_{N_r})$, where $\boldsymbol{y}$ is a vector containing all received symbols. In our simulations in Section IV, we focus on the simple case of an Additive White Gaussian Noise (AWGN) channel. However, we note here that our probabilistic channel model is general, with the only requirement that $p(\boldsymbol{y}|\boldsymbol{c}_1, \ldots, \boldsymbol{c}_{N_r})$ needs to be known. Hence, more complex channel models can be directly integrated into our framework. To model the semantic decoder we use a variational approximation $q_{\boldsymbol{\phi}}(\boldsymbol{m}_r|\boldsymbol{y}, \widetilde{\boldsymbol{m}}_r)$ with NN parameters $\boldsymbol{\phi}$, which outputs the distribution of the semantic variable given the received symbols $\boldsymbol{y}$ and the side information $\widetilde{\boldsymbol{m}}_r$. Finally, to get an estimate of the semantic variable $\widehat{\boldsymbol{m}}_r$, we select the most likely $\boldsymbol{m}_r$ from the posterior distribution $q_{\boldsymbol{\phi}}(\boldsymbol{m}_r|\boldsymbol{y}, \widetilde{\boldsymbol{m}}_r)$. The statistical dependencies among all random variables are summarized in Fig. 4.

D. Semantic Communication Problem Formulation

Our objective is to reconstruct the semantic variable $\boldsymbol{m}_r$ at the decoder of each receiving agent. We formulate the optimization problem of the semantic communication system such that the received symbols $\boldsymbol{y}$ are maximally informative about the semantic variable $\boldsymbol{m}_r$, given the side information $\widetilde{\boldsymbol{m}}_r$. This means that we want to find the parameters $\boldsymbol{\theta}_\ell$ for the encoders $p_{\boldsymbol{\theta}_\ell}(\boldsymbol{c}_\ell|\boldsymbol{s}_\ell)$ that maximize the conditional mutual information $I(\boldsymbol{m}_r; \boldsymbol{y}|\widetilde{\boldsymbol{m}}_r)$. To have a realistic communication scenario, we limit the rate of our communication channel by a limited number of channel uses N_{ch} and an average power

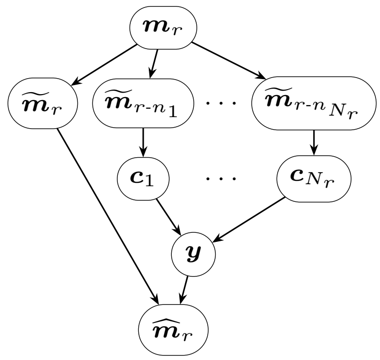


Fig. 4: The statistical dependencies of the semantic variable $\boldsymbol{m}_r$, the observations $\widetilde{\boldsymbol{m}}_\ell$, the transmitted symbols $\boldsymbol{c}_\ell$, the received symbols $\boldsymbol{y}$, and the estimated sematic variable $\widehat{\boldsymbol{m}}_r$.

constraint per channel. This gives us the optimization problem for receiving agent r of

$$\begin{aligned} &\operatorname*{argmax}_{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_{N_r}} I(\boldsymbol{m}_r;\boldsymbol{y}|\widetilde{\boldsymbol{m}}_r) \\ &\text{s.t.} \quad \boldsymbol{y}\in\mathbb{R}^{N_r N_{\text{ch}}},\ P_c\leq 1, \end{aligned} \tag{12}$$

where N_r is the number of transmitters (i.e., the neighboring agents), and P_c is the average signal power available at the decoder per channel use, which is defined differently depending on the particular channel model used (see Section IV-B for the definitions).

The definition of the conditional mutual information between $\boldsymbol{m}_r$ and $\boldsymbol{y}$ given $\widetilde{\boldsymbol{m}}_r$ is given as [40]

$$I(\boldsymbol{m}_r;\boldsymbol{y}|\widetilde{\boldsymbol{m}}_r) = E_{p(\boldsymbol{m}_r,\boldsymbol{y},\widetilde{\boldsymbol{m}}_r)}\left[\log\left(\frac{p(\boldsymbol{m}_r|\boldsymbol{y},\widetilde{\boldsymbol{m}}_r)}{p(\boldsymbol{m}_r|\widetilde{\boldsymbol{m}}_r)}\right)\right], \tag{13}$$

where $\log$ denotes the natural logarithm. When evaluating (13) for the optimization of the conditional mutual information, the optimal decoder $p(\boldsymbol{m}_r|\boldsymbol{y},\widetilde{\boldsymbol{m}}_r)$ is intractable in practice [41]. Thus, the mutual information cannot be maximized directly, but a lower bound on (13) can be derived, which in turn can be optimized. One way of deriving such a lower bound is to use a variational approximation $q_{\boldsymbol{\phi}}(\boldsymbol{m}_r|\boldsymbol{y},\widetilde{\boldsymbol{m}}_r)$, with parameters $\boldsymbol{\phi}$, that approximates $p(\boldsymbol{m}_r|\boldsymbol{y},\widetilde{\boldsymbol{m}}_r)$ [19], [42], [43]. With this approach, a lower bound on the conditional mutual information can be derived (given in Appendix A), which results in

$$\begin{aligned} &I(\boldsymbol{m}_r;\boldsymbol{y}|\widetilde{\boldsymbol{m}}_r) \geq \\ &-E_{p(\boldsymbol{y},\widetilde{\boldsymbol{m}}_r)}[H\left(p(\boldsymbol{m}_r|\boldsymbol{y},\widetilde{\boldsymbol{m}}_r),q_{\boldsymbol{\phi}}(\boldsymbol{m}_r|\boldsymbol{y},\widetilde{\boldsymbol{m}}_r)\right)] \\ &+E_{p(\widetilde{\boldsymbol{m}}_r)}\left[H(p(\boldsymbol{m}_r|\widetilde{\boldsymbol{m}}_r))\right]. \end{aligned} \tag{14}$$

Instead of maximizing the conditional mutual information, now the lower bound in (14) can be maximized. From the derivation it can be seen that the lower bound on the mutual information is tight, if the Kullback-Leibler (KL) divergence $D_{\text{KL}}\left(p(\boldsymbol{m}_r|\boldsymbol{y},\widetilde{\boldsymbol{m}}_r)||q_{\boldsymbol{\phi}}(\boldsymbol{m}_r|\boldsymbol{y},\widetilde{\boldsymbol{m}}_r)\right)=0$, i.e., if the modeled probability distribution of the decoder $q_{\boldsymbol{\phi}}(\boldsymbol{m}_r|\boldsymbol{y},\widetilde{\boldsymbol{m}}_r)$ matches the true posterior distribution $p(\boldsymbol{m}_r|\boldsymbol{y},\widetilde{\boldsymbol{m}}_r)$ almost everywhere [40]. For an intuitive understanding of (14) we can use (25) from Appendix A to get

$$\begin{aligned} &-E_{p(\boldsymbol{y},\widetilde{\boldsymbol{m}}_r)}\left[H\left(p(\boldsymbol{m}_r|\boldsymbol{y},\widetilde{\boldsymbol{m}}_r),q_{\boldsymbol{\phi}}(\boldsymbol{m}_r|\boldsymbol{y},\widetilde{\boldsymbol{m}}_r)\right)\right] \\ &+E_{p(\widetilde{\boldsymbol{m}}_r)}\left[H(p(\boldsymbol{m}_r|\widetilde{\boldsymbol{m}}_r))\right] = \underbrace{I(\boldsymbol{m}_r;\boldsymbol{y}|\widetilde{\boldsymbol{m}}_r)}_{\text{encoder objective}} \\ &\underbrace{-E_{p(\boldsymbol{y},\widetilde{\boldsymbol{m}}_r)}\left[D_{\text{KL}}\left(p(\boldsymbol{m}_r|\boldsymbol{y},\widetilde{\boldsymbol{m}}_r)||q_{\boldsymbol{\phi}}(\boldsymbol{m}_r|\boldsymbol{y},\widetilde{\boldsymbol{m}}_r)\right)\right]}_{\text{decoder objective}}, \end{aligned} \tag{15}$$

which shows that maximizing the conditional mutual information lower bound can be seen as optimizing an implicit encoder and decoder objective. The objective of the encoder is to maximize the mutual information of the encoded symbols after transmission over the channel $\boldsymbol{y}$ to be maximally informative about the semantic variable $\boldsymbol{m}_r$ given the side information $\widetilde{\boldsymbol{m}}_r$. The objective of the decoder is to find the distribution $q_{\boldsymbol{\phi}}(\boldsymbol{m}_r|\boldsymbol{y},\widetilde{\boldsymbol{m}}_r)$ with the minimum KL divergence to the true decoder distribution $p(\boldsymbol{m}_r|\boldsymbol{y},\widetilde{\boldsymbol{m}}_r)$.

