Deep Joint Semantic Coding and Beamforming for Near-Space Airship-Borne Massive MIMO Network

Minghui Wu, Zhen Gao, Zhaocheng Wang, *Fellow, IEEE*, Dusit Niyato, *Fellow, IEEE*,
George K. Karagiannidis, *Fellow, IEEE*, and Sheng Chen, *Life Fellow, IEEE*

***Abstract*—Near-space airship-borne communication network is recognized to be an indispensable component of the future integrated ground-air-space network thanks to airships' advantage of long-term residency at stratospheric altitudes, but it urgently needs reliable and efficient Airship-to-X link. To improve the transmission efficiency and capacity, this paper proposes to integrate semantic communication with massive multiple-input multiple-output (MIMO) technology. Specifically, we propose a deep joint semantic coding and beamforming (JSCBF) scheme for airship-based massive MIMO image transmission network in space, in which semantics from both source and channel are fused to jointly design the semantic coding and physical layer beamforming. First, we design two semantic extraction networks to extract semantics from image source and channel state information, respectively. Then, we propose a semantic fusion network that can fuse these semantics into complex-valued semantic features for subsequent physical-layer transmission. To efficiently transmit the fused semantic features at the physical layer, we then propose the hybrid data and model-driven semantic-aware beamforming networks. At the receiver, a semantic decoding network is designed to reconstruct the transmitted images. Finally, we perform end-to-end deep learning to jointly train all the modules, using the image reconstruction quality at the receivers as a metric. The proposed deep JSCBF scheme fully combines the efficient source compressibility and robust error correction capability of semantic communication with the high spectral efficiency of massive MIMO, achieving a significant performance improvement over existing approaches.**

***Index Terms*—Airship base station, beamforming, massive MIMO, deep learning, semantic communication.**

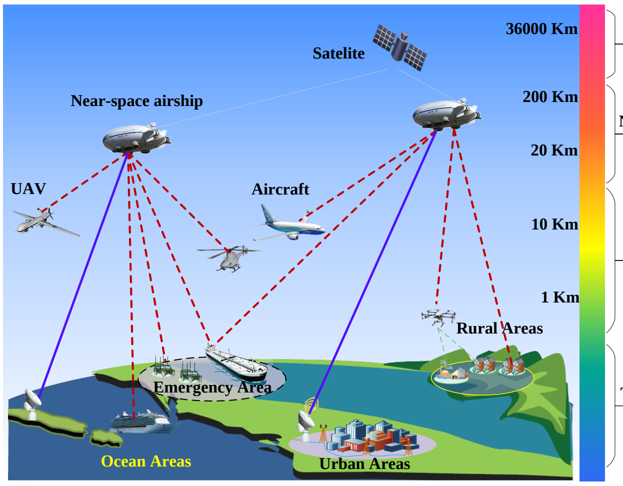


Fig. 1. Schematic diagram of the IGASN architecture, consisting of space satellites, near-space airships, aerial aircraft, and terrestrial BSs.

I. Introduction

In the evolving 6G communications landscape, the integrated ground-air-space network (IGASN), as shown in Fig. 1, is increasingly recognized as a key architecture. Within this framework, the incorporation of an airship-based near-space communications network is seen as a breakthrough extension, offering revolutionary communications capabilities and extending the coverage and efficiency of the network [1]–[4]. Unlike space-based satellite communication systems, which offer wide coverage but suffer from long transmission latency and high costs due to their higher orbital positions, airship-based near-space base stations (BSs) can remain continuously in the stratosphere, closer to Earth, for several years and can be flexibly deployed as needed. This feature provides reduced transmission delays, relatively stable channel propagation conditions, and stronger task-oriented capability for the airship-borne near-space network. Unlike airborne platforms such as unmanned aerial vehicles [5], which are constrained by limited coverage, shorter endurance and regulatory restrictions, airship-based BSs provide broader and uninterrupted network services. In addition, unlike terrestrial BSs that may have obstructed coverage due to geographic features or infrastructure limitations, airship-based BSs can provide unobstructed wide-area coverage [6]. This attribute is particularly beneficial in areas where terrestrial networks are either non-existent, inefficient, damaged or overwhelmed by demand. Given that approximately 70% of the Earth's surface is covered by water, these airship-based BSs are uniquely positioned to provide coverage in vast maritime regions and remote areas beyond the reach of terrestrial BSs. They also offer enhanced robustness and resilience in regions covered by terrestrial BSs, serving as a critical backup for communications support, particularly when ground infrastructure is compromised by natural disasters or other emergencies.

However, the deployment of airship-based near-space communication networks faces the challenge of massive data transmission demands with limited transmission resources. As a novel paradigm in 6G networks, semantic communication can effectively mitigate this problem by achieving higher transmission efficiency [7]–[9]. Building on Weaver and Shannon's definition of semantic information [10], this innovative paradigm shifts the focus to the underlying meaning of symbols rather than the pursuit of precise reconstruction. Semantic communication systems, unlike their traditional counterparts, can significantly compress

The work of Zhen Gao was supported in part by the Natural Science Foundation of China (NSFC) under Grant 62071044 and Grant U2001210; in part by the Shandong Province Natural Science Foundation under Grant ZR2022YQ62; in part by the Beijing Nova Program. The work of Dusit Niyato was supported by the National Research Foundation, Singapore, and Infocomm Media Development Authority under its Future Communications Research & Development Programme, Defence Science Organisation (DSO) National Laboratories under the AI Singapore Programme (AISG Award No: AISG2-RP-2020-019 and FCP-ASTAR-TG-2022-003), Singapore Ministry of Education (MOE) Tier 1 (RG87/22), and the NTU Centre for Computational Technologies in Finance (NTU-CCTF). *(Corresponding author: Zhen Gao.)*

M. Wu is with School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China (e-mail: wuminghui@bit.edu.cn).

Z. Gao is with State Key Laboratory of CNS/ATM, Beijing Institute of Technology (BIT), Beijing 100081, China, also with BIT Zhuhai 519088, China, also with the MIIT Key Laboratory of Complex-Field Intelligent Sensing, BIT, Beijing 100081, China, also with the Advanced Technology Research Institute of BIT (Jinan), Jinan 250307, China, and also with the Yangtze Delta Region Academy, BIT (Jiaxing), Jiaxing 314019, China(e-mail: gaozhen16@bit.edu.cn).

Z. Wang is with Beijing National Research Center for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China, and Z. Wang is also with Tsinghua Shenzhen International Graduate School, Shenzhen 518055, China (e-mail: zcwang@tsinghua.edu.cn).

Dusit Niyato is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798, Singapore (e-mail: dniyato@ntu.edu.sg).

G. K. Karagiannidis is with with the Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece and also with the Artificial Intelligence & Cyber Systems Research Center, Lebanese American University (LAU), Lebanon (e-mail: geokarag@auth.gr).

Sheng Chen is with the School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, U.K. (e-mail: sqc@soton.ac.uk).

source information and reduce associated communication costs. In scenarios where conveying intrinsic meaning is the primary objective, semantic communication will play an indispensable role.

Massive multiple-input multiple-output (MIMO) is a key technology for future wireless communication systems [11]–[15]. By employing large antenna arrays on the airborne BSs coupled with advanced beamforming techniques, the transmission capacity can be significantly increased to meet the massive data transmission requirements of the IGASN. For example, authors in [16] proposed to apply massive MIMO in satellite constellations, which can efficiently support wideband massive user transmission and improve the received signal quality significantly. Therefore, the integration of massive MIMO beamforming and semantic communication techniques in an airship-based near-space communication network [17] represents a very promising communication paradigm by simultaneously exploiting the ability of semantic communication to extract key semantics from the source and the ability of massive MIMO to handle high data rate transmissions for 6G massive communication [18]–[20].

A. State-of the-Art

In the field of semantic communication, the success of deep learning (DL) has inspired the adoption of architectures based on deep neural networks (DNN), such as autoencoders, which have been widely used in semantic communication systems to achieve better performance [9]. Current research in semantic communication is mainly divided into two different directions: the first one focuses on the design of efficient semantic encoding and decoding algorithms [21]–[23], [25]–[28], while the second one is dedicated to the development of algorithms for the physical layer transmission modules of semantic communication [29]–[36].

In the research of semantic coding and decoding algorithms, the transmitter performing semantic coding and the receiver performing semantic decoding are considered as a pair of DNN-based autoencoders, where the encoder at the transmitter semantically extracts and encodes the source data into a complex-valued transmit signal, and the decoder at the receiver decodes the data based on the received symbols. For example, deep joint source-channel coding (JSCC) and its improved versions have been proposed in [21]–[24]. In these approaches, images are mapped into complex-valued transmission symbols by a semantic encoder and reconstructed by a semantic decoder to achieve improved reconstruction performance. Reference [25] further utilizes a diffusion model to enhance the received signal quality, thereby improving the image reconstruction quality in deep JSCC. The authors of [26], [27] designed semantic communication systems based on the attention mechanism for speech transmission, while the authors of [28] proposed the deep JSCC algorithm for multimodal data transmission.

Compared to the first category of research, which focuses on the design of semantic encoding and decoding without considering actual communication scenarios, the second category pays more attention to the enhancement of physical layer transmission modules in various semantic communication scenarios [29]. The authors of [30], [31] studied resource allocation in semantic communication systems. The work [32] investigated the integration of Orthogonal Frequency Division Multiplexing (OFDM) with deep JSCC and optimized the entire semantic communication system in an end-to-end (E2E) manner. The paper [33] designed a semantic-driven constellation to improve the image reconstruction quality in deep JSCC. The authors of [34]–[36] further explored the design of physical layer transmission modules in MIMO semantic communication systems. Specifically, the [34] research focuses on semantic transmission in multi-user MIMO uplink systems, while the [35], [36] research focuses on optimizing power allocation and adaptive channel state information (CSI) feedback code length, respectively, for single-user narrowband MIMO semantic communication systems in the downlink. However, there is currently a lack of research focused on optimizing the downlink multi-user beamforming modules in massive MIMO semantic communication systems.

Traditional non-iterative MIMO beamforming schemes, such as regularized zero forcing (RZF) [37] and signal-to-leakage-and-noise ratio (SLNR) [38] algorithms, are easy to implement, but they tend to achieve suboptimal performance because they do not directly maximize spectral efficiency. To maximize spectral efficiency, the authors of [39]–[41] designed efficient iterative beamforming algorithms based on semidefinite relaxation [39], weighted minimum mean square error (WMMSE) [40], and penalty dual decomposition [41] optimization frameworks. Although these iterative beamforming algorithms offer performance close to theoretical limits, their high computational complexity due to large matrix inversions and high number of iterations, coupled with the dependence on accurate CSI information, hinders their application in multi-user massive MIMO systems under imperfect CSI and unfavorable channel propagation conditions.

Motivated by the success of data-driven DL in various fields, its application to massive MIMO systems has also been actively explored in recent years [42]. The authors of [43]–[45] used convolutional neural networks (CNNs) with supervised training to approximate traditional iterative beamforming algorithms. The authors in [46] proposed a deep learning-based hybrid beamforming scheme for terahertz massive MIMO. The study [47] proposed a beamforming neural network training strategy based on transfer learning for different channel conditions. The authors of [6], [48]–[50] further considered modeling the channel acquisition process and beamforming as an E2E neural network that is jointly trained with the spectral efficiency metric to achieve high spectral efficiency beamforming with limited pilot overhead. These data-driven DL beamforming approaches treat the communication process as a black box and optimize the mapping between inputs and outputs through E2E DL training. This methodology has the advantage of not relying on existing expert knowledge, as it learns transmission strategies directly from data samples, providing better adaptation to imperfect CSI. However, these approaches have limited interpretability and generalization capabilities, and their performance may not be guaranteed.

Consequently, model-driven DL-MIMO beamforming techniques that incorporate expert knowledge have attracted considerable attention in recent years. By unfolding the WMMSE iterative process into a hierarchical network, the authors of [51], [52] developed an efficient iterative design with neural networks for the narrowband MIMO scenario. By integrating traditional successive over-relaxation-based beamforming schemes with DL, the work [53] accelerated the network convergence speed and reduced the number of iterations required. The study [54] derived a WMMSE algorithm in a rate-splitting multiple access scenario and combined it with DNNs to achieve better performance. The authors in [55] applied deep unfolding to the traditional EP algorithm for OTFS systems to enhance performance. However, model-driven DL relies on accurate expert knowledge and may still suffer performance degradation if the expert knowledge does not match the actual channel conditions.

B. Motivation and Contribution

While existing semantic communication schemes have been explored for MIMO systems, they rely on traditional downlink beamforming techniques such as singular value decomposition [35], [36], which requires accurate CSI. In the context of massive MIMO systems, the constraints of limited pilot overhead inevitably lead to inaccurate CSI estimation, posing significant challenges to achieving good performance. In addition, current approaches are mainly designed for single-user narrowband MIMO systems, leaving a research gap in the application of semantic communication in the more complex multi-user massive MIMO systems. Therefore, further research is needed to develop efficient semantic communication schemes tailored for the multi-user case in near-space airship-based broadband massive MIMO communication systems. This motivates our work.

The discussion of DL-based MIMO beamforming schemes in the previous subsection shows that both data-driven and model-driven DL strategies can effectively improve beamforming performance. Each has its own advantages and limitations, making them suitable for different scenarios. This motivates us to merge the strengths of both approaches by introducing a hybrid data-driven and model-driven beamforming strategy for airship-based multi-user massive MIMO systems, with the aim of achieving superior performance. Furthermore, we integrate this beamforming strategy with semantic communication, specifically focusing on the task of image compression and reconstruction. This integration exploits the semantics of both source and channel to design the semantic coding and physical layer beamforming, culminating in a deep Joint Semantic Coding and Beamforming (JSCBF) scheme for airship-based massive MIMO image transmission network in near space.

This paper proposes a deep JSCBF approach for an airship-based massive MIMO image transmission network in near space. Our research focuses on the image modality, which was chosen for its distinctive ability to showcase the efficiency and potential of semantic transmission techniques. The inherent richness of semantic information and the superior compressibility of the image modality provide an exemplary platform to illustrate the effectiveness of semantic transmission. This research lays the foundation for future extensions to more complex modalities, such as video. The main contributions of this work are summarized below.