For the optimization, the conditional entropy term $E_{p(\widetilde{\boldsymbol{m}}_r)}\left[H(p(\boldsymbol{m}_r|\widetilde{\boldsymbol{m}}_r))\right]$ can be omitted, as it is independent of the encoder and decoder parameters. This gives us the following minimization problem of the encoder and decoder parameters $\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_{N_r}$ and $\boldsymbol{\phi}$, respectively, with

$$\begin{aligned} &\operatorname*{argmin}_{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_{N_r},\boldsymbol{\phi}} E_{p(\boldsymbol{y},\widetilde{\boldsymbol{m}}_r)}\left[H\left(p(\boldsymbol{m}_r|\boldsymbol{y},\widetilde{\boldsymbol{m}}_r),q_{\boldsymbol{\phi}}(\boldsymbol{m}_r|\boldsymbol{y},\widetilde{\boldsymbol{m}}_r)\right)\right] \\ &\text{s.t.} \quad \boldsymbol{y}\in\mathbb{R}^{N_r N_{\text{ch}}},\ P_c\leq 1. \end{aligned} \tag{16}$$

Analogously, we can obtain the optimization problem to design the semantic communication system for the transmission of the velocity gradients. This gives us the following minimization problem of the encoder and decoder parameters $\boldsymbol{\vartheta}_1,\ldots,\boldsymbol{\vartheta}_{N_r}$ and $\boldsymbol{\varphi}$, respectively, with

$$\begin{aligned} &\operatorname*{argmin}_{\boldsymbol{\vartheta}_1,\ldots,\boldsymbol{\vartheta}_{N_r},\boldsymbol{\varphi}} E_{p(\boldsymbol{y},\delta\boldsymbol{m}_r)}\left[H\left(p(\overline{\delta\boldsymbol{m}}_r|\boldsymbol{y},\delta\boldsymbol{m}_r),q_{\boldsymbol{\varphi}}(\overline{\delta\boldsymbol{m}}_r|\boldsymbol{y},\delta\boldsymbol{m}_r)\right)\right] \\ &\text{s.t.} \quad \boldsymbol{y}\in\mathbb{R}^{N_r N_{\text{ch}}},\ P_c\leq 1. \end{aligned} \tag{17}$$

In the following section, we discuss how the minimization of (16) and (17) can be implemented in practice using DNNs.

E. Implementation Considerations

We want to optimize the parameters $\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_{N_r}$ and $\boldsymbol{\phi}$ of the encoders and the decoder for the transmission of the velocity models (16), and the parameters $\boldsymbol{\vartheta}_1,\ldots,\boldsymbol{\vartheta}_{N_r}$ and $\boldsymbol{\varphi}$ of the encoders and the decoder for the transmission of the velocity gradients (17), respectively, of our DNN communication systems. For this, training data is required, which we generate as described later in Section IV-D. To optimize the DNNs parameters, we employ Stochastic Gradient Descent (SGD) based optimization and leverage the reparameterization trick to enable efficient gradient estimation with respect to the encoder and decoder parameters (see Appendix B for details).

Furthermore, an efficiently computable loss function is required. However, (16) and (17) are computationally infeasible due to high dimensionality of the distributions $q_{\boldsymbol{\phi}}(\boldsymbol{m}_r|\boldsymbol{y},\widetilde{\boldsymbol{m}}_r)$ and $q_{\boldsymbol{\varphi}}(\overline{\delta\boldsymbol{m}}_r|\boldsymbol{y},\delta\boldsymbol{m}_r)$ with parameters ϕ and φ, respectively.

In the following, we show that under the assumption of a Gaussian distribution for the decoder output distribution, the optimization problems of (16) and (17) become computationally feasible. We show this in the following only for the case of the transmission of the velocity models, as it holds analogously for the case of the velocity gradients. We emphasize here that this assumption is only required for the decoder distribution and is not required for the distribution of the input data.

In many works of semantic communication, where the task is to solve a classification problem, e.g., as in [19], [29], [44], the posterior distribution of the decoder $p(\boldsymbol{m}_r|\boldsymbol{y}, \widetilde{\boldsymbol{m}}_r)$ is a low dimensional discrete probability distribution. However, in our case, $p(\boldsymbol{m}_r|\boldsymbol{y}, \widetilde{\boldsymbol{m}}_r)$ is a continuous, or more precisely, as we operate on float32 numbers on digital computer hardware, a high dimensional discrete distribution. As an example, for a fairly moderate grid size of 100 by 50 float32 numbers for a subsurface image the number of events is about $2^{32\cdot 5000}$, which makes it impossible to estimate a probability for each event. Even assuming that the values at each grid point are independent, it is still infeasible, as there are approximately $5000 \cdot 2^{32} \approx 2 \cdot 10^{10}$ different discrete probability events.

As a consequence, simplification is required to estimate the probability density $q_{\boldsymbol{\phi}}(\boldsymbol{m}_r|\boldsymbol{y}, \widetilde{\boldsymbol{m}}_r)$. Different parametric and non-parametric methods for probability density estimation methods exist [45]. Here, we use a common approach of assuming independent Gaussian distributions for every i-th element m_{r_i} of $\boldsymbol{m}_r \in \mathbb{R}^{N_\mathrm{x} N_\mathrm{z}}$ for $q_{\boldsymbol{\phi}}(\boldsymbol{m}_r|\boldsymbol{y}, \boldsymbol{s}_0)$ with fixed variance σ^2 and means $\boldsymbol{\mu}(\boldsymbol{y}, \widetilde{\boldsymbol{m}}_r)$. The i-th mean $\mu_i(\boldsymbol{y}, \widetilde{\boldsymbol{m}}_r)$ is the i-th output of the decoder $\boldsymbol{\mu}(\boldsymbol{y}, \widetilde{\boldsymbol{m}}_r)$, which depends on the decoder inputs $\boldsymbol{y}$ and $\widetilde{\boldsymbol{m}}_r$. We mention that this assumption of the decoder distribution $q_{\boldsymbol{\phi}}(\boldsymbol{m}_r|\boldsymbol{y}, \boldsymbol{s}_0)$ essentially assumes the Gaussianity and uncorrelatedness of the residual errors of the estimate of the semantic variable $\boldsymbol{m}_r$. In practice this might not be guaranteed to hold, however, our simulation results show that good communication performance can be achieved under this assumption. We leave the investigation of different parametric distributions as further research work.

Given this assumption, our optimization objective of the cross entropy (16) becomes

$$\frac{1}{2\sigma^2} E_{p(\boldsymbol{m}_r, \boldsymbol{y}, \widetilde{\boldsymbol{m}}_r)} \left[\|\boldsymbol{m}_r - \boldsymbol{\mu}(\boldsymbol{y}, \widetilde{\boldsymbol{m}}_r)\|_2^2 \right] + N_\mathrm{x} N_\mathrm{z} \log \sqrt{2\pi\sigma^2}, \tag{18}$$

with its derivation shown in Appendix C. We see that under the given assumption our optimization objective becomes the Mean Square Error (MSE) loss $E_{p(\boldsymbol{m}_r, \boldsymbol{y}, \widetilde{\boldsymbol{m}}_r)} \left[\|\boldsymbol{m}_r - \boldsymbol{\mu}(\boldsymbol{y}, \widetilde{\boldsymbol{m}}_r)\|_2^2 \right]$ [46]. Estimating the gradient for a batch of training data for (18) with respect to the encoder and decoder parameters $\boldsymbol{\phi}, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{N_r}$ is done following (29) and (32) in Appendix B. In an analogous way, the MSE loss can be derived for the optimization problem (17) to design the semantic communication system to exchange the velocity gradients.

IV. Performance Evaluation

In the following, we evaluate the ATC-FWI algorithm together with our proposed semantic DFC. First, we introduce the communication systems that serve as reference systems for our semantic DFC. Then the channel models, the used NN architectures, and the used training dataset are explained. In Section IV-E the communication performance is evaluated, and in Section IV-F the ATC-FWI imaging performance is evaluated for the different communication systems.

A. Description of Reference Systems

Let us introduce the communication methods that serve as reference systems for our proposed semantic DFC. The different systems are shown in Figs. 5(a) to 5(f). To avoid redundancy, Fig. 5 presents only the communication systems for exchanging the velocity models $\widetilde{\boldsymbol{m}}_\ell$, as the setup for the velocity gradient exchange is analogous. We note again, that in the case of transmission of the velocity model, the observations are $\boldsymbol{m}_\ell$, and the semantic variable is $\boldsymbol{m}_r = \sum_{\ell \in \mathcal{N}_r} a_{\ell r} \widetilde{\boldsymbol{m}}_\ell$. For transmission of the gradients, the observations are $\delta \boldsymbol{m}$, and the semantic variable is given as $\overline{\delta \boldsymbol{m}}_r = \sum_{\ell \in \mathcal{N}_r} b_{\ell r} \delta \boldsymbol{m}_\ell$. In our simulations, we set the coefficients $a_{\ell r}$ and $b_{\ell r}$ to be $1/|\mathcal{N}_r|$. This means that the function F in Figs. 5(a) to 5(d) is defined to compute the arithmetic mean.

As simplest baseline scheme we employ a classical digital Separate Source and Channel Coding (SSCC) transmission system and compare it to the proposed approaches, see Fig. 5(a). For source coding of the classical communication system Huffman coding, PNG, or JPEG were considered, together with subsampling, quantization, and for the gradient dataset also cutting part of the image, to reduce the number of bits required to be sent over the channel. In our case with the size of the data $N_\mathrm{x} N_\mathrm{z} = 5\,000$, PNG and JPEG were not suitable, as both require too much overhead data, making compression to very small number of channel uses not possible. We therefore use Huffmann coding for our evaluation. For channel coding, we use a 5G LDPC code implementation from [47]. For the modulation, 16-QAM and 64-QAM were considered. All parameters were optimized using grid search by using 10 000 data samples from the training dataset to find the best parameter setting at the training SNR. The coding rate is selected such that the total number of channel uses of the digital and analog transmission of the JSCC approaches is the same. In the distributed ATC-FWI algorithm, data must be exchanged for every algorithm step, requiring immediate transmission. Therefore, each transmission has to be done in small blocks, which is known to limit the performance of SSCC communication systems [28].

For a more suitable benchmark, a classic JSCC architecture is used where no side information is used at the decoder, which is shown in Fig. 5(b). For the case of JSCC without side information, each encoder/decoder pair is optimized to individually reconstruct the observation $\widehat{\widetilde{\boldsymbol{m}}}_\ell$ for the velocity model dataset or $\widehat{\delta \boldsymbol{m}}_\ell$ for the velocity gradient dataset. To see how the availability of side information impacts the encoder/decoder design, we compare the case of JSCC without side information to JSCC with side information, which is shown in Fig. 5(c). All following methods use the side information available at the decoder. For JSCC (with side information), the encoder pairs are optimized to individually reconstruct the observation $\widehat{\widetilde{\boldsymbol{m}}}_\ell$ or $\widehat{\delta \boldsymbol{m}}_\ell$, respectively. As we consider the side information for the encoder/decoder design for all further compared methods, we do not explicitly mention the consideration of side information to simplify the notation, and only mention it explicitly if no side information is used. The next approach, referred to as distributed JSCC is shown in Fig. 5(d). It uses a joint decoder for all observations $\widehat{\widetilde{\boldsymbol{m}}}_{r\text{-}n_1}, \dots, \widehat{\widetilde{\boldsymbol{m}}}_{r\text{-}n_{N_r}}$ or $\widehat{\delta \boldsymbol{m}}_{r\text{-}n_1}, \dots, \widehat{\delta \boldsymbol{m}}_{r\text{-}n_{N_r}}$, respectively, to exploit the correlation between the observations 1 to N_r.

Furthermore, we compare the proposed semantic JSCC and the semantic JSCC AirComp methods, which are shown in Figs. 5(e) and 5(f), respectively. Both proposed methods directly reconstruct the semantic variables $\boldsymbol{m}_r$ or $\overline{\delta \boldsymbol{m}_r}$, respectively, as DFC, instead of reconstructing the individual observations, as only the semantic variable is required for executing the ATC-FWI algorithm. The benefit of the semantic JSCC AirComp approach compared to the semantic JSCC approach is that all agents can transmit on the same channel resources, potentially using the channel resources more efficiently. This comes with the constraint of formulating the transmitted data as a nomographic function, and with the drawback that synchronization is required [38], [39].

B. Channel Models

For the evaluation, we consider N_r channels from each neighboring agents to the receiving agent as independent and identically distributed (i.i.d.) AWGN channels. The average signal power per channel use at the decoder P_c is given as $P_c = \frac{1}{N_r N_{\text{ch}}} \sum_{l=1}^{N_r N_{\text{ch}}} |c_i|^2$, where c_i is the i-th entry of the vector $\boldsymbol{c}$, which is given as

$$\boldsymbol{c} = \begin{cases} \boldsymbol{c}_1 + \dots + \boldsymbol{c}_{N_r} & \text{for AirComp channel,} \\ \begin{bmatrix} \boldsymbol{c}_1^\top & \dots & \boldsymbol{c}_{N_r}^\top \end{bmatrix}^\top & \text{otherwise.} \end{cases} \tag{19}$$

For the AirComp channel, the power of the additive noise is σ_n^2. Otherwise, for the N_r channels we assume noise powers of $\sigma_{n_1}^2 = \dots = \sigma_{n_{N_r}}^2 = \sigma_n^2$. The noise power σ_n^2 is scaled according to $\text{SNR} = \frac{P_c}{\sigma_n^2}$.

Finally, the total number of channel uses for neighboring agents transmitting to the receiving agent is always $N_r N_{\text{ch}}$, i.e., for the case AirComp channel, each user has $N_r N_{\text{ch}}$ channel uses, and otherwise each of the N_r transmitting agents has N_{ch} channel uses. For AirComp, we neglect synchronization offset errors in this work, thus providing an upper bound on the communication performance of AirComp. We mention that many works exist on agent synchronization from the research area of collaborative distributed beamforming, where precise synchronization is required as in AirComp [48]. Synchronization among agents is commonly achieved using iterative algorithms that incorporate feedback from the receiver [48]. However, we still note that in future work, the effect of synchronization errors should be investigated for semantic communication with AirComp.

C. Neural Network Architecture

For the encoder and decoder we use a Convolutional Neural Network (CNN) architecture [49]. Each encoder consists of five 2D-convolution layers with 60 5×5 convolution filters

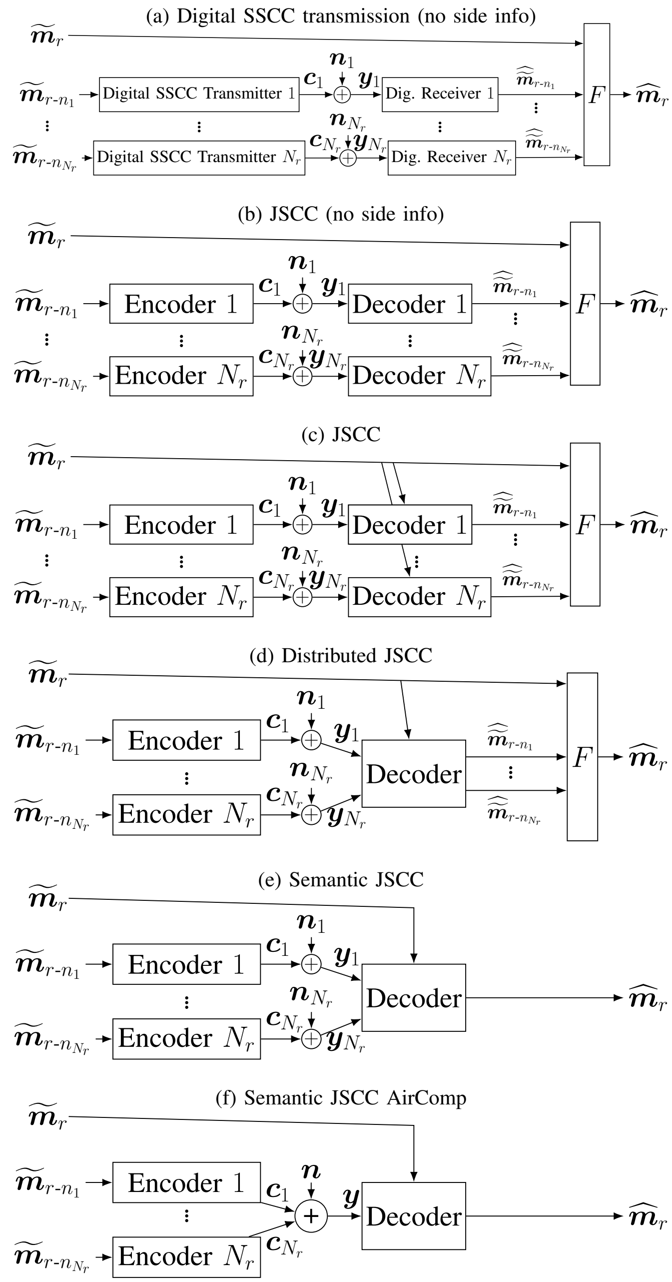


Fig. 5: Comparison of transmission architectures for sharing neighboring velocity models $\widetilde{\boldsymbol{m}}_\ell$ with agent r. Side information is always used, unless stated otherwise: (a) Digital transmission with SSCC without side information (b) JSCC without side information, (c) JSCC, (d) distributed JSCC, (e) semantic JSCC, and (f) semantic JSCC AirComp.

each with batch normalization and PReLU activation function, with respective strides of $\{2, 2, 2, 1, 1\}$, followed by 3 fully connected layers with PReLU activations except for the last layer, where no activation function is used. The same setup is used for the decoder, but at the start 3 fully connected layers are used, followed by transposed convolution layers with respective strides of $\{1, 1, 2, 2, 2\}$. The number of filters of the last encoder layer is selected to match the number of channel uses N_{ch}, and the last decoder layer has just one filter to output a 100 by 50 grid of the velocity model/the gradients. Furthermore, data normalization is done before the first convolution layer and denormalization after the last decoder layer. To match the average power constraint power normalization is done after the last encoder layer.

For simplification, we use parameter sharing for the encoders and decoders, such that not all combinations of sending and receiving agents need to be trained individually. For the transmissions of the velocity models and the gradients the same neural network architecture is used, but the training is done with different datasets.

D. Training Data Generation and Training

As we use NNs for the encoder and decoder and do end-to-end training, we require training data. To generate the training data, we first generate several scenarios of velocity models as ground truth P-wave velocity models. We consider a $200\,\mathrm{m}$ wide and $100\,\mathrm{m}$ deep two dimensional slice of the subsurface with associated velocity values at grid points every two meters in both directions. We consider three different scenarios of ground truth velocity models. The first two scenarios are rectangular and elliptically shaped velocity anomalies with a constant background, and the third scenario consists of three layers with varying width with different velocities.

To generate the training data for the communication system consisting of the velocity models $\widetilde{\boldsymbol{m}}_\ell$ and the velocity gradients $\delta \boldsymbol{m}_\ell$ for all agents ℓ, the ATC-FWI algorithm is executed with error-free communication links, and the respective velocity models and gradients are stored as training data.

$14\,000$ different velocity ground truth models were generated, and the ATC-FWI algorithm was run for 20 iterations with an initial model of a Gaussian filtered ground truth for each scenario. For the training in total only $80\,000$ training data samples were used due to memory constraints. For the velocity dataset, samples were removed where all agents had nearly identical data to avoid overfitting to the side information, and for the gradient dataset $80\,000$ samples were selected randomly. All models are trained for 300 epochs with a learning rate of 10^{-4}. For the last few epochs the learning rate is lowered stepwise until 10^{-7}.