- To the best of our knowledge, the proposed deep JSCBF scheme is the first to integrate semantic communication with downlink multi-user broadband massive MIMO beamforming. In this scheme, the semantics derived from both images and CSI are jointly used to guide the design of semantic coding and beamforming.
- For semantic coding, we adopt the transformer architecture to extract semantic features from both images and CSI. In addition, we design a semantic fusion network to fuse the CSI semantics and image semantics to obtain the fused semantic features for the subsequent physical layer transmission. By fusing CSI semantics and image semantics, our deep JSCBF scheme can make the fused semantic features to better adapt to different channel conditions and mitigate the damage to semantic information caused by uncertainties in physical layer transmission.
- To address the unique challenges of airship-to-X links, where channel estimation overheads are often limited due to the high-speed movement of airships and the complexity of the aerial environment, we designed our scheme to handle imperfect CSI conditions effectively. Specifically, to better transmit the fused semantic features in the physical layer link with limited CSI accuracy, we propose a hybrid data-driven and model-driven semantic-aware beamforming scheme. The data-driven semantic-aware beamforming network takes the fused semantic features as input to effectively exploit the embedded image and CSI semantics and ultimately outputs transmission signals in the dimensionality of the BS antennas. To further ensure the performance of the network and improve its interpretability, we first derive a new beamforming algorithm based on WMMSE for multi-user massive MIMO systems under imperfect CSI. Then, we integrate this derived WMMSE beamforming algorithm with DL to propose a model-driven semantic-aware beamforming network. In this network, a transformer is used to replace the iterative parameter update process in WMMSE. Finally, by weighting and summing the outputs of both beamforming networks, we obtain the final transmission signal, effectively combining the advantages of both data-driven and model-driven DL for improved performance.
- By adopting a combined loss function that integrates pixel-level distortion loss, i.e., mean square error (MSE), with perceptual metrics, i.e., multi-scale structural similarity (MS-SSIM) and learned perceptual image patch similarity (LPIPS), we achieve the E2E joint training of the proposed networks. Therefore, our deep JSCBF scheme achieves the joint optimization of semantic encoding/decoding and massive MIMO beamforming, resulting in a significant improvement in image reconstruction performance.

Notation: We use lower-case letters for scalars, lower-case boldface letters for column vectors, and upper-case boldface letters for matrices. Superscripts $(\cdot)^*$, $(\cdot)^{\mathrm{T}}$, $(\cdot)^{\mathrm{H}}$, $(\cdot)^{-1}$ and $(\cdot)^{\dagger}$ denote the conjugate, transpose, conjugate transpose, inversion and Moore-Penrose inversion operators, respectively. $\|\mathbf{A}\|_F$ is the Frobenius norm of $\mathbf{A}$. $\mathrm{vec}(\mathbf{A})$ and $\mathrm{angle}(\mathbf{A})$ denote the vectorization operation and the phase values of $\mathbf{A}$, respectively. $\mathbf{I}_n$ denotes the $n \times n$ identity matrix, while $\mathbf{1}_n$ ($\mathbf{0}_n$) denotes the vector of size n with all the elements being 1 (0). $\Re\{\cdot\}$ and $\Im\{\cdot\}$ denote the real part and imaginary part of the corresponding argument, respectively. $[\mathbf{A}]_{m,n}$ denotes the mth-row and nth-column element of $\mathbf{A}$, while $[\mathbf{A}]_{[:,m:n]}$ is the sub-matrix containing the mth to nth columns of $\mathbf{A}$. The expectation is denoted by $\mathbb{E}(\cdot)$. $\frac{\partial a}{\partial b}$ is the partial derivative of a with respect to b.

II. System and Channel Model

We investigate the multi-user downlink image transmission in near-space airship-borne massive MIMO network, where the airship-borne BS transmits image data to K single-antenna user equipment (UEs). The system adopts OFDM with N_c subcarriers, and the airship-borne BS is equipped with a massive antenna array with N_t antennas. The transmitted red-green-blue three-dimensional (3D) image data of the kth UE is denoted as $\mathbf{D}[k] \in \mathbb{R}^{3\times M_x \times M_y}$, where M_x and M_y denote the width and height of the image, respectively. The airborne BS encodes the image data into a complex-valued symbol of dimension N_s $\mathbf{s}[k] \in \mathbb{C}^{N_s \times 1}$. This coding process can be expressed as

$$\mathbf{s}[k] = \mathcal{E}(\mathbf{D}[k]), \tag{1}$$

where $\mathcal{E}(\cdot)$ denotes the encoding function that maps the image data $\mathbf{D}[k]$ onto the complex-valued vector $\mathbf{s}[k]$. To mitigate inter-user interference during multi-user image data transmission, the airship-borne BS employs multi-user beamforming, converting $\mathbf{s}[k]$ for $1 \leq k \leq K$ into the 3D transmit signals $\mathbf{X} \in \mathbb{C}^{N_c \times N_t \times Q}$ spanning Q OFDM symbols, which can be expressed as

$$\mathbf{X} = \mathcal{P}\left(\mathbf{s}[1], \cdots, \mathbf{s}[K], \hat{\mathbf{H}}\right), \tag{2}$$

where $\hat{\mathbf{H}} \in \mathbb{C}^{K\times N_c \times N_t}$ is the estimate of the true 3D CSI matrix $\mathbf{H} = [[\mathbf{h}[1,1], \cdots, \mathbf{h}[1,N_c]]; \cdots; [\mathbf{h}[K,N_c], \cdots, \mathbf{h}[K,N_c]]]^{\mathrm{T}} =$

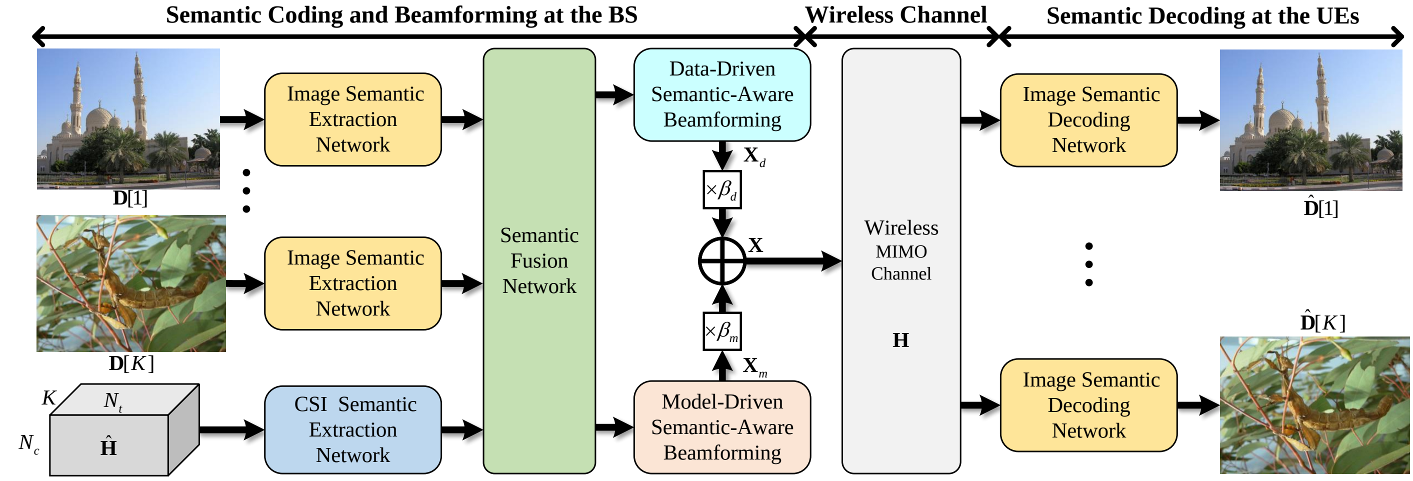

Semantic Coding and Beamforming at the BS
Wireless Channel
Semantic Decoding at the UEs
Image Semantic Extraction Network
$\mathbf{D}[1]$
Image Semantic Extraction Network
$\mathbf{D}[K]$
K
N_t
N_c
$\hat{\mathbf{H}}$
CSI Semantic Extraction Network
Semantic Fusion Network
Data-Driven Semantic-Aware Beamforming
$\mathbf{X}_d$
$\times\beta_d$
$\mathbf{X}$
$\times\beta_m$
$\mathbf{X}_m$
Model-Driven Semantic-Aware Beamforming
Wireless MIMO Channel
$\mathbf{H}$
Image Semantic Decoding Network
$\hat{\mathbf{D}}[1]$
$\hat{\mathbf{D}}[K]$
Image Semantic Decoding Network

Fig. 2. Proposed deep JSCBF scheme for near-space airship-borne massive MIMO image transmission network.

$\left[\mathbf{H}^{\mathrm{T}}[1];\cdots;\mathbf{H}^{\mathrm{T}}[K]\right]^{\mathrm{T}} \in \mathbb{C}^{K\times N_c\times N_t}$, $\mathbf{h}[k,n] \in \mathbb{C}^{N_t\times 1}$ is the channel vector from the airship BS to the kth UE on the nth subcarrier, and $\mathbf{H}[k]=\left[\mathbf{h}[k,1],\cdots,\mathbf{h}[k,N_c]\right]^{\mathrm{T}}\in\mathbb{C}^{N_c\times N_t}$. Since multi-user beamforming can be linear or nonlinear, we use $\mathcal{P}(\cdot)$ to represent the generic transformation from the encoded data and estimated CSI onto the corresponding transmit signals.

The signal received by the kth UE on the nth subcarrier of the qth OFDM symbol can be expressed as

$$y[k,q,n] = \mathbf{h}^{\mathrm{H}}[k,n]\mathbf{x}[q,n] + z[k,q,n], \tag{3}$$

where $\mathbf{x}[q,n] \in \mathbb{C}^{N_t\times 1}$ is the airship-borne BS's transmit signal on the nth subcarrier of the qth OFDM symbol, and $z[k,q,n] \sim \mathcal{CN}\left(0,\sigma_n^2\right)$ is the complex-valued additive white Gaussian noise (AWGN) with zero mean and variance σ_n^2. By aggregating the received signals across the N_c subcarriers and Q OFDM symbols, the kth UE obtains the overall received signal $\mathbf{Y}[k] \in \mathbb{C}^{N_c\times Q}$. Based on $\mathbf{Y}[k]$, the UE decodes and reconstructs the image as

$$\hat{\mathbf{D}}[k] = \mathcal{D}\left(\mathbf{Y}[k]\right), \tag{4}$$

where $\mathcal{D}(\cdot)$ denotes the decoding mapping from the received signal $\mathbf{Y}[k]$ onto the reconstructed 3D image $\hat{\mathbf{D}}[k]\in\mathbb{R}^{3\times M_x\times M_y}$.

We consider a typical multipath massive MIMO channel model. Specifically, given $L_p[k]$ as the number of multipath components between the airship-borne BS and the kth UE, the corresponding channel vector on the nth subcarrier can be described as [6]

$$\mathbf{h}[k,n] = \frac{1}{\sqrt{L_p[k]}}\sum_{l=1}^{L_p[k]} \alpha_{l,k}\mathbf{a}_t\left(\theta_{l,k},\phi_{l,k}\right)e^{-\mathrm{j}\frac{2\pi n\tau_{l,k}}{N_c T_s}}. \tag{5}$$

In (5), $\alpha_{l,k} \sim \mathcal{CN}(0,1)$ is the complex gain of the lth path, $\theta_{l,k} \in [-\pi,\ \pi]$ and $\phi_{l,k} \in [0,\ \pi/4]$ are the lth path's azimuth and zenith angles of departures between the airship-borne BS and the kth UE, respectively, while $\tau_{l,k}$ is the delay of the lth path, T_s is the OFDM sampling interval, and $\mathbf{a}_t(\cdot) \in \mathbb{C}^{N_t\times 1}$ denotes the normalized transmit array response vector.

For the airship BS equipped with a half-wavelength uniform planar array (UPA) of dimension $N_t = N_y \times N_z$, the array response vector can be expressed as

$$\begin{aligned}\mathbf{a}_t(\theta,\phi) =& \Big[1,\cdots,e^{\mathrm{j}\frac{2\pi}{\lambda}d\left(n\sin(\theta)\cos(\phi)+m\sin(\phi)\right)},\\ &\cdots,e^{\mathrm{j}\frac{2\pi}{\lambda}d\left((N_y-1)\sin(\theta)\cos(\phi)+(N_z-1)\sin(\phi)\right)}\Big]^{\mathrm{T}},\end{aligned} \tag{6}$$

where λ is the wavelength, and the adjacent antenna spacing d is given by $d = \frac{\lambda}{2}$.

III. Deep Joint Semantic Coding and Beamforming (JSCBF)

The block diagram of the proposed deep JSCBF scheme is shown in Fig. 2. In this section, we first introduce the proposed deep JSCBF problem for the airship-based massive MIMO image transmission network in near space. Second, based on the transformer architecture, we design image and CSI semantic extraction networks, respectively, and further develop a semantic fusion network to fuse the extracted source semantic and CSI semantic into complex-valued semantic features for subsequent physical layer transmission. Third, we propose data-driven and model-driven semantic-aware beamforming networks to map these fused semantic features onto transmission signals. Finally, we perform joint training of all proposed networks to achieve high-quality image compression and reconstruction.

A. Problem Formulation

To achieve efficient and accurate image transmission over an airborne massive MIMO communication network with limited wireless resources, it is crucial to jointly design the image coding and beamforming at the transmitter and the image decoding at the receiver by minimizing the semantic loss between the original image and the reconstructed image. This optimization problem can be formulated as

$$\begin{aligned}\min_{\mathcal{E}(\cdot),\mathcal{P}(\cdot),\mathcal{D}(\cdot)}\quad & \sum_{k=1}^{K}\mathcal{L}\big(\hat{\mathbf{D}}[k],\mathbf{D}[k]\big),\\ \text{s.t.}\quad & \mathbf{s}[k] = \mathcal{E}(\mathbf{D}[k]),\ \forall k,\\ & \mathbf{X} = \mathcal{P}\left(\mathbf{s}[1],\cdots,\mathbf{s}[K],\hat{\mathbf{H}}\right),\\ & \hat{\mathbf{D}}[k] = \mathcal{D}(\mathbf{Y}[k]),\ \forall k,\\ & \|\mathbf{X}\|_F^2) \leq QP_t,\end{aligned} \tag{7}$$

where the transmit signals $\mathbf{X}$ should satisfy the power constraint, and $\mathcal{L}\big(\hat{\mathbf{D}}[k],\mathbf{D}[k]\big)$ is the loss function to measure the semantic similarity between the reconstructed and original images. Various metrics such as Peak Signal to Noise Ratio (PSNR), MS-SSIM, or LPIPS can be used to quantify the semantic similarity between $\hat{\mathbf{D}}[k]$ and $\mathbf{D}[k]$, which will be discussed in detail in the following subsections. The use of semantic loss as the optimization goal in our problem aims to capture the meaningful information conveyed by the transmitted signals. By minimizing semantic loss, we ensure that essential information is preserved during transmission, which directly enhances transmission efficiency. The goal is to maximize the semantic content received by the users while using minimal resources, thus improving overall transmission efficiency.

Transmission capacity is traditionally defined as the maximum data transfer rate achievable under given channel conditions. We

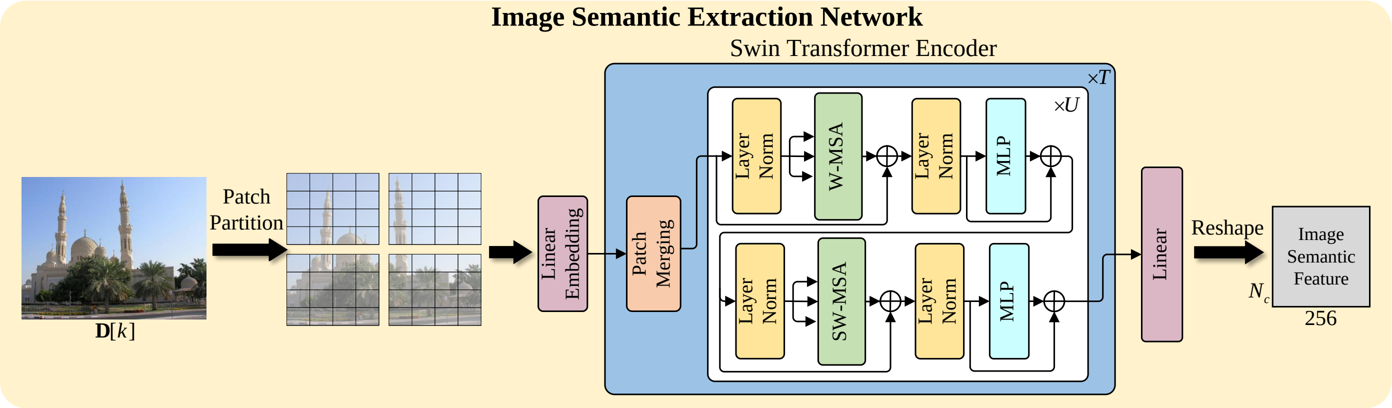


Fig. 3. Image semantic extraction network, where MLP stands for multilayer perceptron, and $\times U$ ($\times T$) means the module repeats U times (T times).

consider the semantic similarity metric as an indicator of the effective capacity of the communication link. Higher semantic similarity means closer utilization of the channel's capacity for meaningful information transfer. This interpretation bridges the gap between semantic communication and traditional capacity metrics, suggesting that optimizing semantic similarity can lead to optimal utilization of the channel's capacity.

The problem (7) is a complicated joint optimization. Conventional methods typically focus on optimizing each module separately using different metrics, thereby i) failing to achieve joint optimization based on the evaluation metric of semantic reconstruction, and ii) failing to jointly exploit the embedded semantic information in both source and CSI for designing communication systems. As a result, conventional methods cannot achieve optimal performance, especially under imperfect CSI. How to effectively utilize the semantic information inherent in both source and CSI to facilitate the joint design of all modules and construct an E2E massive MIMO semantic communication system for improved performance is the core problem to be solved in this paper.

B. Transformer-based Semantic Extraction and Fusion of Image and CSI semantics

1) Image semantic extraction: As shown in Fig. 3, this paper introduces a Swin Transformer-based image semantic extraction network designed for efficient image semantic extraction. The Swin Transformer has unique advantages, such as hierarchical structure processing and shifted window-based self-attention [56]. These features facilitate detailed contextual analysis of images, making Swin Transformer an ideal network backbone for the task of image semantic extraction.

To process $3 \times 256 \times 256$ images with Swin Transformer, the images are first segmented into non-overlapping patches of dimension $N \times N$ by patch partitioning, which converts their dimension to $3N^2 \times \frac{256}{N} \times \frac{256}{N}$, where N is the patch size. The images are further divided into $\frac{256^2}{N^2 M^2}$ non-overlapping windows of dimension $M \times M$, where M is the window size. A linear embedding layer is applied to the raw patches in each window to create a $\mathbf{F}_p \in \mathbb{R}^{C\times M\times M}$ feature map with C channels. The feature map $\mathbf{F}_p$ is transformed into a sequence $\mathbf{F}_s \in \mathbb{R}^{M^2\times C}$, where M^2 is the number of patches in each window. Next, the Swin Transformer encoder applies window-based multi-head self-attention (W-MSA) layers within each window to increase focus on local features. This attention mechanism is applied within the windows of these patches and is critical for capturing intricate details, preserving feature locality while reducing computational complexity compared to global attention mechanisms. To capture a broader context, the Swin Transformer encoder uses shifted window-based multi-head self-attention (SW-MSA) layers. This allows interaction between adjacent windows, facilitating more comprehensive global semantic feature extraction from the image. As the image propagates through the Swin Transformer encoder layers, patch merging is applied, scaling the channel dimensions while reducing the spatial dimensions. This merging process is essential for constructing a hierarchical representation, starting with fine-grained details and progressing to more abstract features. Through patch merging, the Swin Transformer encoder ensures comprehensive and nuanced processing of visual data, deftly balancing local feature extraction with global contextual understanding. Finally, the extracted features are passed through a fully connected linear layer, resulting in the semantic feature matrix of the image with a dimension of $N_c \times 256$.

2) CSI semantic extraction: Based on the estimated CSI $\hat{\mathbf{H}}$, the airship BS extracts CSI semantic information for subsequent semantic fusion and physical layer beamforming design. This is achieved by the proposed transformer-based CSI semantic extraction network shown in Fig. 4. Unlike the convolution operation in CNNs [57], which is limited to feature extraction from local regions, the self-attention mechanism inherent in the transformer structure excels at global feature extraction [58], [59]. This capability allows it to detect the inter-subcarrier correlation within the input signal on a global scale and assign appropriate weighting coefficients to the components within each

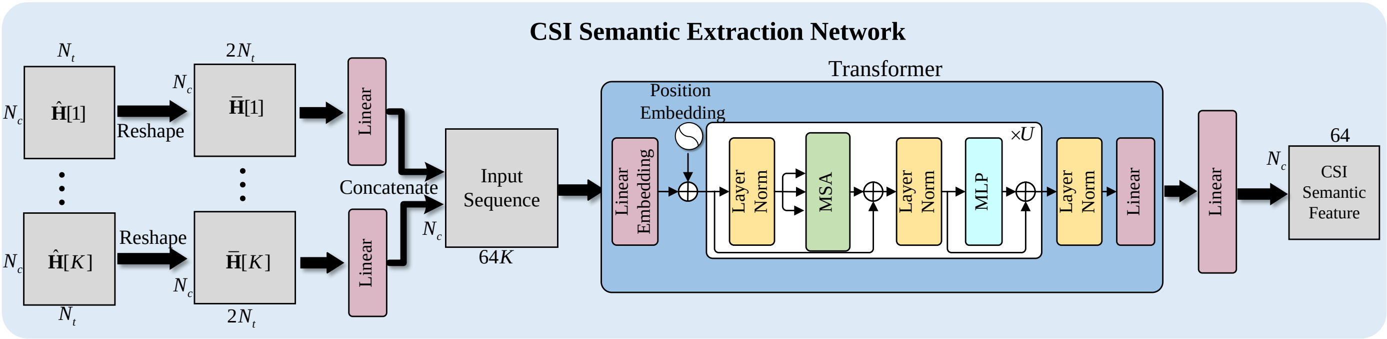


Fig. 4. CSI semantic extraction network.

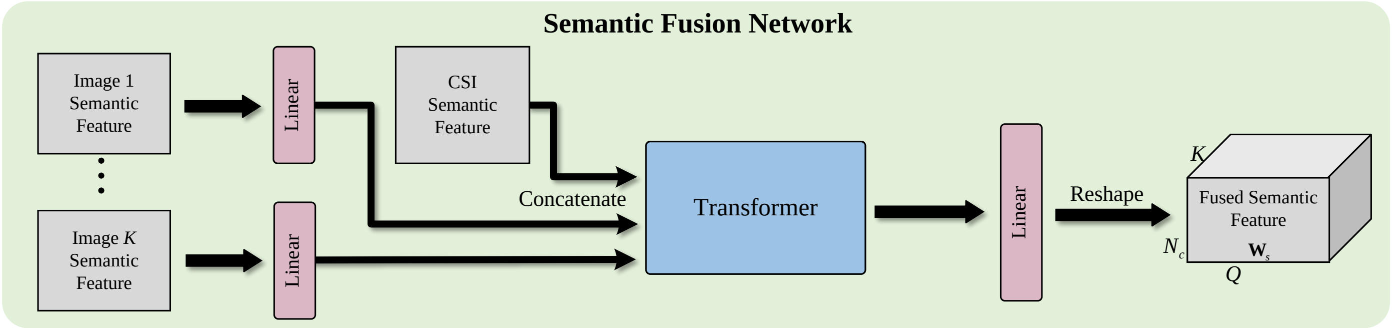


Fig. 5. Semantic fusion network.

subcarrier, thereby improving overall performance. A typical transformer takes a one-dimensional (1D) real-valued sequence as input and produces an output in a similar format [58], [59]. To process the complex-valued input $\hat{\mathbf{H}}[k] \in \mathbf{R}^{N_c \times N_t}$, where $\hat{\mathbf{H}}[k]$ is the estimate of $\mathbf{H}[k]$, the first step is to convert it into a real-valued matrix $\bar{\mathbf{H}}[k] \in \mathbb{R}^{N_c \times 2N_t}$:

$$\begin{cases} \left[\bar{\mathbf{H}}[k]\right]_{[:,1:N_t]} = \Re\left\{\hat{\mathbf{H}}[k]\right\}, \\ \left[\bar{\mathbf{H}}[k]\right]_{[:,1+N_t:2N_t]} = \Im\left\{\hat{\mathbf{H}}[k]\right\}. \end{cases} \tag{8}$$

As shown in Fig. 4, the real-valued CSI matrices corresponding to all K UEs are first dimensionally compressed to $N_c \times 64$ by a fully connected linear layer, and then concatenated into a 1D real-valued sequence of dimension $N_c \times 64K$. This sequence serves as the input to the transformer, where the effective input sequence length is determined by the number of subcarriers N_c. Inside the transformer, the input sequence is first transformed into a vector sequence of dimension d_{model} via a fully connected linear embedding layer followed by a position embedding layer. Different frequencies of sine functions are used to denote the positions of different subcarriers, and the position information of the subcarriers is embedded by summing with the vector sequence. The transformer then employs U identical layers to extract semantic features from the input sequence, where each layer consists of a multi-head self-attention (MSA) sublayer and an MLP sublayer. The extracted features are then processed through a fully connected linear layer to obtain a CSI semantic feature matrix of dimension $N_c \times 64$.

3) Semantic fusion of image and CSI semantics: After extracting semantic features from both images and CSI, the obtained features remain in the hidden space of the neural network and cannot be directly used for physical layer transmission. To this end, we employ a semantic fusion network as shown in Fig. 5 to fuse the image and CSI semantic features from multiple UEs for semantic coding to form a semantic data stream tailored for physical layer transmission. In contrast, traditional physical layer processing schemes adopt a separate module design approach, i.e., the source coding typically considers only the distribution of the source itself, neglecting the influence of CSI. In particular, small-scale channel fading would introduce significant variations in channel conditions between different subcarriers and spatial subchannels. This can lead to degraded performance of conventional coding schemes.

The proposed approach addresses this issue by incorporating additional semantic information from the CSI at the semantic coding stage, thereby facilitating the semantic coding module to adapt to the MIMO physical layer transmission link state. This integration facilitates E2E joint optimization with the subsequent physical layer transmission modules, thereby improving the system's adaptability to channel variations. Specifically, we first process the image semantic feature of each UE through a linear layer and concatenate them with the CSI semantic feature. Subsequently, we the concatenated data stream using a transformer to obtain a semantic matrix of dimension $N_c \times 2QK$, denoted as $\tilde{\mathbf{W}}_{\text{s}}$. Then, we decompose $\tilde{\mathbf{W}}_{\text{s}}$ to form a complex-valued matrix $\mathbf{W}_{\text{s}}$ according to

$$\begin{cases} \Re\left\{\mathbf{W}_{\text{s}}\right\} = \left[\tilde{\mathbf{W}}_{\text{s}}\right]_{[:,1:QK]}, \\ \Im\left\{\mathbf{W}_{\text{s}}\right\} = \left[\tilde{\mathbf{W}}_{\text{s}}\right]_{[:,1+QK:2QK]}. \end{cases} \tag{9}$$

Finally, we reshape $\mathbf{W}_{\text{s}}$ into the dimension of $N_c \times K \times Q$, namely, $\mathbf{W}_{\text{s}} = \left[\mathbf{W}_{\text{s}}[1]; \cdots; \mathbf{W}_{\text{s}}[N_c]\right]$ with $\mathbf{W}_{\text{s}}[n] \in \mathbb{C}^{K \times Q}$, for subsequent physical-layer transmission.

This proposed semantic encoding scheme leverages semantic information from both images and CSI, enabling adaptive optimization of semantic transmission based on channel conditions.

C. Semantic-Aware Beamforming

After the semantic fusion is completed, the semantic features are precoded in the spatial domain and transmitted to the target UE over the MIMO channel. Since our ultimate goal is to ensure the similarity of the original and reconstructed images at the semantic level, we need to optimize the beamforming process based on the semantic information. To this end, we model the beamforming module as a DNN with semantic features as inputs, which is jointly trained E2E with the other modules using the final semantic metrics. Specifically, we propose a beamforming module as a data-driven semantic-aware beamforming network, as shown in Fig. 6, and a model-driven semantic-aware beamforming network, as shown in Fig. 7. The outputs of these two networks are then summed to produce the final transmit signal.

The data-driven approach leverages extensive training data to train deep learning models, capturing the complex characteristics of the channels and images. This method excels in automatically learning the nonlinear features of the channel and adapting to various channel conditions. The data-driven approach relies on

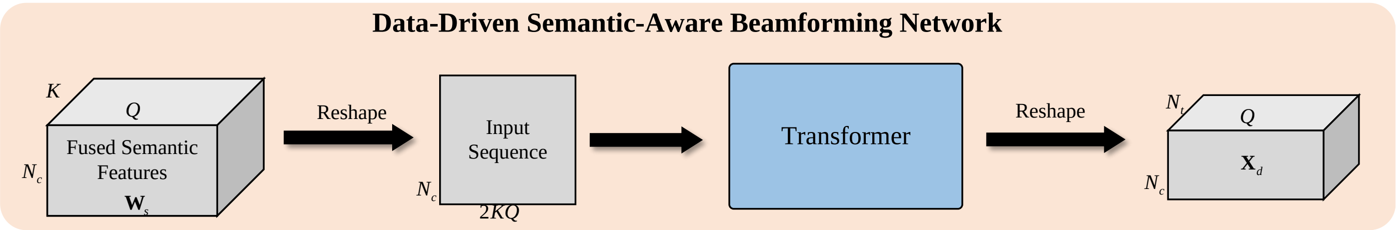


Fig. 6. Data-driven semantic-aware beamforming network.

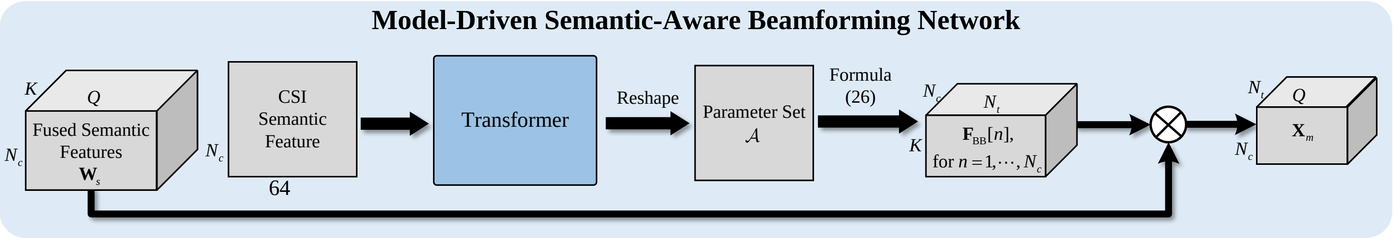

Model-Driven Semantic-Aware Beamforming Network
K
Q
Fused Semantic Features $\mathbf{W}_s$
N_c
CSI Semantic Feature
N_c
64
Transformer
Reshape
Parameter Set $\mathcal{A}$
Formula (26)
N_c
N_t
K
$\mathbf{F}_{\text{BB}}[n]$, for $n=1,\cdots,N_c$
N_t
Q
N_c
$\mathbf{X}_m$

Fig. 7. Model-driven semantic-aware beamforming network.

powerful computational resources and large amounts of training data, performing best when the training data matches the actual channel environment.