E. Results - Communication Performance

To evaluate the communication performance for the velocity model dataset, we employ the Normalized Mean Square Error (NMSE) of the semantic variable $\frac{\|\widehat{\boldsymbol{m}}_r-\boldsymbol{m}_r\|_2^2}{\|\boldsymbol{m}_r\|_2^2}$ where $\widehat{\boldsymbol{m}}_r$ is the reconstructed semantic variable at the receiving agent, and $\boldsymbol{m}_r$ is the ground truth semantic variable. Analogously, to evaluate the communication performance for the velocity gradient dataset, we employ the NMSE of the semantic variable $\frac{\|\widehat{\overline{\delta \boldsymbol{m}}}_r-\overline{\delta \boldsymbol{m}}_r\|_2^2}{\|\overline{\delta \boldsymbol{m}}_r\|_2^2}$, where $\widehat{\overline{\delta \boldsymbol{m}}}_r$ is the reconstructed semantic variable at the receiving agent, and $\overline{\delta \boldsymbol{m}}_r$ is the ground truth semantic variable. Fig. 6 shows the NMSE of the semantic variable for all combinations of $N_r = 2$, $N_r = 7$ and $N_{\text{ch}} = 91$, $N_{\text{ch}} = 45$ over different channel SNRs for $15\,000$ evaluation data samples of the velocity model and velocity gradient datasets that were not used for training. The training SNR is set to $10\,\mathrm{dB}$.

For the velocity dataset, the results are shown in Fig. 6(a)–(d). It is observed that considering the side information at the decoder significantly improves the NMSE of the semantic variable for the velocity dataset, especially for $N_r = 7$. This is expected as the correlation between the velocity models from the transmitting agents to the receiving agent is high. The distributed JSCC, the semantic JSCC, and the semantic JSCC AirComp approach achieve a lower NMSE than the JSCC approach, as they consider the correlation of the velocity models between all agents. The three models all achieve nearly the same NMSE.

The achieved NMSE for the semantic JSCC approach can be expected to be the same as the distributed JSCC approach, as the pixel-wise correlation of the velocity models is always non-negative. At least for the case of linear function computation for Gaussian sources, the following result is available in the literature: When the linear function is the sum of two variables and the correlation between the two variables is non-negative, the required sum rate for asymptotically error-free reconstruction is the same regardless of whether the individual variables are reconstructed or the sum is directly reconstructed [50], [51]. Furthermore, if the correlation between the two variables is negative, the required sum rate is smaller if the decoder directly reconstructs the function [51].

For the gradient dataset, where the results are shown in Fig. 6(e)–(h), the absolute value of the correlation of the gradients to be transmitted is lower compared to the velocity dataset. However, there is some negative pixel-wise correlation for the gradient dataset. The digital transmission optimized for $10\,\mathrm{dB}$ SNR fails completely for smaller SNRs. The required compression from $N_{\mathrm{x}} N_{\mathrm{z}} = 5\,000$ grid point values with 32 bits float values per pixel down to $N_{\text{ch}} = 91$ or $N_{\text{ch}} = 45$ channel uses is not possible with high accuracy. For the gradient dataset, using the side information at the decoder resulted in overfitting of the encoders and decoder during training. As a consequence, the side information was not used for the JSCC approach. For the gradient dataset, the improvement in NMSE of the semantic JSCC approach to the distributed JSCC approach is significantly larger compared to the case of the velocity dataset. As discussed above, for the special case of Gaussian sources, the required sum rate can be smaller if the decoder directly reconstructs the function in the semantic JSCC approach compared to the distributed JSCC approach, where all individual variables are reconstructed if the correlation is negative. This effect is observed to be larger for the case of seven transmitters in Fig. 6(g)–(h). Finally, the semantic JSCC AirComp approach outperforms all other methods for the gradient dataset, which is expected here, as AirComp is more advantageous for a larger number of agents, as the shared bandwidth for all agents increases [17].

F. Results - ATC-FWI Algorithm Performance With Communication

We now integrate the various communication systems from the previous section into the ATC-FWI and evaluate its overall imaging performance. Furthermore, we compare our proposed semantic DFC systems with the case where no errors during

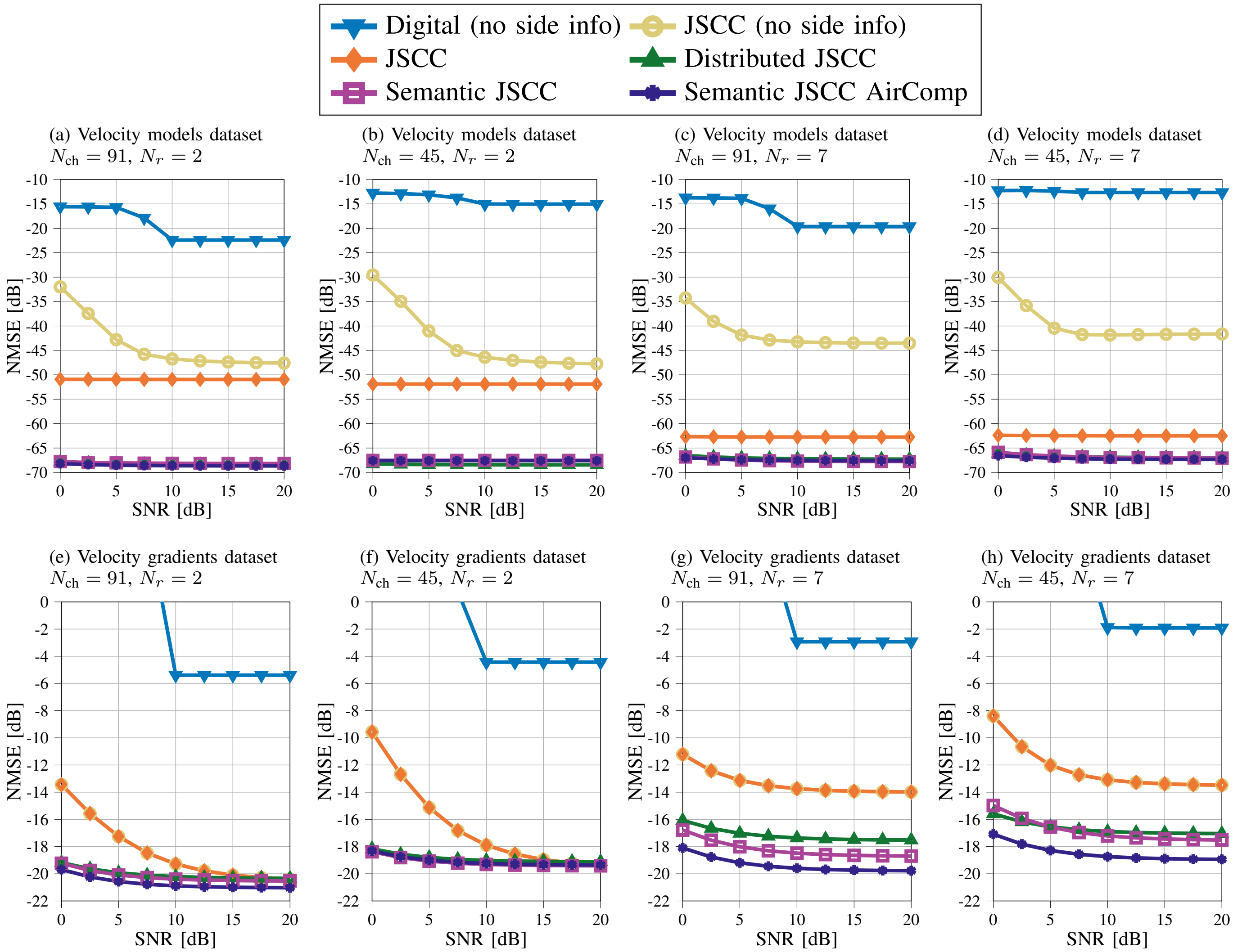


Fig. 6: Communication performance: NMSE of the semantic variable in dB for different channel SNRs is shown. The training SNR is 10 dB. The results are shown for the velocity model evaluation dataset (top half) and the velocity gradient evaluation dataset (bottom half). The number of senders is $N_r = 2$ on the left and $N_r = 7$ on the right. The number of channel uses is $N_{\text{ch}} = 91$ for (a), (c), (e), and (g), and $N_{\text{ch}} = 45$ otherwise.

communication occur, which is referred to as error-free communication in the following.

Our evaluation metric for the ATC-FWI algorithm at iteration k is the NMSE of the estimated velocity averaged over all N_{R} agents, given as

$$\frac{1}{N_{\mathrm{R}}} \sum_{r=1}^{N_{\mathrm{R}}} \frac{\|\boldsymbol{m}_r^{[k]} - \boldsymbol{m}^*\|_2^2}{\|\boldsymbol{m}^*\|_2^2}, \tag{20}$$

where $\boldsymbol{m}^*$ is the velocity model ground truth. Furthermore, we show the global cost (6) of the ATC-FWI algorithm over the number of iterations k. The results are averaged over 100 independent runs with different velocity ground truths from the evaluation dataset. The seismic source locations are fixed with equidistant positions for all simulations with $N_{\mathrm{S}} = 16$. We employ $N_{\mathrm{R}} = 8$ agents for all simulations, where for each run, the agents' positions are selected randomly on a line of the surface of the two dimensional slice of the ground. We investigate a line and a full mesh communication topology. For the line topology every agent is able to communicate only with its direct neighbors, i.e, each agent can communicate with a maximum of two neighbors. This corresponds to the investigation in the previous section with two transmitting agents. Note that the agents at the end of the line array can only communicate with one neighbor. For the full mesh topology, every agent can communicate with every other agent. With eight agents, this corresponds to the investigation in the previous section with seven transmitting neighbors.