The model-driven approach combines physical models with neural networks. Specifically, the physical model provides prior knowledge and structured information, while the neural network handles complex nonlinear parts. This approach maintains efficiency and reliability while better adapting to dynamically changing channel environments. This approach performs best when the physical model matches the actual channel environment.

Our proposed dual-driven beamforming network combines data-driven and model-driven approaches, leveraging the strengths of both to enhance system robustness and performance. The data-driven method captures complex channel characteristics through deep learning models, while the model-driven method uses prior knowledge from physical models to ensure reliable and efficient beamforming.

1) Data-driven semantic-aware beamforming: As shown in Fig. 6, our data-driven semantic-aware beamforming network maps the 3D fused semantic feature matrix $\mathbf{W}_{\text{s}} \in \mathbb{C}^{N_c \times K \times Q}$ onto the 3D transmit signal matrix $\mathbf{X}_d \in \mathbb{C}^{N_c \times N_t \times Q}$. Specifically, the complex-valued semantic feature matrix $\mathbf{W}_{\text{s}}$ is first transformed into a 1D real-valued input sequence with dimension $N_c \times 2QK$, then processed by a transformer for further refinement, and finally transformed into the transmit signals $\mathbf{X}_d$ by a fully connected linear layer and subsequent transform operation.

Current massive MIMO beamforming primarily uses linear beamforming, and different linear beamforming strategies are used for different scenarios. For example, space division multiple access beamforming schemes perform well in the case of sufficiently accurate CSI and small channel correlation between UEs, while non-orthogonal multiple access and rate-split multiple access beamforming techniques show superior performance in the case of imperfect CSI or severe channel interference between UEs [54]. In contrast, we implement beamforming through the nonlinear mapping provided by DNN, and exploit the powerful representational capability of DNN to provide the proposed beamforming scheme with enhanced adaptability.

2) Model-driven semantic-aware beamforming: We initially derive a novel WMMSE beamforming method tailored for multi-user MIMO scenarios under imperfect CSI. Subsequently, we incorporate DL techniques into training to enhance performance, culminating in the establishment of a model-driven semantic-aware beamforming network.

2.1) Considering the utilization of linear beamforming based on WMMSE with sum rate as the optimization objective, the optimization problem can be articulated as

$$\begin{aligned} \max_{\mathbf{F}_{\text{BB}}[n], \forall n} \quad & R = \frac{1}{N_c} \sum_{k=1}^{K} \sum_{n=1}^{N_c} \\ & \log_2 \left(1 + \frac{\left| \mathbf{h}^{\text{H}}[k,n] \mathbf{f}_{\text{BB}}[k,n] \right|^2}{\sum\limits_{m \neq k} \left| \mathbf{h}^{\text{H}}[k,n] \mathbf{f}_{\text{BB}}[m,n] \right|^2 + \sigma_n^2} \right), \\ \text{s.t.} \quad & \sum_{n=1}^{N_c} \left\| \mathbf{F}_{\text{BB}}[n] \right\|_F^2 \leq P_t, \end{aligned} \tag{10}$$

where $\mathbf{f}_{\text{BB}}[k,n] \in \mathbb{C}^{N_t \times 1}$ represents the beamforming vector for the kth UE on the nth subcarrier, $\mathbf{F}_{\text{BB}}[n] = \left[\mathbf{f}_{\text{BB}}[1,n], \cdots, \mathbf{f}_{\text{BB}}[K,n] \right] \in \mathbb{C}^{N_t \times K}$, and the beamforming matrices $\mathbf{F}_{\text{BB}}[n]$, $\forall n$, are required to satisfy a power constraint with a maximum power of P_t.

In the optimization problem (10), the power allocation in the frequency domain and the design of beamforming in the spatial domain are intricately coupled, rendering the resolution of this optimization problem a challenging task. Consequently, we employ the concept of alternating optimization, decoupling the power allocation in the frequency domain and the beamforming in the spatial domain, and optimizing them in an alternating fashion. Specifically, reformulating (10) as

$$\begin{aligned} \max_{\mathbf{F}_{\text{BB}}[n], p[n], \forall n} \quad & R = \frac{1}{N_c} \sum_{k=1}^{K} \sum_{n=1}^{N_c} \\ & \log_2 \left(1 + \frac{\left| \mathbf{h}^{\text{H}}[k,n] \mathbf{f}_{\text{BB}}[k,n] \right|^2}{\sum\limits_{m \neq k} \left| \mathbf{h}^{\text{H}}[k,n] \mathbf{f}_{\text{BB}}[m,n] \right|^2 + \sigma_n^2} \right), \\ \text{s.t.} \quad & \left\| \mathbf{F}_{\text{BB}}[n] \right\|_F^2 \leq p[n], \ \forall n, \\ & \sum_{n=1}^{N_c} p[n] \leq P_t, \end{aligned} \tag{11}$$

where $p[n]$ denotes the power allocation coefficient on the nth subcarrier. Given the fixed beamforming matrices $\mathbf{F}_{\text{BB}}[n], \forall n$, we can optimize the power allocation coefficients $p[n], \forall n$ in the frequency domain, utilizing the water-filling algorithm. Given the power allocation coefficients, the optimization of the beamforming matrices can be decoupled on per subcarrier base as

$$\begin{aligned} \max_{\mathbf{F}_{\text{BB}}[n]} \quad & R[n] = \sum_{k=1}^{K} \\ & \log_2 \left(1 + \frac{\left| \mathbf{h}^{\text{H}}[k,n] \mathbf{f}_{\text{BB}}[k,n] \right|^2}{\sum\limits_{m \neq k} \left| \mathbf{h}^{\text{H}}[k,n] \mathbf{f}_{\text{BB}}[m,n] \right|^2 + \sigma_n^2} \right), \\ \text{s.t.} \quad & \left\| \mathbf{F}_{\text{BB}}[n] \right\|_F^2 \leq p[n]. \end{aligned} \tag{12}$$

That is, given $p[n], \forall n$, the beamforming matrices for different subcarriers can be designed independently.

Consider the relationship between the true CSI $\mathbf{h}[k,n]$ and the estimated CSI $\hat{\mathbf{h}}[k,n]$ as

$$\mathbf{h}[k,n] = \hat{\mathbf{h}}[k,n] + \Delta \mathbf{h}[k,n], \tag{13}$$

where $\Delta \mathbf{h}[k,n] \in \mathbb{C}^{N_t \times 1}$ represents the CSI error, which follows the complex Gaussian distribution with the auto-correlation matrix $\mathbf{R}_e \in \mathbb{C}^{N_t \times N_t}$. Denote

$$\hat{r}[k,n] = e[k,n] \big(\mathbf{h}^{\text{H}}[k,n] \mathbf{F}_{\text{BB}}[n] \mathbf{r}[n] + z[k,n] \big) \tag{14}$$

as the estimate of the transmit data $r[k,n]$ to the kth UE on the nth subcarrier, where $\mathbf{r}[n] = \left[r[1,n], \cdots, r[K,n] \right]^{\text{T}} \in \mathbf{C}^{K \times 1}$ is the transmit data vector on the nth subcarrier, $e[k,n]$ is the equalizer of the kth UE on the nth subcarrier, and $z[k,n] \sim \mathcal{CN}(0, \sigma_n^2)$ is the complex-valued AWGN. Then the MSE of decoding the data for the kth UE can be approximately expressed as

$$\varepsilon[k,n] = \mathbb{E} \left\{ \left| \hat{r}[k,n] - r[k,n] \right|^2 \right\} \approx \varepsilon^{(1)}[k,n] + \varepsilon^{(2)}[k,n], \tag{15}$$

where

$$\varepsilon^{(1)}[k,n] = |e[k,n]|^2 T[k,n] - 2\Re\left\{e[k,n]\hat{\mathbf{h}}^{\mathrm{H}}[k,n]\mathbf{f}_{\mathrm{BB}}[k,n]\right\} + 1 \tag{16}$$

represents the MSE under the assumption that the CSI error is negligible, and

$$\varepsilon^{(2)}[k,n] = |e[k,n]|^2 \sum_{m\neq k} \mathbf{f}_{\mathrm{BB}}^{\mathrm{H}}[m,n]\mathbf{R}_e\mathbf{f}_{\mathrm{BB}}[m,n] \tag{17}$$

represents the MSE introduced by the inter-user interference due to the CSI error, while

$$T[k,n] = \sum_{m=1}^{K} \left|\hat{\mathbf{h}}^{\mathrm{H}}[k,n]\mathbf{f}_{\mathrm{BB}}[m,n]\right|^2 + \sigma_n^2 \tag{18}$$

denotes the average received power.

We now elaborate the formula (15). Since the airship BS lacks perfect CSI, the MSE at the beamforming stage is a random variable to the airship BS that does not facilitate the calculation of the actual achievable rates. Therefore, we propose to optimize using an approximate MSE in lieu of the actual MSE. We assume the availability of equivalent CSI at the UE side post-linear beamforming, enabling the realization of the rates derived below. Additionally, we presume that the MSE is exclusively attributed to inter-user interference and noise. For the kth UE, the expected received signal is $e[k,n]\big(\Delta\mathbf{h}[k,n] + \hat{\mathbf{h}}[k,n]\big)^{\mathrm{H}}\mathbf{f}_{\mathrm{BB}}[k,n]r[k,n]$, while the inter-user interference and noise are $\sum\limits_{m\neq k} e[k,n]\big(\Delta\mathbf{h}[k,n] + \hat{\mathbf{h}}[k,n]\big)^{\mathrm{H}}\mathbf{f}_{\mathrm{BB}}[m,n]r[m,n] + e[k,n]z[k,n]$. Based on the above reasoning, the MSE can be represented by (15). It is important to note that we have made a simplistic assumption that the CSI error $\Delta\mathbf{h}[k,n]$ follows a complex Gaussian distribution and is independent of the true CSI $\mathbf{h}[k,n]$, hence the use of the approximation symbol in (15).

By setting $\frac{\partial\varepsilon[k,n]}{\partial e[k,n]} = 0$, the minimum mean square error (MMSE) equalizer can be obtained as:

$$e^{\mathrm{MMSE}}[k,n] = \frac{\mathbf{f}_{\mathrm{BB}}^{\mathrm{H}}[k,n]\hat{\mathbf{h}}[k,n]}{\left(T[k,n] + \sum\limits_{m\neq k}^{K} \mathbf{f}_{\mathrm{BB}}^{\mathrm{H}}[m,n]\mathbf{R}_e\mathbf{f}_{\mathrm{BB}}[m,n]\right)}. \tag{19}$$

Substituting (19) into (15), the MMSE can be obtained as

$$\varepsilon^{\mathrm{MMSE}}[k,n] = 1 - \frac{\left|\mathbf{f}_{\mathrm{BB}}^{\mathrm{H}}[k,n]\hat{\mathbf{h}}[k,n]\right|^2}{\left(T[k,n] + \sum\limits_{m\neq k} \mathbf{f}_{\mathrm{BB}}^{\mathrm{H}}[m,n]\mathbf{R}_e\mathbf{f}_{\mathrm{BB}}[m,n]\right)}. \tag{20}$$

Then, the signal-to-interference-plus-noise ratio (SINR) for the kth UE on the nth subcarrier can be expressed as

$$\gamma[k,n] = \frac{1}{\varepsilon^{\mathrm{MMSE}}[k,n]} - 1, \tag{21}$$

with the corresponding achievable rate given by

$$\hat{R}[k,n] = -\log_2\left(\varepsilon^{\mathrm{MMSE}}[k,n]\right). \tag{22}$$

Since the logarithmic rate-MSE relationship cannot be directly used for solving rate optimization problems, we introduce the augmented weighted MSE (WMSE)

$$\xi[k,n] = \lambda[k,n]\varepsilon[k,n] - \log_2(\lambda[k,n]), \tag{23}$$

where $\lambda[k,n]$ represents the weight of MSE for the kth UE on the nth subcarrier. By setting $\frac{\partial\xi[k,n]}{\partial\lambda[k,n]} = 0$, the optimal weight can be obtained as

$$\lambda^{\mathrm{MMSE}}[k,n] = \frac{1}{\varepsilon^{\mathrm{MMSE}}[k,n]}. \tag{24}$$

Substituting (19) and (24) into (23), the relationship for the rate-WMMSE can be established as

$$\begin{aligned} \min_{\mathbf{F}_{\mathrm{BB}}[n],\mathbf{e}[n],\boldsymbol{\lambda}[n]} \quad & \xi[n] = \sum_{k=1}^{K} \xi[k,n], \\ \text{s.t.} \quad & \|\mathbf{F}_{\mathrm{BB}}[n]\|_F^2 \leq p[n], \end{aligned} \tag{25}$$

where $\mathbf{e}[n] = [e[1,n],\cdots,e[K,n]]^{\mathrm{T}} \in \mathbb{C}^{K\times 1}$ and $\boldsymbol{\lambda}[n] = [\lambda[1,n],\cdots,\lambda[K,n]]^{\mathrm{T}} \in \mathbb{R}^{K\times 1}$. By fixing $\mathbf{e}[n]$ and $\boldsymbol{\lambda}[n]$ and applying Lemma 1 of [51] to eliminate the power constraints, we obtain the closed-form solution for $\mathbf{F}_{\mathrm{BB}}[n]$ from $\frac{\partial\xi[n]}{\partial\mathbf{f}_{\mathrm{BB}}[k,n]} = \mathbf{0}_{N_t}$, i.e.,

$$\begin{aligned} \mathbf{f}_{\mathrm{BB}}[k,n] = \Bigg(& \frac{\sigma_n^2}{p[n]} \sum_{m=1}^{K} \lambda[m,n]\,|e[m,n]|^2\,\mathbf{I}_{N_t} \\ & + \sum_{m\neq k} \lambda[m,n]\,|e[m,n]|^2\,\mathbf{R}_e \\ & + \sum_{m=1}^{K} \lambda[m,n]\,|e[m,n]|^2\,\mathbf{h}[m,n]\mathbf{h}^{\mathrm{H}}[m,n]\Bigg)^{-1} \\ & \times e^*[k,n]\lambda[k,n]\mathbf{h}[k,n]. \end{aligned} \tag{26}$$

The beamforming matrices $\mathbf{F}_{\mathrm{BB}}[n], \forall n$, depend on the parameter set $\mathcal{A} = \big\{p[n], \lambda[k,n], e[k,n], \mathbf{R}_e; \text{for } k = 1,\cdots,K; n = 1,\cdots,N_c\big\}$. Hence, the beamforming process can be decomposed into the block coordinate descent iterative optimization of the power allocation coefficient $p[n]$, weighting factor $\lambda[k,n]$, equalizer coefficient $e[k,n]$ and beamformer $\mathbf{f}_{\mathrm{BB}}[k,n]$, as summarized in Algorithm 1, where I_1 is the number of iterations.

2.2) However, the above WMMSE algorithm is based on the assumption that the CSI error follow a known complex Gaussian distribution and are independent of the CSI, which may not be true in the actual scenario. To this end, we propose a model-driven semantic-aware beamforming network by deep unfolding the proposed WMMSE algorithm, as depicted in Fig. 7. This network is capable of perceiving the semantic information from the input and adaptively adjust the critical parameter set $\mathcal{A}$

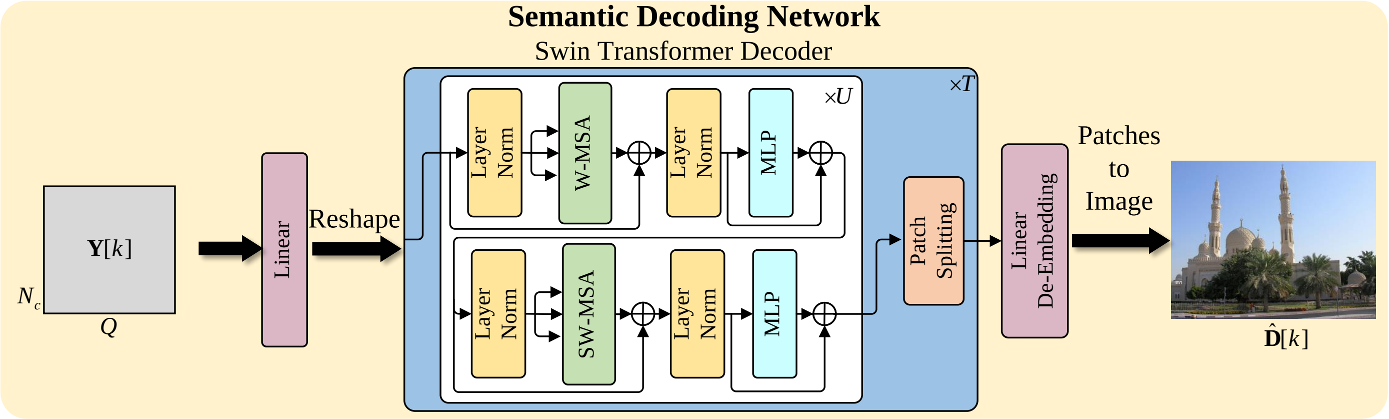


Fig. 8. Semantic decoding network.

through DL training for improved performance. Specifically, this network takes the CSI semantic feature as input. A transformer is employed to process the CSI semantic feature and generate the critical parameter set $\mathcal{A}$ as outlined in the proposed WMMSE algorithm. The closed-form solution (26) is then utilized to obtain the beamformer based on the critical parameter set $\mathcal{A}$.

Based on the obtained beamformer, we perform linear beamforming on the fused semantic features to obtain the 3D transmit signal matrix $\mathbf{X}_m = \left[\mathbf{X}_m[1];\cdots;\mathbf{X}_m[N_c]\right] \in \mathcal{C}^{N_c\times N_t\times Q}$ with

$$\mathbf{X}_m[n] = \mathbf{F}_{\mathrm{BB}}[n]\mathbf{W}_{\mathrm{s}}[n] \in \mathbb{C}^{N_t\times Q}, \text{ for } n = 1,\cdots,N_c. \quad (27)$$

2.3) Next, we provide a qualitative theoretical analysis of the impact of CSI errors on transmission capacity. Based on Equation (20), and considering that MSE mainly arises from user interference and noise due to CSI errors, we can approximate the post-beamforming MSE as:

$$\varepsilon^{\mathrm{MSE}}[k,n] \approx 1- \frac{\left|\mathbf{f}_{\mathrm{BB}}^{H}[k,n]\widehat{\mathbf{h}}[k,n]\right|^2}{\left(\left|\mathbf{f}_{\mathrm{BB}}^{H}[k,n]\widehat{\mathbf{h}}[k,n]\right|^2 + \sum\limits_{m\neq k}\mathbf{f}_{\mathrm{BB}}^{H}[m,n]\mathbf{R}_e\mathbf{f}_{\mathrm{BB}}[m,n]\right)}. \quad (28)$$

Assuming the CSI error follows an independent and identically distributed (i.i.d.) complex Gaussian distribution with variance σ_H^2, this can be further simplified to:

$$\varepsilon^{\mathrm{MSE}}[k,n] \approx 1- \frac{\left|\mathbf{f}_{\mathrm{BB}}^{H}[k,n]\widehat{\mathbf{h}}[k,n]\right|^2}{\left(\left|\mathbf{f}_{\mathrm{BB}}^{H}[k,n]\widehat{\mathbf{h}}[k,n]\right|^2 + \sum\limits_{m\neq k}\mathbf{f}_{\mathrm{BB}}^{H}[m,n]\mathbf{f}_{\mathrm{BB}}[m,n]\sigma_H^2\right)}. \quad (29)$$

Thus, we can derive an approximate relationship between the channel estimation error energy σ_H^2 and the achievable rate:

$$R[k,n] \approx -\log_2 \left(1 - \frac{\left|\mathbf{f}_{\mathrm{BB}}^{H}[k,n]\widehat{\mathbf{h}}[k,n]\right|^2}{\left(\left|\mathbf{f}_{\mathrm{BB}}^{H}[k,n]\widehat{\mathbf{h}}[k,n]\right|^2 + \sum\limits_{m\neq k}\mathbf{f}_{\mathrm{BB}}^{H}[m,n]\mathbf{f}_{\mathrm{BB}}[m,n]\sigma_H^2\right)}\right). \quad (30)$$

2.4) To leverage the advantages of both data-driven DL and model-driven DL, we adopt a weighted summation method to accomplish the merging of two types of beamforming, yielding the 3D transmit signals

$$\mathbf{X}_c = \beta_d\mathbf{X}_d + \beta_m\mathbf{X}_m, \quad (31)$$

where β_d and β_m are the learnable weighting coefficients for weighting the two beamformers, respectively. Through the E2E DL training, the network can strike a balance between data-driven DL and model-driven DL for achieving the optimal beamforming performance.

To ensure that the average power does not exceed P_t, power constraints are applied to yield the final 3D transmit signals as

$$\mathbf{X} = \min\left\{\sqrt{QP_t}, \|\mathbf{X}_c\|_F\right\}\frac{\mathbf{X}_c}{\|\mathbf{X}_c\|_F}. \quad (32)$$

3) Semantic Decoding: At the receiver of each UE, a semantic decoding network based on the Swin Transformer is used to reconstruct images from the received signals, as shown in Fig. 8. Specifically, each UE first processes the received signal through a fully connected layer and transforms it into a semantic feature map of dimension $N_c\times 16\times 16$. Then, the UE uses a Swin Transformer decoder that is symmetric with the Swin Transformer encoder to process the feature map. Unlike patch merging in the Swin Transformer encoder, the Swin Transformer decoder uses patch splitting to upsample the feature map resolution and reduce the number of feature map channels. A final linear deembedding layer is used to map the feature map onto the reconstructed image.

```
Algorithm 1: WMMSE-Based Beamforming with Imperfect CSI
Initialize Set beamformer $\mathbf{F}_{\mathrm{BB}}$ to zero forcing beamformer;
for $i = 1$ to $I_1$ do
    Update $p[n]$ using water-filling algorithm with fixed $\mathbf{F}_{\mathrm{BB}}[n]$, $\mathbf{e}[n]$ and $\boldsymbol{\lambda}[n]$, for $n = 1,\cdots,N_c$;
    Update $\mathbf{e}[n]$ and $\boldsymbol{\lambda}[n]$ using (19) and (24) with fixed $\mathbf{F}_{\mathrm{BB}}[n]$ and $p[n]$, for $n = 1,\cdots,N_c$;
    Update $\mathbf{F}_{\mathrm{BB}}[n]$ using (19) with fixed $\mathbf{e}[n]$, $\boldsymbol{\lambda}[n]$ and $p[n]$, for $n = 1,\cdots,N_c$;
end for
```

4) Loss Function: To enhance the precision of image reconstruction and improve the visual quality, we propose a composite loss function that combines the MSE, MS-SSIM [60], and LPIPS [61] based on pre-trained VGG-16 [62], denoted as $\mathcal{L}_{\mathrm{MSE}}(\cdot)$, $\mathcal{L}_{\mathrm{MS-SSIM}}(\cdot)$, and $\mathcal{L}_{\mathrm{VGG}}(\cdot)$, respectively.

The MSE loss is utilized to minimize the average squared difference between the estimated values and the ground truth, ensuring pixel-level accuracy.

By contrast, the MS-SSIM loss is employed to preserve the structural information across multiple scales, thereby preserving the perceptually relevant parts of the image [60].

Although MSE and MS-SSIM are the most widely used metrics for image similarity measurement, they are simplistic functions that fail to account for many nuances of human perception [61]. To better achieve semantic communication, we further adopt the emerging DL-based LPIPS metric [61] as a perceptual loss to quantify image transmission performance. This metric is capable of emulating the human perceptual evaluation process, thus providing an LPIPS loss score. The LPIPS is computed using features extracted from a neural network, typically a pre-trained VGG network. The LPIPS value ranges from 0 to 1, with a value closer to 0 indicating less distortion. More specifically, the LPIPS metric can be expressed as

$$\mathcal{L}_{\mathrm{LPIPS}}\big(\hat{\mathbf{D}},\mathbf{D}\big) = \sum_j w_j\left\|\phi_j\big(\hat{\mathbf{D}}\big) - \phi_j\big(\mathbf{D}\big)\right\|_2^2, \quad (33)$$

where $\phi_j\big(\hat{\mathbf{D}}\big)$ and $\phi_j\big(\mathbf{D}\big)$ denote the feature maps extracted from the generated image $\hat{\mathbf{D}}$ and the target image $\mathbf{D}$ by the jth layer of the neural network, respectively, and w_j represents the learned weight for the j-th layer, emphasizing its importance in the perceptual similarity measure. The selection of layers and the calculation of weights are typically based on empirical evaluations consistent with studies of human perception.

The overall loss function is a weighted sum of these three components:

$$\begin{aligned}\mathcal{L}_{\mathrm{total}}\big(\hat{\mathbf{D}},\mathbf{D}\big) =& \lambda_{\mathrm{MSE}}\mathcal{L}_{\mathrm{MSE}}\big(\hat{\mathbf{D}},\mathbf{D}\big)\\ &+ \lambda_{\mathrm{MS-SSIM}}\mathcal{L}_{\mathrm{MS-SSIM}}\big(\hat{\mathbf{D}},\mathbf{D}\big)\\ &+ \lambda_{\mathrm{LPIPS}}\mathcal{L}_{\mathrm{LPIPS}}\big(\hat{\mathbf{D}},\mathbf{D}\big),\end{aligned} \quad (34)$$

where λ_{MSE}, $\lambda_{\mathrm{MS-SSIM}}$ and λ_{LPIPS} are the weights that balance the contribution of each loss component. By combining these loss functions, our model is trained to achieve pixel-level accuracy, capture multi-scale structural similarity and maintain high-level perceptual qualities, thereby generating images that are appealing to the human visual system.

TABLE I: NMSE performance of GMMV-LAMP at $P_t = 20\,\text{dBm}$.

Number of pilot OFDM symbols L	4	8	16	32	64
NMSE	0.205	0.063	0.020	0.011	0.006

IV. Numerical Results

A. Baseline Schemes

For the performance evaluation, we compare our proposed scheme with the following baseline schemes.

- **BPG/bmshj2018 + LDPC + QAM + RZF**: In this case, the airship-borne BS employs the advanced image coding method better portable graphics (BPG) [63] or the DL-based bmshj2018 model [64] for image source coding and decoding, while low-density parity-check (LDPC) code with 1/2, 2/3, or 3/4 code rate is utilized for channel coding and decoding. Constellation modulation includes binary phase shift keying (BPSK), quadrature phase shift keying (QPSK), or 16 quadrature amplitude modulation (QAM), and regularized zero forcing (RZF) is employed for MIMO beamforming. The search for optimal combinations of source coding compression rate, LDPC code rate, and constellation modulation order is conducted to obtain the best performance configuration for this baseline scheme.
- **BPG/bmshj2018 + E2E transmission network**: In this case, the airship BS adopts BPG or bmshj2018 model to complete the image source coding and decoding, LDPC to complete the channel coding and decoding, and extends the autoencoder-based E2E OFDM transmission network in [65] to a multi-user MIMO system for achieving the physical-layer transmission, where the airship BS transmits 2 bits to each UE on each subcarrier during each OFDM symbol.
- **BPG/bmshj2018 +capacity**: In this case, the airship BS adopts the BPG or bmshj2018 model for image source coding and decoding, and is able to transmit the coded bits perfectly at the rate of the channel capacity. These two schemes can be considered as the ideal performance upper bounds.
- **ADJSCC**: In this case, we adopt the attention DL based JSCC (ADJSCC) scheme [24] for the E2E transmission of images. Notably, the original AWGN channel is substituted with an equivalent channel after RZF beamforming.

B. Dataset for Neutral Network Training

The channel dataset is generated following the sparse multipath channel model for airship-borne massive MIMO communication scenarios, as outlined in Subsection II. Specifically, the airship-borne BS is equipped with an 8×8 UPA maintaining a half-wavelength antenna spacing, the number of subcarriers is $N_c = 64$, and the subcarrier spacing is 60 kHz. The noise power spectral density is -174 dBm/Hz, corresponding to a noise power of $\sigma_n = -108$ dBm. For the path gain $\alpha_{l,k}$, it can be represented as the product of large-scale fading and small-scale fading, where the large-scale fading can be expressed as

$$\frac{G_t G_r \lambda_c^2 \Gamma[k,n]}{(4\pi)^3 d_t^2[k,n] d_r^2[k,n]}$$

where G_t and G_r are the transmit and receive antenna gains, $\Gamma[k,n]$ is the scattering cross-section, $d_t[k,n]$ is the distance from the BS to the scatterer, and $d_r[k,n]$ is the distance from the scatterer to the UE. We assume the distance from the airship to the scatterer is about 20 km, the distance from the scatterer to the UE is about 10 m, the transmit and receive antenna gains are both 15 dB, and the scattering cross-section is about 0.1 m^2, resulting in large-scale fading of approximately $\alpha = -138$ dB. To simplify the simulation, we set the path gain as an i.i.d. complex Gaussian distribution, i.e., $\alpha_{l,k} \sim \mathcal{CN}(0, \alpha)$. The number of paths and UEs are set to 2 and 4, respectively. Furthermore, throughout the process of training and validating the proposed deep JSCBF scheme, we consistently employ an advanced channel estimation method based on generalized multiple measurement-vectors (GMMV)-learnable approximate message passing (LAMP) [66] to acquire the estimated CSI as the input of the deep JSCBF network. This ensures that the neural network can learn to mitigate the performance degradation caused by CSI estimation errors. Table I shows the performance of the GMMV-LAMP channel estimation scheme in terms of normalized mean square error (NMSE) at 50 dBm transmit power, evaluated over different numbers of OFDM pilot symbols. It can be observed that the channel estimation results are very poor in the case of low pilot overhead. For the proposed deep JSCBF scheme, as well as for all baseline schemes, a fair comparison is made based on the imperfect CSI estimated by GMMV-LAMP.