The ATC-FWI algorithm requires an adequate initial velocity model. To this end, we use a Gaussian smoothed version of the true velocity model. In practice, distributed traveltime tomography algorithm from [52] can be used to obtain a low resolution subsurface reconstruction that serves as an initial model. Furthermore, we locally stop the ATC-FWI algorithm at each agent if the local cost (7) increases by any value larger than 0 compared to the local cost of the previous iteration, to avoid divergence of the algorithm. Such behavior can arise

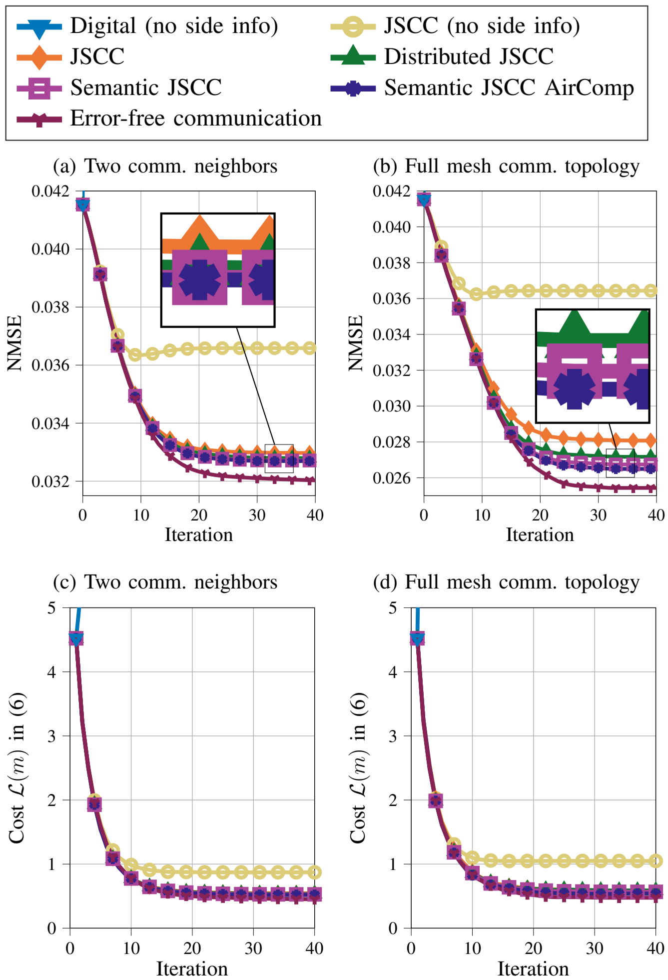


Fig. 7: ATC-FWI algorithm performance: NMSE of the estimated velocity averaged over all agents and the ATC-FWI cost function over the algorithm iterations is shown for a MAS of eight agents with a line communication topology, i.e. two communicating neighbors for each agent, except at the edges only one communicating neighbor (a), (c), and with a full mesh communication topology, i.e. seven communicating neighbors for each agent (b), (d).

from errors occurring in the communication between agents. Figs. 7(c) and 7(d) depict the cost of the ATC-FWI over the iterations.

1) Results over number of iterations: The results for the line topology, where the agents can only communicate with their direct neighbors for $N_{\mathrm{ch}} = 91$, are shown in Fig. 7(a). The SSCC digital communication scheme fails completely for the compression from $N_{\mathrm{x}}N_{\mathrm{z}} = 5\,000$ grid point values with 32 bits per pixel to $N_{\mathrm{ch}} = 91$ channel uses. The transmission errors are larger than the NMSE at initialization leading to algorithm divergence. If no side information is used at the decoder, the performance is also significantly limited by the compression to $N_{\mathrm{ch}} = 91$ channel uses. The semantic JSCC and the semantic JSCC AirComp perform similarly and slightly outperform the other methods in terms of NMSE of the estimated velocity model.

For a full mesh topology (Fig. 7(b)) with $N_{\mathrm{ch}} = 91$, the overall NMSE is significantly lower compared to the line

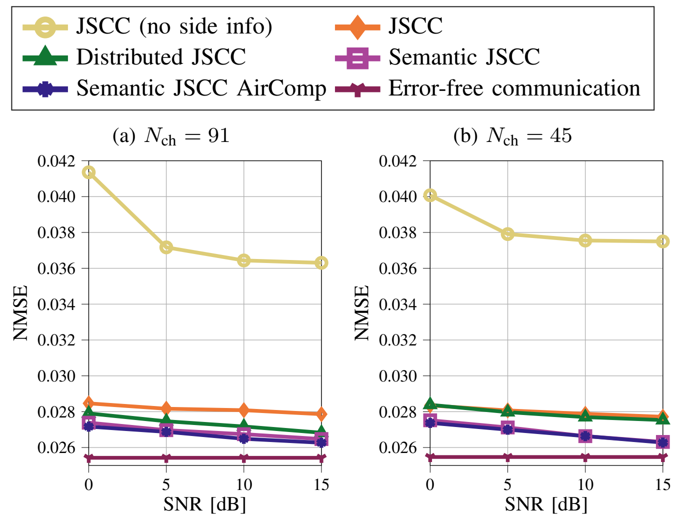


Fig. 8: ATC-FWI algorithm performance: NMSE of the estimated velocity averaged over all agents and the ATC-FWI cost function after 35 algorithm iterations over the communication channel SNR is shown for a MAS of eight agents with a full mesh communication topology (seven communicating neighbors for each agent).

topology after 40 iterations, as the gradients and velocity models can be exchanged more easily between the agents. Furthermore, the advantage of the semantic JSCC and the semantic JSCC AirComp approaches is larger and they approach the performance of the case with error-free communication, which can be explained by the previously observed lower NMSE for the gradient transmission in Fig. 6.

2) Results over SNR: We depict the results of the ATC-FWI algorithm performance for different channel SNRs using a full mesh topology in Fig. 8. The results show the algorithm behavior after 35 iterations. Again, the results are averaged over 100 independent runs with different velocity ground truths from the evaluation dataset. The two cases of $N_{\mathrm{ch}} = 91$ (Fig. 8(a)) and $N_{\mathrm{ch}} = 45$ (Fig. 8(b)) are compared to each other. For a larger SNR the NMSE is smaller, which is as expected. Furthermore, comparing the cases of $N_{\mathrm{ch}} = 91$ to $N_{\mathrm{ch}} = 45$, it can be seen that the advantage of the semantic JSCC and semantic JSCC AirComp is larger compared to the other methods, especially to distributed JSCC, for $N_{\mathrm{ch}} = 45$, which indicates that the semantic approach works well for very high compression.

3) Exemplary result of ATC-FWI for block anomaly: Finally, we show the estimated velocity models for a model with a squared anomaly of $3.20\,\mathrm{km/s}$ over a background velocity of $1.45\,\mathrm{km/s}$. The results shown in Fig. 9 are after 40 iterations of the ATC-FWI algorithm. For all compared methods the estimated velocity models are shown in absolute terms. Furthermore, we depict the difference of each communication method in the imaging results when using error-free communication. The results demonstrate that when errors occur during communication, the velocity of the anomaly is underestimated on average. We observe that our proposed semantic JSCC AirComp outperforms all other methods. In particular, the standard digital communication system fails completely.

V. Conclusion

In this work, we proposed a semantic communication system for distributed function computation within the ATC-FWI, a method that allows cooperative subsurface imaging within a MAS. By means of numerical simulations, we showed that our proposed semantic JSCC AirComp can outperform both classical SSCC digital communication systems and state-of-the-art JSCC methods. For the imaging performance of the distributed ATC-FWI algorithm, we showed that our proposed semantic methods using side information significantly outperform the generic JSCC communication approaches for a strict compression rate. The classic SSCC digital communication method even failed in the imaging task for the strict compression rates. For the case of seven transmitting neighboring agents it was demonstrated that the distributed JSCC approach could be outperformed by the proposed semantic JSCC approach, which again was outperformed by the proposed semantic JSCC AirComp approach. We note that the proposed approach scales to larger MASs, since parameter sharing is employed for the encoders and decoders during training. This prevents the number of trainable parameters from growing with the number of agents. Finally, we mention again, that the investigation on how much a synchronization offset would limit the performance of the proposed semantic JSCC AirComp approach is left as future work.

Appendix A
Derivation of Lower Bound of Conditional Mutual Information

For the following derivation, we use $\boldsymbol{z}$ for $\overline{\delta \boldsymbol{m}_r}$ and $\boldsymbol{s}_0$ for $\delta \boldsymbol{m}_r$ in case of the velocity gradients, and $\boldsymbol{z}$ for $\boldsymbol{m}_r$ and $\boldsymbol{s}_0$ for $\widetilde{\boldsymbol{m}}_r$, in case of the velocity models.

$$
\begin{aligned}
I(\boldsymbol{z};\boldsymbol{y}|\boldsymbol{s}_0) &= E_{p(\boldsymbol{z},\boldsymbol{y},\boldsymbol{s}_0)}\left[\log\left(\frac{p(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{s}_0)}{p(\boldsymbol{z}|\boldsymbol{s}_0)}\right)\right] && (21)\\
&= E_{p(\boldsymbol{z},\boldsymbol{y},\boldsymbol{s}_0)}\left[\log\left(\frac{p(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{s}_0)}{p(\boldsymbol{z}|\boldsymbol{s}_0)}\frac{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{s}_0)}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{s}_0)}\right)\right] && (22)\\
&= E_{p(\boldsymbol{z},\boldsymbol{y},\boldsymbol{s}_0)}\left[\log\left(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{s}_0)\right)\right] && (23)\\
&\quad - E_{p(\boldsymbol{z},\boldsymbol{y},\boldsymbol{s}_0)}\left[\log\left(p(\boldsymbol{z}|\boldsymbol{s}_0)\right)\right]\\
&\quad + E_{p(\boldsymbol{z},\boldsymbol{y},\boldsymbol{s}_0)}\left[\log\left(\frac{p(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{s}_0)}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{s}_0)}\right)\right]\\
&= E_{p(\boldsymbol{y},\boldsymbol{s}_0)}\left[E_{p(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{s}_0)}\left[\log\left(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{s}_0)\right)\right]\right] && (24)\\
&\quad - E_{p(\boldsymbol{z},\boldsymbol{s}_0)}\left[\log\left(p(\boldsymbol{z}|\boldsymbol{s}_0)\right)\right]\\
&\quad + E_{p(\boldsymbol{z},\boldsymbol{y},\boldsymbol{s}_0)}\left[\log\left(\frac{p(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{s}_0)}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{s}_0)}\right)\right]\\
&= -E_{p(\boldsymbol{y},\boldsymbol{s}_0)}\left[H\left(p(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{s}_0), q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{s}_0)\right)\right] && (25)\\
&\quad + E_{p(\boldsymbol{s}_0)}\left[H(p(\boldsymbol{z}|\boldsymbol{s}_0))\right]\\
&\quad + E_{p(\boldsymbol{y},\boldsymbol{s}_0)}\left[D_{\mathrm{KL}}\left(p(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{s}_0)||q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{s}_0)\right)\right]\\
&\geq -E_{p(\boldsymbol{y},\boldsymbol{s}_0)}\left[H\left(p(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{s}_0), q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{s}_0)\right)\right] && (26)\\
&\quad + E_{p(\boldsymbol{s}_0)}\left[H(p(\boldsymbol{z}|\boldsymbol{s}_0))\right],
\end{aligned}
$$

where $E_{p(x)}[\cdot]$ is the expected value with respect to $x \sim p(x)$, $H(\cdot)$ is the Shannon entropy, $H(q,p)$ is the cross entropy