TABLE II: Channel parameter settings

Parameter	Value
Carrier frequency [GHz]	2.6
Subcarrier spacing [kHz]	60
Number of subcarriers	64
Antenna array size	64
Number of single-antenna UEs	4
Azimuth angle range	$[-\pi,\ \pi]$
Zzenith angles range	$[0,\ \pi/4]$
Height of the BS	20 km
noise power spectral density	-174 dBm/Hz

For the image dataset, we adopt the open-source ImageNet dataset [67], which is divided into training, validation, and test sets containing 100000, 10000, and 10000 image samples, respectively.

C. Training Settings

We use the open-source DL library PyTorch to train and validate the proposed deep JSCBF scheme on a computer with dual Nvidia GeForce GTX 3090 GPUs. During the training process, we adopt the Adam optimizer and use the loss function introduced in the previous section to train the entire neural network, with the training set batch size set to 16. The learning rate is set to 10^{-4}. During the training process, we do not impose any constraints on the network's output. However, during the validation phase, we clip the network's output to the range of 0 to 1 to ensure that the resulting images are displayed correctly. To speed up the training process, we first pre-train both the image semantic extraction network and the semantic decoding network without considering the physical layer transmission process. Then, these networks are integrated with other modules for joint E2E training.

Regarding the network parameters, the Swin Transformer in the image semantic extraction network and the semantic decoding network has a window size of $M = 4$, a patch size of $N = 2$, an embedding dimension of 48, and a total of 4 layers. For the other networks using the Transformer model, the entire Transformer structure consists of 4 layers, with each layer comprising a multi-head self-attention mechanism and an MLP. The multi-head self-attention layers have a dimension of 256 with 8 heads, and the MLP dimensions are set to 1024.

D. Performance Comparison

In Fig. 9, we compare the performance of the proposed deep JSCBF scheme with the baseline schemes in terms of the pixel-level distortion metric PSNR as well as the perceptual metrics

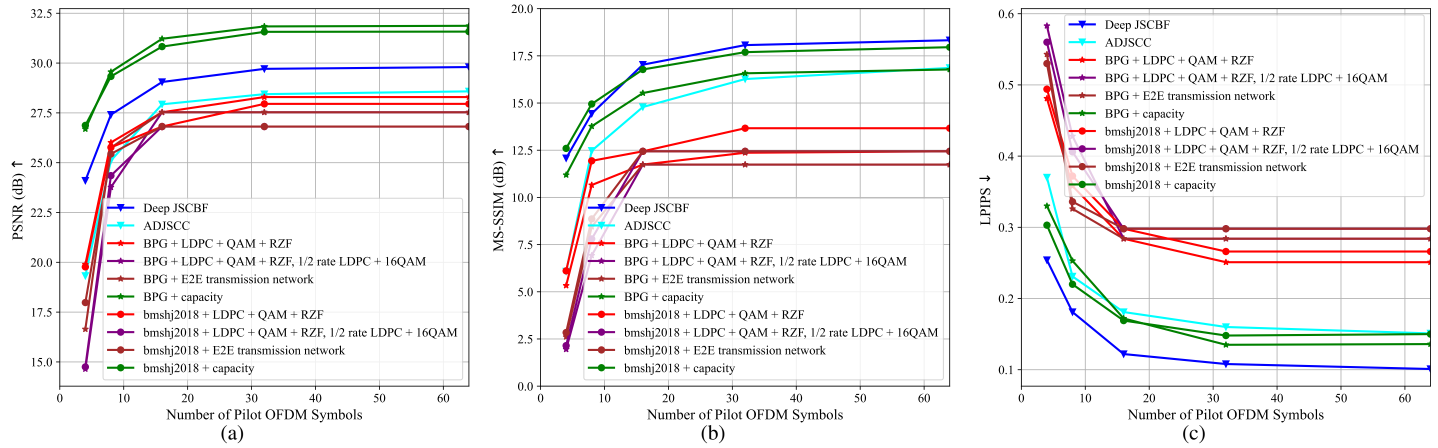


Fig. 9. Performance comparison of different solutions versus the number of pilot OFDM symbols L at $P_t = 50$ dBm, given $Q = 128$: (a) PSNR performance comparison; (b) MS-SSIM performance comparison; (c) LPIPS performance comparison.

MS-SSIM and LPIPS over different numbers of pilot OFDM symbols L, given $P_t = 50\,\text{dBm}$ and the number of transmitted OFDM symbols $Q = 128$. It is important to note that since both BPG and bmshj2018 are variable-length source coding schemes, they cannot achieve an arbitrarily specified compression rate. Therefore, we exhaustively search the quality factor of BPG and bmshj2018, as well as the LDPC code rate and the QAM modulation order, to obtain a number of transmitted OFDM symbols Q that is closest to that of the proposed deep JSCBF scheme to ensure a fair comparison.

In Fig. 9(a), it is observed that the 'BPG/bmshj2018+LDPC +QAM+RZF, 1/2 rate LDPC+16QAM' schemes experience significant performance degradation under inaccurate CSI estimation (i.e., $L \leq 8$), a phenomenon often referred to as the 'cliff effect'. This is attributed to a large mismatch between the actual CSI and the estimated CSI, leading to a large number of errors that interfere with the normal functioning of the source decoding module. Although the 'BPG/bmshj2018+E2E transmission network' schemes adopt a DL-based E2E physical-layer design, their source coding and physical-layer design remain separate and are thus still susceptible to the 'cliff effect'. In contrast, the 'BPG/bmshj2018+LDPC+QAM+RZF' schemes perform better because they can ensure image compression quality while keeping the bit error rate as low as possible by exhaustively searching the source coding compression rate, LDPC code rate, and QAM modulation order. The existing ADJSCC scheme improves the robustness to imperfect CSI with E2E training, and performs close to the 'bmshj2018/BPG+LDPC+QAM+RZF' schemes. In comparison, the proposed deep JSCBF scheme jointly exploits the channel and image semantics to design a hybrid data and model-driven beamforming scheme, and integrates the beamforming module with the semantic extraction and decoding modules for joint E2E training, thereby achieving optimal performance. Especially under imprecise CSI estimation, the PSNR of the proposed deep JSCBF scheme outperforms the other baseline schemes by nearly 5 dB. However, there is still a gap of about 2 dB in PSNR compared to the idealized 'BPG/bmshj2018+capacity'.

It can be seen from Fig. 9(b) and Fig. 9(c) that the existing ADJSCC scheme has a significant advantage over traditional separate design approaches in perceptual metrics such as MS-SSIM and LPIPS. The proposed deep JSCBF scheme exhibits the best perceptual performance. In particular, the MS-SSIM of the proposed deep JSCBF scheme exceeds that of the separate design approaches by more than 5 dB, and the LPIPS of the proposed deep JSCBF scheme is less than half that of the separate design approaches. In addition, compared to the idealized 'BPG/bmshj2018 + capacity', the proposed deep JSCBF scheme exhibits better MS-SSIM performance and significantly better LPIPS performance. This is because traditional source coding optimizes images based only on pixel-level distortion, which fails to adequately extract semantic information that is more relevant to perceptual metrics. In contrast, the proposed deep JSCBF scheme, through the E2E training process, can extract semantic information most relevant to a loss function composed of pixel-level distortion loss ($\mathcal{L}_{\text{MSE}}$) and perceptual losses ($\mathcal{L}_{\text{MS-SSIM}}$ and $\mathcal{L}_{\text{LPIPS}}$), filtering out redundant information. This approach balances image pixel-level distortion performance with perceptual performance, while reducing image transmission overhead. Therefore, although the proposed deep JSCBF scheme may not perform as well as the idealized 'BPG/bmshj2018 + capacity' in terms of pixel-level distortion, it has a significant advantage in perceptual metrics.

Fig. 10 shows the performance as a function of transmit power P_t achieved by the different schemes. It can be seen that the traditional separate design approaches experience significant

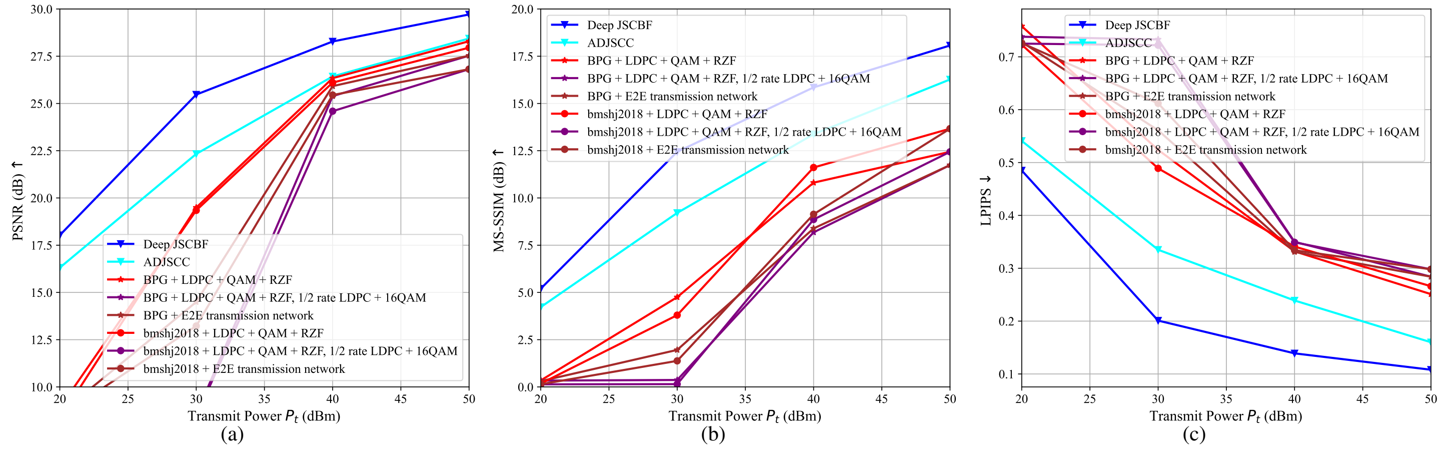
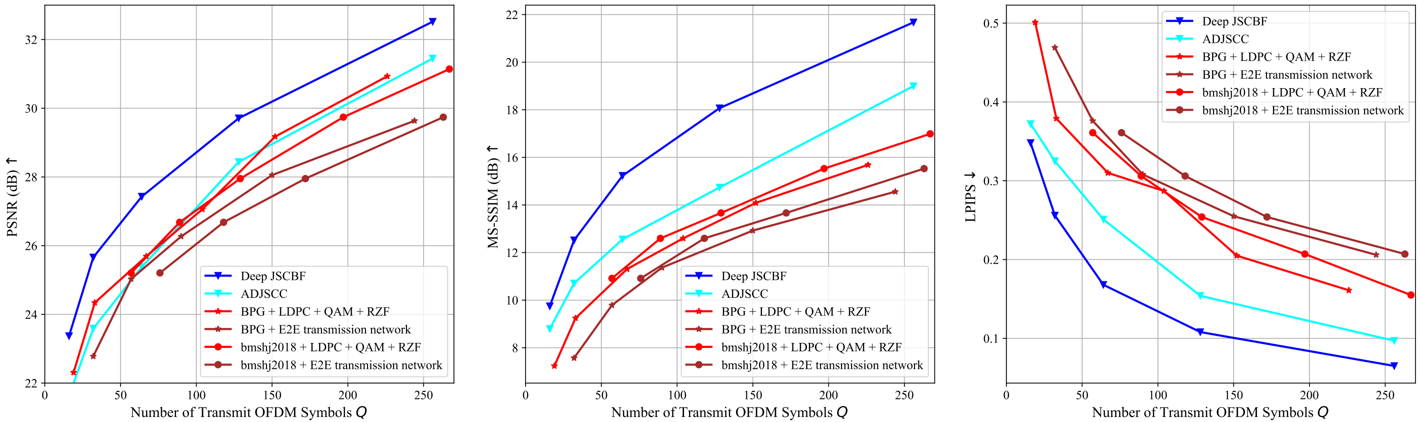


Fig. 10. Performance comparison of different solutions versus transmit power P_t at $L = 32$, given $Q = 128$: (a) PSNR performance comparison; (b) MS-SSIM performance comparison; (c) LPIPS performance comparison.



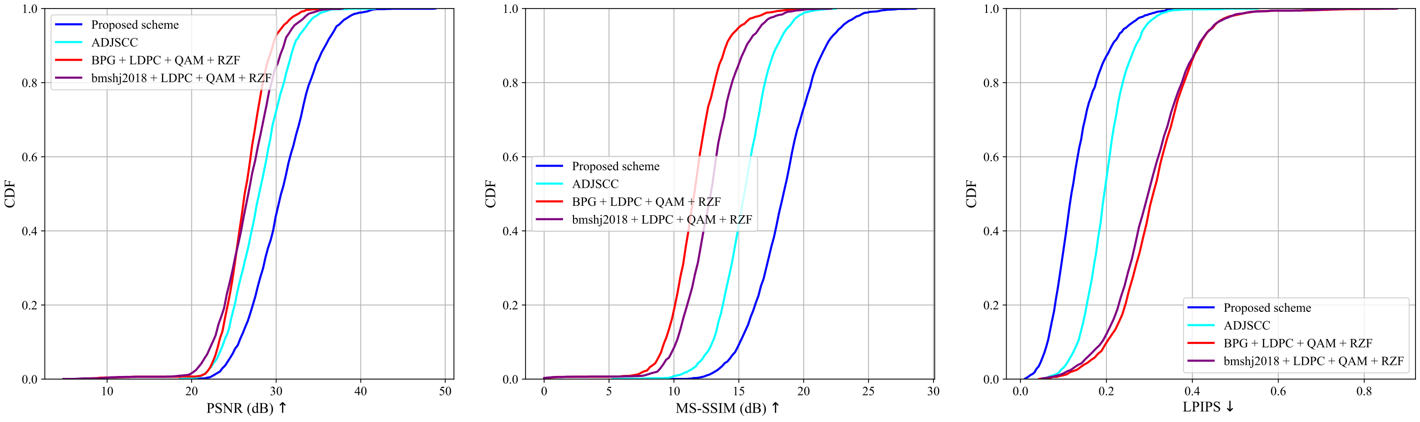
(a) (b) (c)

Fig. 11. Performance comparison of different solutions versus transmit OFDM symbols Q, given $P_t = 50\,\text{dBm}$ and $L = 32$: (a) PSNR performance comparison; (b) MS-SSIM performance comparison; (c) LPIPS performance comparison.

performance degradation under low transmit power conditions. In contrast, the ADJSCC scheme, with its integrated optimization of source and channel coding, shows superior robustness to low transmit power compared to the traditional separate designs. In particular, the proposed deep JSCBF scheme achieves the best performance. In terms of PSNR, as shown in Fig. 10(a), the proposed deep JSCBF scheme demonstrates a gain of about 2 dB over the traditional separate design approaches at high transmit power levels (i.e., $P_t \geq 40\,\text{dBm}$), while it outperforms the traditional separate design approaches by more than 5 dB at low transmit power levels, highlighting its robustness to challenging signal conditions. In addition, the proposed deep JSCBF scheme consistently exhibits a 2 dB improvement in PSNR over all transmit power levels compared to the ADJSCC, which is mainly attributed to the optimization of the physical layer beamforming in our approach. In terms of perceptual metrics, as shown in Fig. 10(b) and Fig. 10(c), the proposed deep JSCBF scheme exhibits significant improvements over the conventional separate design approaches. The MS-SSIM of our scheme surpasses that of the separate designs by more than 5 dB, and its LPIPS is less than half that of the separate designs. Furthermore, the perceptual metrics of the proposed deep JSCBF scheme are consistently better than the ADJSCC at all transmit power levels. These results confirm the effectiveness of the proposed deep JSCBF scheme in exploiting semantically aware hybrid data and model-driven beamforming for improved image reconstruction performance.

Fig. 11 shows the performance of the different schemes as a function of the number of transmitted OFDM symbols Q, where a smaller Q implies the use of a higher compression rate to compress the images, thereby reducing the communication overhead. It is evident that for all values of Q, the proposed deep JSCBF scheme significantly outperforms both the separate design approaches and ADJSCC in terms of pixel-level distortion metrics and perceptual metrics. Therefore, to achieve the same level of performance, our deep JSCBF scheme imposes significantly lower communication overhead than the other schemes.

Fig. 12 shows the cumulative distribution functions (CDFs) of PSNR, MS-SSIM, and LPIPS performance for the reconstructed images under the different schemes, given $L = 16$, $Q = 128$ and $P_t = 50\,\text{dBm}$. Unlike the previous simulations, which focus only on the average power, this analysis considers the power distribution of all samples in the data set, providing a more comprehensive understanding beyond the aggregate average power. It is observed from Fig. 12(a) that with the application of the proposed deep JSCBF scheme, approximately 80% of the reconstructed images exceed a PSNR of 27dB, while less than 50% of the images reconstructed by the other baseline schemes achieve this performance level. In terms of perceptual metrics, as can be seen from Fig. 12(b) and Fig. 12(c), 80% of the images reconstructed by the proposed deep JSCBF scheme exceed an MS-SSIM of 17 dB and maintain an LPIPS loss below 0.17, while only about 20% of the images reconstructed by the ADJSCC scheme meet this benchmark, and even fewer, about 5%, of the images reconstructed by the other separate design



(a) (b) (c)

Fig. 12. The CDFs of the performance distributions for the reconstructed images achieved by different schemes, given $Q = 128$, $L = 16$ and $P_t = 50\,\text{dBm}$: (a) PSNR performance comparison; (b) MS-SSIM performance comparison; (c) LPIPS performance comparison.

TABLE III: Ablation study.

Number of pilot OFDM symbols L		4	8	16	32	64
PSNR (dB)/ MS-SSIM (dB)/ LPIPS	Deep JSCBF	24.10/12.07/0.254	27.41/14.43/1.188	29.05/17.03/0.125	29.71/18.0/0.113	29.80/19.32/0.106
	Deep JSCBF (model-driven beamforming only)	22.89/10.51/0.306	26.56/13.87/0.215	28.74/16.57/0.139	29.54/17.81/0.121	29.63/19.15/0.112
	ADJSCC	19.09/5.56/0.392	24.91/12.28/0.249	27.33/14.58/0.194	28.04/15.98/0.170	28.15/16.67/0.162

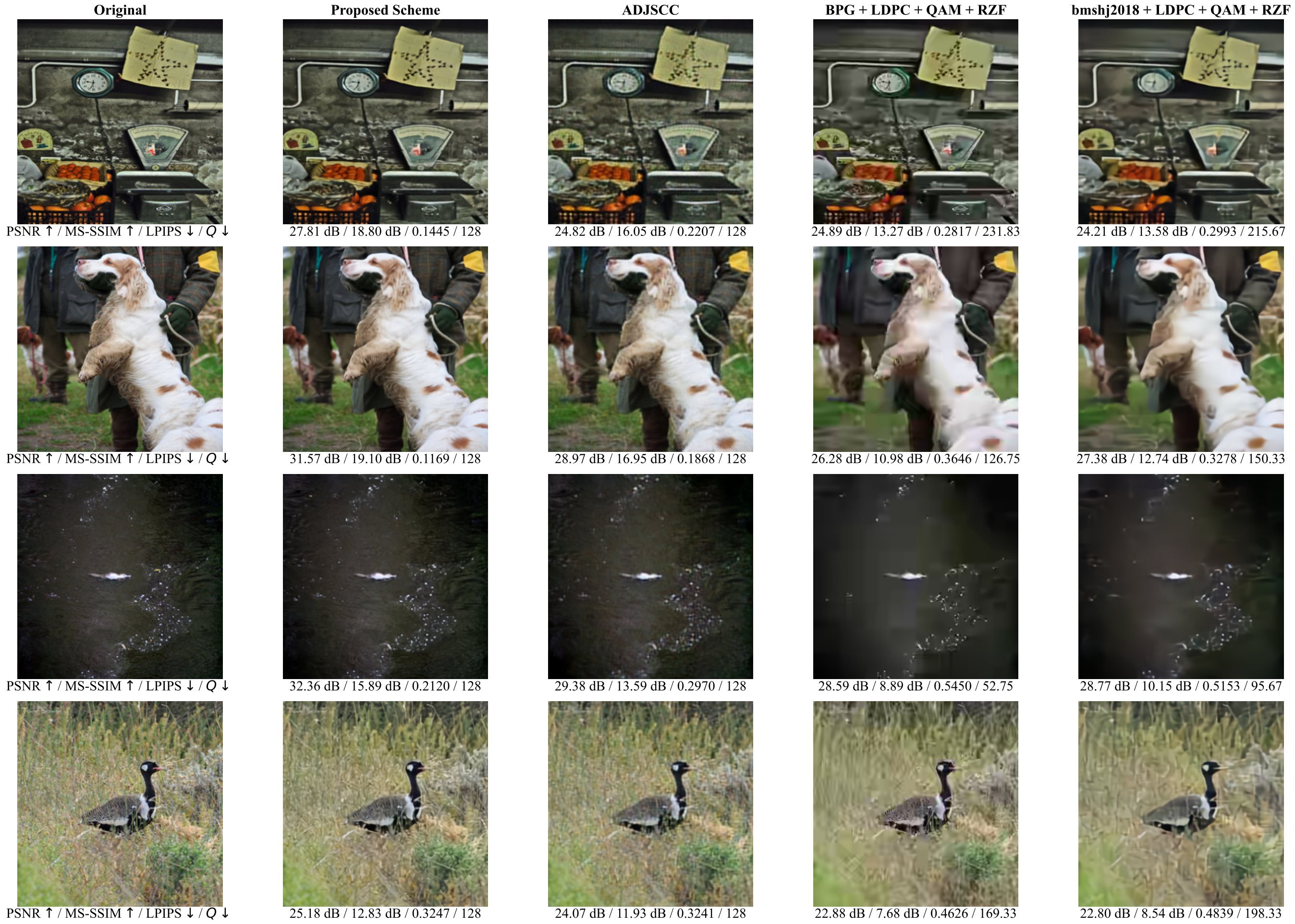

Original
Proposed Scheme
ADJSCC
BPG + LDPC + QAM + RZF
bmshj2018 + LDPC + QAM + RZF
PSNR ↑ / MS-SSIM ↑ / LPIPS ↓ / Q ↓
27.81 dB / 18.80 dB / 0.1445 / 128
24.82 dB / 16.05 dB / 0.2207 / 128
24.89 dB / 13.27 dB / 0.2817 / 231.83
24.21 dB / 13.58 dB / 0.2993 / 215.67
PSNR ↑ / MS-SSIM ↑ / LPIPS ↓ / Q ↓
31.57 dB / 19.10 dB / 0.1169 / 128
28.97 dB / 16.95 dB / 0.1868 / 128
26.28 dB / 10.98 dB / 0.3646 / 126.75
27.38 dB / 12.74 dB / 0.3278 / 150.33
PSNR ↑ / MS-SSIM ↑ / LPIPS ↓ / Q ↓
32.36 dB / 15.89 dB / 0.2120 / 128
29.38 dB / 13.59 dB / 0.2970 / 128
28.59 dB / 8.89 dB / 0.5450 / 52.75
28.77 dB / 10.15 dB / 0.5153 / 95.67
PSNR ↑ / MS-SSIM ↑ / LPIPS ↓ / Q ↓
25.18 dB / 12.83 dB / 0.3247 / 128
24.07 dB / 11.93 dB / 0.3241 / 128
22.88 dB / 7.68 dB / 0.4626 / 169.33
22.80 dB / 8.54 dB / 0.4839 / 198.33

Fig. 13. Examples of visual comparison for different schemes.

approaches reach this level. These observations attest to the ability of the proposed deep JSCBF scheme to reconstruct images with a high probability of significantly superior performance compared to traditional schemes.

Finally, Fig. 13 presents a visual comparison of the reconstructed images obtained by the different schemes, given $L = 16$, $Q = 128$, and $P_t = 50\,\text{dBm}$. It can be seen that the visual representation of the images reconstructed by the proposed deep JSCBF scheme has higher clarity and better detail representation compared to the baseline schemes. This further confirms the fact that the proposed deep JSCBF scheme has significant advantages over the other baseline approaches, both in terms of pixel-level distortion and perceptual metrics.

E. Ablation Study

To illustrate the purpose of using a semantic-aware hybrid of data and model-driven beamforming, and to investigate the impact of estimated CSI, Table III presents relevant ablation experiments. Here, 'deep JSCBF (model-driven beamforming only)' refers to a simplified version of the proposed deep JSCBF scheme with the semantic-aware data-driven beamforming module removed. It is observed that the proposed deep JSCBF scheme, when using model-driven beamforming only, still exhibits performance advantages over the traditional ADJSCC. This is attributed to the use of semantic information to optimize the key parameter set $\mathcal{A}$ in beamforming, thus better addressing the performance degradation problem caused by the use of CSI estimation. The model-driven beamforming approach that integrates expert knowledge can achieve satisfactory performance when the number of OFDM pilot symbols is sufficient. However, in the case of insufficient number of pilot OFDM symbols, the model-driven beamforming approach suffers from significant performance degradation due to inaccurate CSI estimation. In contrast, the proposed deep JSCBF scheme builds on the model-driven approach by adding an additional data-driven branch to compensate for the shortcoming of the model-driven branch. This exploitation of the data-driven method enables the proposed deep JSCBF to achieve better performance under inaccurate CSI estimation.

F. Analysis of the effectiveness and performance improvements

The proposed networks utilize Transformer or Swin Transformer architectures, which excel at capturing long-range dependencies and global context. This capability is crucial for semantic feature extraction and fusion. The self-attention mechanism within Transformers allows the network to focus on relevant parts of the input data, enhancing the quality of semantic features extracted from both images and CSI.

Unlike previous Deep JSCC methods, our approach integrates semantic communication with large-scale MIMO beamforming. By jointly optimizing the processes of semantic extraction, fusion, and beamforming, our method achieves superior end-to-end performance. Previous works on Deep JSCC primarily addressed single-user, single-antenna, narrowband systems. In contrast, our approach extends these concepts to multi-user, broadband MIMO systems, delivering significant performance improvements in more complex scenarios.

Additionally, our proposed precoding scheme combines data-driven and model-driven approaches. By integrating data-driven deep learning models with traditional model-driven methods, we enhance system robustness and performance. The data-driven approach leverages extensive training data to capture complex channel characteristics, while the model-driven approach uses prior knowledge from physical models to ensure reliable and

TABLE IV: Complexity and Running Times of Different Schemes

Schemes	Complexity	Running times of each sample
bmshj2018 or ADJSCC	$\mathcal{O}\left(UM_xM_yC^2k_s^2\right)$	1.54ms-6.72ms
bpg	$\mathcal{O}\left(M_xM_yN_{\text{bpg}}P_{\text{bpg}}\right)$	0.86s
LDPC	$\mathcal{O}\left(I\left(d_vn+d_cn\left(1-a\right)\right)\right)$	4.37s
RZF	$\mathcal{O}\left(N_cN_t^2K+N_cN_tK^2\right)$	9.01ms
Proposed image semantic extraction network or image semantic decoding network	$\mathcal{O}\left(UM_xM_yM^2d_{\text{model}}/N^2+UM_xM_yd_{\text{model}}^2/N^2\right)$	2.09ms-2.35ms
Proposed CSI semantic extraction network semantic fusion network data-driven semantic-aware beamforming or model-driven semantic-aware beamforming	$\mathcal{O}\left(UN_c^2d_{\text{model}}+UN_cd_{\text{model}}^2\right)$	2.01ms-9.72ms

efficient precoding. This dual-driven precoding strategy demonstrates remarkable performance gains in complex environments.

G. Computational Complexity Analysis

This subsection investigates the computational complexity of different schemes. For the DL-based schemes, since there is no strict time limit during the offline training stage, we only consider the computational complexity at the online testing stage. Additionally, except for the LDPC and BPG schemes, which run on the CPU, all other schemes are executed on the GPU. The computational complexity analysis and running times of the different schemes are presented in Table IV. The details are as follows.

- The computational complexity of the bmshj2018 and ADJSCC scheme is primarily derived from convolution operations, with a complexity of $\mathcal{O}\left(UM_xM_yC^2k_s^2\right)$, where C represents the number of feature channels and k_s denotes the convolution kernel size.
- The computational complexity of LDPC codes is primarily due to the decoding process. The decoding complexity is $\mathcal{O}\left(I\left(d_vn+d_cn\left(1-a\right)\right)\right)$, where I is the number of iterations, d_v and d_c are the average numbers of non-zero entries per column and row in the parity-check matrix, respectively, n is the total number of bits, and a is the code rate.
- The computational complexity of the RZF precoding scheme primarily arises from the matrix inversion of the channel, with a complexity of $\mathcal{O}\left(N_cN_t^2K+N_cN_tK^2\right)$.
- The proposed image semantic extraction network and image semantic decoding network leverage the Swin Transformer architecture. The complexity of these networks is dominated by the self-attention mechanism within local windows with the complexity of $\mathcal{O}\left(UM_xM_yM^2d_{\text{model}}/N^2\right)$ and the FFN with a complexity of $\mathcal{O}\left(UM_xM_yd_{\text{model}}^2/N^2\right)$, where d_{model} is the model dimensionality.
- The proposed CSI semantic extraction network, semantic fusion network, data-driven semantic-aware beamforming, and model-driven semantic-aware beamforming utilize the Transformer architecture. The primary sources of computational complexity in these networks are the self-attention mechanism with a complexity of $\mathcal{O}\left(UN_c^2d_{\text{model}}\right)$ and and the FFN with a complexity of $\mathcal{O}\left(UN_cd_{\text{model}}^2\right)$.

To provide a clear view of the computational complexity of different schemes, Table IV also presents the running times of each scheme. The results indicate that the computational overhead of the proposed scheme is acceptable.