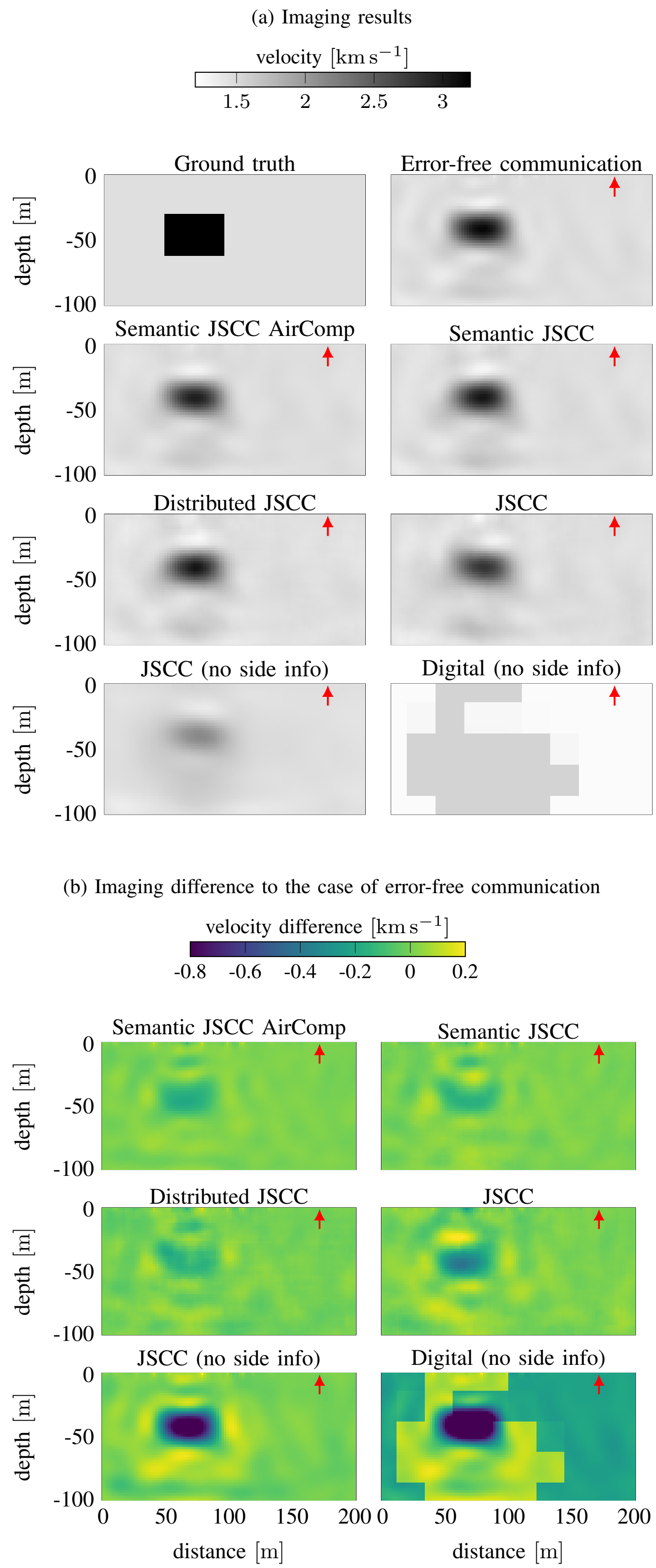


Fig. 9: In (a), the imaging result of the ATC-FWI algorithm for the case of a block anomaly (ground truth, top left image) is shown. In (b), the imaging difference to the error-free communication imaging result is shown. A MAS of eight agents is used with a full mesh communication topology with $N_{\text{ch}} = 91$. The results are shown for the agent positioned at the red arrow after 40 iterations of the ATC-FWI algorithm.

between p and q, and $D_{\mathrm{KL}}(p||q)$ is the KL divergence from p to q. The step for (26) uses the fact that the KL divergence is non-negative [40].

Appendix B
Neural Network Training

In this section, we show how the gradients with respect to the encoder and decoder parameters can be estimated efficiently from the training data. Here, we use $\boldsymbol{z}$ for $\overline{\delta \boldsymbol{m}_r}$ and $\boldsymbol{s}_0, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_{N_r}$ for $\delta \boldsymbol{m}_r, \delta \boldsymbol{m}_{r\text{-}n_1}, \ldots, \delta \boldsymbol{m}_{r\text{-}n_{N_r}}$ in case of the velocity gradients, and $\boldsymbol{z}$ for $\boldsymbol{m}_r$ and $\boldsymbol{s}_0, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_{N_r}$ for $\widetilde{\boldsymbol{m}}_r, \widetilde{\boldsymbol{m}}_{r\text{-}n_1}, \ldots, \widetilde{\boldsymbol{m}}_{r\text{-}n_{N_r}}$, in case of the velocity models. To train the NN, training data samples of the observations $\boldsymbol{s}_0, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_{N_r}$ are used as the inputs, and training data samples of the semantic variable $\boldsymbol{z}$ are used as the desired output.

Using SGD based optimization of the autoencoder, the gradient with respect to the decoder parameters $\boldsymbol{\phi}$ can be approximated for a batch of training data by

$$\frac{\partial}{\partial \boldsymbol{\phi}} E_{p(\boldsymbol{y},\boldsymbol{s}_0)}\left[H\left(p(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{s}_0), q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{s}_0)\right)\right] \tag{27}$$

$$= -E_{p(\boldsymbol{z},\boldsymbol{y},\boldsymbol{s}_0)}\left[\frac{\partial \log\left(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{s}_0)\right)}{\partial \boldsymbol{\phi}}\right] \tag{28}$$

$$\approx -\frac{1}{L}\sum_{l=1}^{L} \frac{\partial \log\left(q_{\boldsymbol{\phi}}(\boldsymbol{z}_l|\boldsymbol{y}_l,\boldsymbol{s}_{0_l})\right)}{\partial \boldsymbol{\phi}}, \tag{29}$$

where L is the batch size and $\boldsymbol{z}_l$, $\boldsymbol{y}_l$, and $\boldsymbol{s}_{0_l}$ are training data samples from the distribution $p(\boldsymbol{z},\boldsymbol{y},\boldsymbol{s}_0)$.

To estimate the gradients with respect to the encoders with parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{N_r}$, we sample from $p_{\boldsymbol{\theta}_\ell}(\boldsymbol{c}_\ell|\boldsymbol{s}_\ell)$, which depends itself on $\boldsymbol{\theta}_\ell$. This leads to high variance in the gradient estimation [41].

To get less noisy gradient estimates for the SGD optimizer, the reparametrization trick can be used by separating $\boldsymbol{y} \sim p(\boldsymbol{y}|\boldsymbol{c}_1, \ldots, \boldsymbol{c}_{N_r})$ into a deterministic and differentiable function $f_{\boldsymbol{\theta}}(\boldsymbol{s},\boldsymbol{n})$ with random variable $\boldsymbol{n}$, which is independent of the $\boldsymbol{\theta}_\ell$, and $\boldsymbol{\theta} = \left[\boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_{N_r}^\top\right]^\top$, $\boldsymbol{s} = \left[\boldsymbol{s}_1^\top, \ldots, \boldsymbol{s}_{N_r}^\top\right]^\top$ [41]. In our case, we can apply the reparametrization trick by assuming deterministic encoders $p_{\boldsymbol{\theta}_\ell}(\boldsymbol{c}_\ell|\boldsymbol{s}_\ell)$ and additive noise $\boldsymbol{n}$ for the wireless channel. Then the gradients with respect to $\boldsymbol{\theta}_\ell$ can be written as

$$\frac{\partial}{\partial \boldsymbol{\theta}_\ell} E_{p(\boldsymbol{y},\boldsymbol{s}_0)}\left[H\left(p(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{s}_0), q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{s}_0)\right)\right] \tag{30}$$

$$= -E_{p(\boldsymbol{z},\boldsymbol{s}_0,\boldsymbol{s}_1,\ldots,\boldsymbol{s}_{N_r})p(\boldsymbol{n})}\left[\frac{\partial f_{\boldsymbol{\theta}}(\boldsymbol{s},\boldsymbol{n})}{\partial \boldsymbol{\theta}_\ell}\frac{\partial \log\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{s}_0)\right]}{\partial \boldsymbol{y}}\right] \tag{31}$$

$$\approx -\frac{1}{L}\sum_{l=1}^{L} \frac{\partial f_{\boldsymbol{\theta}}(\boldsymbol{s}_l,\boldsymbol{n}_l)}{\partial \boldsymbol{\theta}_\ell}\left.\frac{\partial \log\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}_l|\boldsymbol{y},\boldsymbol{s}_{0_l})\right]}{\partial \boldsymbol{y}}\right|_{\boldsymbol{y}=f_{\boldsymbol{\theta}}(\boldsymbol{s}_l,\boldsymbol{n}_l)}, \tag{32}$$

where $\boldsymbol{s}_l$ and $\boldsymbol{n}_l$ are training data samples of $\boldsymbol{s}$ and $\boldsymbol{n}$, respectively.

Appendix C
Derivation of the optimization objective under Gaussian assumption

Here, we use $\boldsymbol{z}$ for $\overline{\delta \boldsymbol{m}_r}$ and $\boldsymbol{s}_0$ for $\delta \boldsymbol{m}_r$ in case of the velocity gradients, and $\boldsymbol{z}$ for $\boldsymbol{m}_r$ and $\boldsymbol{s}_0$ for $\widetilde{\boldsymbol{m}}_r$, in case of the velocity models. To be able to optimize our communication system, a simplification for the decoder probability density function $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{s}_0)$ is required. Here, we assume independent Gaussian distributions for every i-th element z_i of $\boldsymbol{z} \in \mathbb{R}^{N_\mathrm{x}N_\mathrm{z}}$ for $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{s}_0)$ with fixed variance σ^2 and means $\boldsymbol{\mu}(\boldsymbol{y},\boldsymbol{s}_0)$. The i-th mean $\mu_i(\boldsymbol{y},\boldsymbol{s}_0)$ is the i-th output of the decoder $\boldsymbol{\mu}(\boldsymbol{y},\boldsymbol{s}_0)$, which depends on the decoder inputs $\boldsymbol{y}$ and $\boldsymbol{s}_0$. Given this assumption, our optimization objective of the cross entropy can be simplified to

$$E_{p(\boldsymbol{y},\boldsymbol{s}_0)}\left[H\left(p(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{s}_0), q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{s}_0)\right)\right] \tag{33}$$

$$= E_{p(\boldsymbol{z},\boldsymbol{y},\boldsymbol{s}_0)}\left[\log\left(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{s}_0)\right)\right] \tag{34}$$

$$= E_{p(\boldsymbol{z},\boldsymbol{y},\boldsymbol{s}_0)}\left[\log\left(\prod_{i=1}^{N_\mathrm{x}N_\mathrm{z}} \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(z_i-\mu_i(\boldsymbol{y},\boldsymbol{s}_0))^2}{2\sigma^2}\right)\right)\right] \tag{35}$$

$$= E_{p(\boldsymbol{z},\boldsymbol{y},\boldsymbol{s}_0)}\left[\frac{1}{2\sigma^2}\sum_{i=1}^{N_\mathrm{x}N_\mathrm{z}}(z_i-\mu_i(\boldsymbol{y},\boldsymbol{s}_0))^2\right] + N_\mathrm{x}N_\mathrm{z}\log\sqrt{2\pi\sigma^2} \tag{36}$$

$$= \frac{1}{2\sigma^2}E_{p(\boldsymbol{z},\boldsymbol{y},\boldsymbol{s}_0)}\left[\|\boldsymbol{z}-\boldsymbol{\mu}(\boldsymbol{y},\boldsymbol{s}_0)\|_2^2\right] + N_\mathrm{x}N_\mathrm{z}\log\sqrt{2\pi\sigma^2}. \tag{37}$$

References

[1] S. Zhang, R. Pohlmann, T. Wiedemann, A. Dammann, H. Wymeersch, and P. A. Hoeher, "Self-aware swarm navigation in autonomous exploration missions," *Proceedings of the IEEE*, vol. 108, no. 7, p. 1168–1195, Jul. 2020. [Online]. Available: http://dx.doi.org/10.1109/JPROC.2020.2985950

[2] R. J. Amala Arokia Nathan, S. Strand, D. Mehrwald, D. Shutin, and O. Bimber, "An autonomous drone swarm for detecting and tracking anomalies among dense vegetation," *Communications Engineering*, vol. 4, no. 1, p. 205, 2025.

[3] G. Rabideau, J. Russino, A. Branch, N. Dhamani *et al.*, "Planning, scheduling, and execution on the moon: the cadre technology demonstration mission," 2025. [Online]. Available: https://arxiv.org/abs/2502.14803

[4] T. Wiedemann, D. Shutin, and A. J. Lilienthal, "Model-based gas source localization strategy for a cooperative multi-robot system—a probabilistic approach and experimental validation incorporating physical knowledge and model uncertainties," *Robotics and Autonomous Systems*, vol. 118, p. 66–79, Aug. 2019. [Online]. Available: http://dx.doi.org/10.1016/j.robot.2019.03.014

[5] B.-S. Shin and D. Shutin, "Joint distributed traveltime and full waveform tomography for enhanced subsurface imaging in seismic networks," *IEEE Transactions on Computational Imaging*, vol. 10, p. 600–612, 2024. [Online]. Available: http://dx.doi.org/10.1109/TCI.2024.3366754

[6] E. Rizk, K. Yuan, and A. H. Sayed, "Diffusion learning with partial agent participation and local updates," 2025. [Online]. Available: https://arxiv.org/abs/2505.11307

[7] J. Zhang, Z. Liu, Y. Zhu, E. Shi, B. Xu, C. Yuen, D. Niyato, M. Debbah, S. Jin, B. Ai, and S. Xuemin, "Multi-agent reinforcement learning in wireless distributed networks for 6g," 2025. [Online]. Available: https://arxiv.org/abs/2502.05812

[8] 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.

[9] W. Yang, H. Du, Z. Q. Liew, W. Y. B. Lim, Z. Xiong, D. Niyato, X. Chi, X. Shen, and C. Miao, "Semantic communications for future internet: Fundamentals, applications, and challenges," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 1, pp. 213–250, 2022.

[10] D. Wheeler and B. Natarajan, “Engineering semantic communication: A survey,” *IEEE Access*, vol. 11, pp. 13 965–13 995, 2023.

[11] W. Xu, Z. Yang, D. W. K. Ng, M. Levorato, Y. C. Eldar, and M. Debbah, “Edge learning for b5g networks with distributed signal processing: Semantic communication, edge computing, and wireless sensing,” *IEEE journal of selected topics in signal processing*, vol. 17, no. 1, pp. 9–39, 2023.

[12] H. Kalita, A. Quintero, A. Wissing, B. Haugh, C. Angie, G. Nail, J. Wilson, J. Richards, J. Landin, K. Kukkala, M. Vazquez, N. Tan, Q. Lamey, R. Lu, R. Peralta, V. Vilvanathan, and J. Thangavelautham, “Evaluation of lunar pits and lava tubes for use as human habitats,” in *Earth and Space 2021*. American Society of Civil Engineers, Apr. 2021, p. 944–957. [Online]. Available: http://dx.doi.org/10.1061/9780784483374.086

[13] J.-P. de la Croix, F. Rossi, R. Brockers, D. Aguilar, K. Albee, E. Boroson, A. Cauligi, J. Delaune, R. Hewitt, D. Kogan, G. Lim, B. Morrell, Y. Nakka, V. Nguyen, P. Proença, G. Rabideau, J. Russino, M. S. da Silva, G. Zohar, and S. Comandur, “Multi-agent autonomy for space exploration on the cadre lunar technology demonstration,” in *2024 IEEE Aerospace Conference*. IEEE, Mar. 2024, p. 1–14. [Online]. Available: http://dx.doi.org/10.1109/AERO58975.2024.10521425

[14] A. H. Sayed *et al.*, “Diffusion strategies for adaptation and learning over networks,” *IEEE Signal Processing Magazine*, pp. 155–171, 2013.

[15] W. Shi, Q. Ling, G. Wu, and W. Yin, “Extra: An exact first-order algorithm for decentralized consensus optimization,” *SIAM Journal on Optimization*, vol. 25, no. 2, p. 944–966, Jan. 2015. [Online]. Available: http://dx.doi.org/10.1137/14096668X

[16] N. Ma and P. Ishwar, “Some results on distributed source coding for interactive function computation,” *IEEE Transactions on Information Theory*, vol. 57, no. 9, pp. 6180–6195, 2011.

[17] O. Abari, H. Rahul, and D. Katabi, “Over-the-air function computation in sensor networks,” *arXiv preprint arXiv:1612.02307*, 2016.

[18] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, and L. Bottou, “Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion.” *Journal of machine learning research*, vol. 11, no. 12, 2010.

[19] E. Beck, C. Bockelmann, and A. Dekorsy, “Semantic information recovery in wireless networks,” *Sensors*, vol. 23, no. 14, p. 6347, 2023.

[20] J. Whang, A. Nagle, A. Acharya, H. Kim, and A. G. Dimakis, “Neural distributed source coding,” *IEEE Journal on Selected Areas in Information Theory*, vol. 5, pp. 493–508, 2024.

[21] J. Virieux and S. Operto, “An overview of full-waveform inversion in exploration geophysics,” *Geophysics*, vol. 74, no. 6, 2009.

[22] A. Fichtner, F. Bleibinhaus, and Y. Capdeville, *Full seismic waveform modelling and inversion*. Springer Berlin, 2011.

[23] F. Li, M. Valero, Y. Cheng, T. Bai, and W. Song, “Distributed sensor networks based shallow subsurface imaging and infrastructure monitoring,” *IEEE Transactions on Signal and Information Processing over Networks*, vol. 6, p. 241–250, 2020. [Online]. Available: http://dx.doi.org/10.1109/TSIPN.2020.2975349

[24] S. Wang, M. P. Panning, S. D. Vance, and W. Song, “Underground microseismic event monitoring and localization within sensor networks,” *Sensors*, vol. 21, no. 8, p. 2830, Apr. 2021. [Online]. Available: http://dx.doi.org/10.3390/s21082830

[25] R. E. Plessix, “A review of the adjoint-state method for computing the gradient of a functional with geophysical applications,” *Geophys. J. Int.*, 2006.

[26] M. Louboutin, M. Lange, F. Luporini, N. Kukreja, P. A. Witte, F. J. Herrmann, P. Velesko, and G. J. Gorman, “Devito (v3.1.0): an embedded domain-specific language for finite differences and geophysical exploration,” *Geoscientific Model Development*, vol. 12, no. 3, p. 1165–1187, Mar. 2019. [Online]. Available: http://dx.doi.org/10.5194/gmd-12-1165-2019

[27] J. Xu, T.-Y. Tung, B. Ai, W. Chen, Y. Sun, and D. Gündüz, “Deep joint source-channel coding for semantic communications,” *IEEE communications Magazine*, vol. 61, no. 11, pp. 42–48, 2023.

[28] D. Gündüz, M. A. Wigger, T.-Y. Tung, P. Zhang, and Y. Xiao, “Joint source–channel coding: Fundamentals and recent progress in practical designs,” *Proceedings of the IEEE*, pp. 1–32, 2024.

[29] A. Halimi Razlighi, M. H. V. Tillmann, E. Beck, C. Bockelmann, and A. Dekorsy, “Cooperative and collaborative multi-task semantic communication for distributed sources,” in *ICC 2025-IEEE International Conference on Communications*. IEEE, 2025, pp. 3966–3971.

[30] S. Wang, K. Yang, J. Dai, and K. Niu, “Distributed image transmission using deep joint source-channel coding,” in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 5208–5212.

[31] N. Mital, E. Özyilkan, A. Garjani, and D. Gündüz, “Neural distributed image compression with cross-attention feature alignment,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2023, pp. 2498–2507.

[32] H. Yang, T. Ding, and X. Yuan, “Federated learning with lossy distributed source coding: Analysis and optimization,” *IEEE Transactions on Communications*, vol. 71, no. 8, pp. 4561–4576, 2023.

[33] K. Yang, T. Jiang, Y. Shi, and Z. Ding, “Federated learning via over-the-air computation,” *IEEE Transactions on Wireless Communications*, vol. 19, no. 3, pp. 2022–2035, 2020.

[34] H. Ye, G. Y. Li, and B.-H. F. Juang, “Deep over-the-air computation,” in *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*, 2020, pp. 1–6.

[35] S. Wei, C. Feng, C. Guo, B. Zhang, and J. Chen, “Deep over-the-air computation based multi-user semantic communication for smart cities,” in *2024 IEEE/CIC International Conference on Communications in China (ICCC Workshops)*, 2024, pp. 622–626.

[36] P. Yi, Y. Cao, X. Kang, and Y.-C. Liang, “Integrated distributed semantic communication and over-the-air computation for cooperative spectrum sensing,” *IEEE Transactions on Communications*, pp. 1–1, 2024.

[37] N. Agrawal, R. L. G. Cavalcante, M. Yukawa, and S. Stańczak, “Distributed convex optimization “over-the-air” in dynamic environments,” *IEEE Transactions on Signal and Information Processing over Networks*, vol. 10, pp. 610–625, 2024.

[38] Z. Wang, Y. Zhao, Y. Zhou, Y. Shi, C. Jiang, and K. B. Letaief, “Over-the-air computation for 6g: Foundations, technologies, and applications,” *IEEE Internet of Things Journal*, vol. 11, no. 14, pp. 24 634–24 658, 2024.

[39] A. Şahin and R. Yang, “A survey on over-the-air computation,” *IEEE Communications Surveys & Tutorials*, vol. 25, no. 3, pp. 1877–1908, 2023.

[40] T. M. Cover, *Elements of information theory*. John Wiley & Sons, 1999.

[41] O. Simeone, *Machine learning for engineers*. Cambridge university press, 2022.

[42] B. Poole, S. Ozair, A. Van Den Oord, A. Alemi, and G. Tucker, “On variational bounds of mutual information,” in *International conference on machine learning*. PMLR, 2019, pp. 5171–5180.

[43] D. Barber and F. Agakov, “The im algorithm: a variational approach to information maximization,” *Advances in neural information processing systems*, vol. 16, no. 320, p. 201, 2004.

[44] A. Halimi Razlighi, C. Bockelmann, and A. Dekorsy, “Semantic communication for cooperative multi-task processing over wireless networks,” *IEEE Wireless Communications Letters*, vol. 13, no. 10, pp. 2867–2871, 2024.

[45] B. W. Silverman, *Density estimation for statistics and data analysis*. Routledge, 2018.

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

[47] J. Hoydis, S. Cammerer, F. A. Aoudia, A. Vem, N. Binder, G. Marcus, and A. Keller, “Sionna: An open-source library for next-generation physical layer research,” *arXiv preprint arXiv:2203.11854*, 2022.

[48] S. Jayaprakasam, S. K. A. Rahim, and C. Y. Leow, “Distributed and collaborative beamforming in wireless sensor networks: Classifications, trends, and research directions,” *IEEE Communications Surveys & Tutorials*, vol. 19, no. 4, pp. 2092–2116, 2017.

[49] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, “A survey of convolutional neural networks: analysis, applications, and prospects,” *IEEE transactions on neural networks and learning systems*, vol. 33, no. 12, pp. 6999–7019, 2021.

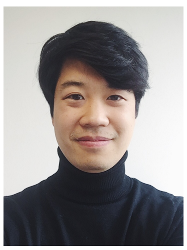
[50] A. B. Wagner, S. Tavildar, and P. Viswanath, “Rate region of the quadratic gaussian two-encoder source-coding problem,” *IEEE Transactions on Information Theory*, vol. 54, no. 5, pp. 1938–1961, 2008.

[51] Y. Yang and Z. Xiong, “Distributed compression of linear functions: Partial sum-rate tightness and gap to optimal sum-rate,” *IEEE Transactions on Information Theory*, vol. 60, no. 5, pp. 2835–2855, 2014.

[52] B.-S. Shin and D. Shutin, “Distributed traveltime tomography using kernel-based regression in seismic networks,” *IEEE Geoscience and Remote Sensing Letters*, vol. 19, p. 1–5, 2022.

Maximilian H. V. Tillmann (Graduate Student Member, IEEE) received the B.Sc. degree in electrical engineering from RWTH Aachen University, Aachen, Germany, in 2019, the M.E. degree from Keio University, Yokohama, Japan, in 2022, with the T.I.M.E double degree program, and the M.Sc. degree from RWTH Aachen University in 2023. He is currently working towards the Ph.D. degree in electrical engineering with the Department of Communications Engineering (ANT), University of Bremen, Bremen, Germany. His research interests include semantic communication, machine learning for wireless communication, and distributed optimization.

Mr. Tillmann was a recipient of the Friedrich Wilhelm Prize from RWTH Aachen University in 2024 for his master's thesis.



Ban-Sok Shin (Member, IEEE) received the Dipl. Ing. (M.Sc.) and Dr.-Ing. (Ph.D.) degrees in electrical engineering from the University of Bremen, Bremen, Germany, in 2013 and 2020, respectively. From 2013 to 2019, he was a Research Assistant with the Department of Communications Engineering, University of Bremen, working on distributed estimation schemes for environmental monitoring applications. Since 2019, he has been a postdoctoral researcher in the Swarm Exploration Group at the Institute of Communications and Navigation, German Aerospace Center, where he is working on subsurface exploration and imaging using autonomous multi-agent networks. His research interests include subsurface imaging, in particular full waveform inversion, inverse problems, and distributed estimation, and their application to autonomous multi-agent networks.

Dr. Shin was the recipient of the OHB Award for the best Ph.D. dissertation from the Faculty of Electrical Engineering, University of Bremen, in 2021.

Dmitriy Shutin (Senior Member, IEEE) received the master's degree in computer science from Dnipropetrovsk State University, Ukraine, in 2000, and the Ph.D. degree in electrical engineering from the Graz University of Technology, Graz, Austria, in 2006. From 2001 to 2006 and from 2006 to 2009, he was a Teaching Assistant and an Assistant Professor with the Signal Processing and Speech Communication Laboratory, Graz University of Technology. In 2009, he joined the Department of Electrical Engineering, Princeton University, where he worked as a Research Associate until 2011. In 2011, he joined the Institute of Communications and Navigation, German Aerospace Center, where he is currently a Leader of the Swarm Exploration Group. His research interests include machine learning for signal processing, distributed information and signal processing, and swarm exploration.

Dr. Shutin was a recipient of the Best Student Paper Award at the 2005 IEEE International Conference on Information, Communications and Signal Processing (ICICS). In 2009, he was awarded the Erwin Schroedinger Research Fellowship. From 2012 to 2014, he acted as a selected Advisor of German Air Navigation Service Provider within the Navigational System Panel of ICAO.

Armin Dekorsy (Senior Member, IEEE) is a professor at the University of Bremen, where he is the director of the Gauss-Olbers Space Technology Transfer Center and heads the Department of Communications Engineering. With over eleven years of industry experience, including distinguished research positions such as DMTS at Bell Labs and Research Coordinator Europe at Qualcomm, he has actively participated in more than 65 international research projects, with leadership roles in 17 of them. He is a Senior Member of the IEEE Communications and Signal Processing Society and a member of the VDE/ITG Expert Committee on Information and System Theory. He co-authored the textbook *Nachrichtenübertragung, Release 6, Springer Vieweg*, which is a bestseller in the field of communication technologies in German-speaking countries. His research focuses on signal processing and wireless communications for 5G/6G, industrial radio, and 3D networks.