V. Conclusions

This paper has proposed a novel deep JSCBF approach for airship-based massive MIMO near space image transmission network, where the semantic encoding/decoding and MIMO beamforming modules are collectively modeled as a unified E2E neural network, which includes image and CSI semantic extraction networks, a semantic fusion network, hybrid data and model-driven semantic-aware beamforming networks, and a semantic decoding network. Specifically, we have built the image and CSI semantic extraction networks based on the transformer architecture to extract semantics from both image source data and CSI, which are used to support the subsequent semantic fusion and beamforming. A semantic fusion network has been developed to fuse the semantics of image source data and CSI to form complex-valued semantic features for subsequent physical layer transmission. To balance the advantages of data-driven and model-driven DL, we further designed the hybrid data and model-driven semantic-aware beamforming networks. The results of these two beamforming networks are then weighted and merged to improve the beamforming performance. Finally, a semantic decoding network based on the Swin Transformer architecture was employed at the UE side to reconstruct images from the received signals. We have performed E2E joint training for all the modules using a loss function that combines MSE, MS-SSIM, and LPIPS. Numerous simulation results have demonstrated that the proposed deep JSCBF scheme significantly outperforms existing separation module design schemes as well as the existing deep JSCC, especially in the case of low transmit power or insufficient pilot overhead.

References

[1] H. Liu, *et al.*, "Near-space communications: The last piece of 6G space-air-ground-sea integrated network puzzle," accepted by *Space: Science and Technology*, 2023. Appeared in: https://spj.science.org/doi/10.34133/space.0176

[2] X. Cao, P. Yang, and X. Su, "Survey on near-space information networks: Channel modeling, networking, and transmission perspectives," Oct. 2023. [Online]. Available: https://arxiv.org/abs/2310.09025

[3] B. Jiang, *et al.*, "Total and minimum energy efficiency tradeoff in robust multigroup multicast satellite communications," *Space Sci. Technol.*, 2023.

[4] Y. Liao, *et al.*, " Integration of communication and navigation technologies toward LEO-enabled 6G networks: A survey," *Space Sci. Technol.*, 2023.

[5] L. Qiao, J. Zhang, Z. Gao, D. Zheng, M. J. Hossain, Y. Gao, D. W. K. Ng, and M. Di Renzo, "Joint activity and blind information detection for UAV-assisted massive IoT access," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 5, pp. 1489-1508, May 2022.

[6] Z. Gao, *et al.*, "Data-driven deep learning based hybrid beamforming for aerial massiveMIMO-OFDM systems with implicit CSI," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 10, pp. 2894–2913, Oct. 2022.

[7] D. Gündüz, *et al.*, "Beyond transmitting bits: Context, semantics, and task-oriented communications," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 5–41, Jan. 2023.

[8] K. Niu, *et al.*, "A paradigm shift toward semantic communications," *IEEE Commun. Mag.*, vol. 60, no. 11, pp. 113–119, No.v 2022.

[9] W. Yang, *et al.*, "Semantic communications for future Internet: Fundamentals, applications, and challenges," *IEEE Commun. Surveys and Tutorials*, vol. 25, no. 1, pp. 213–250, Firstquarter 2023.

[10] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical J.*, vol. 27, no. 3, pp. 379–423, 1948.

[11] Y. Mei, Z. Gao, D. Mi, M. Zhou, D. Zheng, M. Matthaiou, P. Xiao, and R. Schober, "Massive access in extra large-scale MIMO With mixed-ADC over near-field channels," *IEEE Trans. Veh. Technol.*, vol. 72, no. 9, pp. 12373-12378, Sep. 2023.

[12] M. Ke, Z. Gao, Y. Wu, X. Gao, and R. Schober, "Compressive sensing-based adaptive active user detection and channel estimation: Massive access meets massive MIMO," *IEEE Trans. Signal Process.*, vol. 68, pp. 764-779, Jan. 2020.

[13] J. Zhang, *et al.*, "Prospective multiple antenna technologies for beyond 5G," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1637–1660, Aug. 2020.

[14] Y. F. Liu *et al.* "A survey of advances in optimization methods for wireless communication system design." to appear in *IEEE J. Sel. Areas Commun.*, arXiv preprint arXiv:2401.12025, 2024.

[15] M. Ke, Z. Gao, M. Zhou, D. Zheng, D. W. K. Ng, and H. V. Poor, "Next-generation URLLC with massive devices: A unified semi-blind detection framework for sourced and unsourced random access," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 7, pp. 2223-2244, Jul. 2023.

[16] K. Ying, Z. Gao, S. Chen, M. Zhou, D. Zheng, S. Chatzinotas, B. Ottersten, and H. V. Poor, "Quasi-synchronous random access for massive MIMO-based LEO satellite constellations," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 6, pp. 1702-1722, Jun. 2023.

[17] M. Ke, Z. Gao, Y. Huang, G. Ding, D. W. K. Ng, Q. Wu, and J. Zhang, "An edge computing paradigm for massive IoT connectivity over high-altitude platform networks," *IEEE Wireless Commun.*, vol. 28, no. 5, pp. 102-109, Oct. 2021.

[18] Y. Mei, Z. Gao, Y. Wu, W. Chen, D. W. K. Ng, and M. Di Renzo, "Compressive sensing-based joint activity and data detection for grant-free massive IoT access," *IEEE Trans. Wireless Commun.*, vol. 21, no. 3, pp. 1851-1869, Mar. 2022.

[19] L. Qiao, J. Zhang, Z. Gao, D. W. K. Ng, M. D. Renzo, and M. -S. Alouini, "Massive access in media modulation based massive machine-type communications," *IEEE Trans. Wireless Commun.*, vol. 21, no. 1, pp. 339-356, Jan. 2022.

[20] Z. Gao, M. Ke, Y. Mei, L. Qiao, S. Chen, Derrick W. K. Ng, and H. V. Poor, "Compressive sensing-based grant-free massive access for 6G massive communication," *IEEE Internet Things J.*, vol. 11, no. 5, pp. 7411-7435, Mar. 2024.

[21] E. Bourtsoulatze, D. B. Kurka, and D. Gündüz, "Deep joint source-channel coding for wireless image transmission," *IEEE Trans. Cogn. Commun. Netw.*, vol. 5, no. 3, pp. 567–579, Sep. 2019.

[22] J. Dai, *et al.*, "Nonlinear transform source-channel coding for semantic communications," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 8, pp. 2300–2316, Aug. 2022.

[23] J. Dai, *et al.*, "Toward adaptive semantic communications: Efficient data transmission via online learned nonlinear transform source-channel coding," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 8, pp. 2609–2627, Aug. 2023.

[24] J. Xu, B. Ai, W. Chen, A. Yang, P. Sun, and M. Rodrigues, "Wireless image transmission using deep source channel coding with attention modules," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 2315–2328, Apr. 2022.

[25] T. Wu, *et al.*, "CDDM: channel denoising diffusion models for wireless semantic communications," *IEEE Trans. Wireless Commun.*, 2024.

[26] Z. Weng, *et al.*, "Deep learning enabled semantic communications with speech recognition and synthesis," *IEEE Trans. Wireless Commun.*, vol. 22, no. 9, pp. 6227–6240, Sep. 2023.

[27] Z. Weng and Z. Qin, "Semantic communication systems for speech transmission," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 8, pp. 2434–2444, Aug. 2021.

[28] G. Zhang, *et al.*, "A unified multi-task semantic communication system for multimodal data," *arXiv preprint arXiv:2209.07689*, 2022.

[29] P. Jiang, C.-K. Wen, S. Jin, and G. Y. Li, "Wireless semantic transmission via revising modules in conventional communications," *IEEE Wireless Commun.*, vol. 30, no. 3, pp. 28–34, Jun. 2023.

[30] Z. Yang, M. Chen, Z. Zhang, and C. Huang, "Energy efficient semantic communication over wireless networks with rate splitting," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 5, pp. 1484–1495, May 2023.

[31] H. Zhang, *et al.*, "DRL-driven dynamic resource allocation for task-oriented semantic communication," *IEEE Trans. Commun.*, vol. 71, no. 7, pp. 3992–4004, Jul. 2023.

[32] M. Yang, C. Bian, and H.-S. Kim, "OFDM-guided deep joint source channel coding for wireless multipath fading channels," *IEEE Trans. Cogn. Commun. Netw.*, vol. 8, no. 2, pp. 584–599, Jun. 2022.

[33] M. Wang, J. Li, M. Ma, and X. Fan, "Constellation design for deep joint source-channel coding," *IEEE Signal Process. Lett.*, vol. 29, pp. 1442–1446, Jul. 2022.

[34] X. Luo, *et al.*, "Multi-modal and multi-user semantic communications for channel-level information fusion," *IEEE Wireless Commun.*, Early Access pp. 1–18, Oct. 2022.

[35] G. Zhang, Q. Hu, Y. Cai, and G. Yu, "SCAN: Semantic communication with adaptive channel feedback," *arXiv preprint arXiv:2306.15534*, 2023.

[36] H. Wu, *et al.*, "Vision transformer for adaptive image transmission over MIMO channels," in *Proc. ICC 2023* (Rome, Italy), May 28-Jun. 1, 2023, pp. 3702–3707.

[37] C. B. Peel, B. M. Hochwald, and A. L. Swindlehurst, "A vector-perturbation technique for near-capacity multiantenna multiuser communication–part I: Channel inversion and regularization," *IEEE Trans. Commun.*, vol. 53, no. 1, pp. 195–202, Jan. 2005.

[38] M. Sadek, A. Tarighat, and A. H. Sayed, "A leakage-based precoding scheme for downlink multi-user MIMO channels," *IEEE Trans. Wireless Commun.*, vol. 6, no. 5, pp. 1711–1721, May 2007.

[39] N. Sidiropoulos, T. Davidson, and Z.-Q. Luo, "Transmit beamforming for physical-layer multicasting," *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 2239–2251, Jun. 2006.

[40] Q. Shi, M. Razaviyayn, Z.-Q. Luo, and C. He, "An iteratively weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4331–4340, Sep. 2011.

[41] Q. Shi and M. Hong, "Spectral efficiency optimization for millimeter wave multiuser MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 3, pp. 455–468, Jun. 2018.

[42] Z. Qin, H. Ye, G. Y. Li, and B.-H. F. Juang, "Deep learning in physical layer communications," *IEEE Wireless Commun.*, vol. 26, no. 2, pp. 93–99, Apr. 2019.

[43] H. Sun, *et al.*, "Learning to optimize: Training deep neural networks for interference management," *IEEE Trans. Signal Process.*, vol. 66, no. 20, pp. 5438–5453, Oct. 2018.

[44] W. Lee, M. Kim, and D.-H. Cho, "Deep power control: Transmit power control scheme based on convolutional neural network," *IEEE Commun. Lett.*, vol. 22, no. 6, pp. 1276–1279, Jun. 2018.

[45] W. Xia, *et al.*, "A deep learning framework for optimization of MISO downlink beamforming," *IEEE Trans. Commun.*, vol. 68, no. 3, pp. 1866–1880, Mar. 2020.

[46] Q. Yuan, H. Liu, M. Xu, Y. Wu, L. Xiao, and T. Jiang, "Deep learning-based hybrid precoding for terahertz massive MIMO communication with beam squint," *IEEE Commun. Lett.*, vol. 27, no. 1, pp. 175–179, Jan. 2023.

[47] Y. Yuan, *et al.*, "Transfer learning and meta learning-based fast downlink beamforming adaptation," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 1742–1755, Mar. 2021.

[48] F. Sohrabi, K. M. Attiah, and W. Yu, "Deep learning for distributed channel feedback and multiuser precoding in FDD massive MIMO," *IEEE Trans. Wireless Commun.*, vol. 20, no. 7, pp. 4044–4057, Jul. 2021.

[49] K. M. Attiah, F. Sohrabi, and W. Yu, "Deep learning for channel sensing and hybrid precoding in TDD massive MIMO OFDM systems," *IEEE Trans. Wireless Commun.*, vol. 21, no. 12, pp. 10839–10853, Dec. 2022.

[50] K. Kong, W.-J. Song, and M. Min, "Knowledge distillation-aided end-to-end learning for linear precoding in multiuser MIMO downlink systems with finite-rate feedback," *IEEE Trans. Veh. Technol.*, vol. 70, no. 10, pp. 11095–11100, Oct. 2021.

[51] Q. Hu, *et al.*, "Iterative algorithm induced deep-unfolding neural networks: Precoding design for multiuser MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1394–1410, Feb. 2021.

[52] J. Shi, *et al.*, "Robust WMMSE precoder with deep learning design for massive MIMO," *IEEE Trans. Commun.*, vol. 71, no. 7, pp. 3963–3976, Jul. 2023.

[53] Z. Gao, *et al.*, "Hybrid knowledge-data driven channel semantic acquisition and beamforming for cell-free massive MIMO," *IEEE J. Sel. Topics Signal Process.*, vol. 17, no. 5, pp. 964–979, Sep. 2023.

[54] M. Wu, *et al.*, "Deep learning-based rate-splitting multiple access for reconfigurable intelligent surface-aided Tera-hertz massive MIMO," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 5, pp. 1431–1451, May 2023.

[55] S. Li, C. Ding, L. Xiao, X. Zhang, G. Liu, and T. Jiang, "Expectation propagation aided model driven learning for OTFS signal detection," *IEEE Trans. Veh. Technol.*, vol. 72, no. 9, pp. 12407-12412, Seq. 2023.

[56] Z. Liu, *et al.*, "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. ICCV 20021*, Oct. 11-17, 2021, pp. 10012-10022.

[57] C.-K. Wen, W.-T. Shih, and S. Jin, "Deep learning for massive MIMO CSI feedback," *IEEE Wireless Commun. Lett.*, vol. 7, no. 5, pp. 748–751, Oct. 2018.

[58] A. Vaswani, *et al.*, "Attention is all you need," in *Proc. NIPS 2017* (Long Beach, CA, USA), Dec. 4-9, 2017, pp. 1–11.

[59] K. Han, *et al.*, "A survey on vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, Jan. 2023.

[60] A. Horé and D. Ziou, "Image quality metrics: PSNR vs. SSIM," in *Proc. 20th Int. Conf. Pattern Recognit.* (Istanbul, Turkey), Aug. 23-26, 2010, pp. 2366–2369.

[61] R. Zhang, *et al.*, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. CVPR 2018* (Salt Lake City, UT, USA), Jun. 18-22, 2018, pp. 586–595.

[62] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

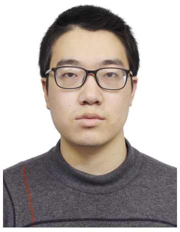
[63] F. Bellard, "BPG image format," *URL https://bellard.org/bpg*, vol. 1, no. 2, p. 1, 2015.

[64] J. Ballé, *et al.*, "Variational image compression with a scale hyperprior," in *Proc. ICLR 2018* (Vancouver, BC, Canada), Apr. 30-May 3, 2018, pp. 1–47.

[65] A. Felix, *et al.*, "OFDM-autoencoder for end-to-end learning of communications systems," in *Proc. SPAWC 2018* (Kalamata, Greece), Jun. 25-28, 2018, pp. 1–5.

[66] X. Ma, Z. Gao, F. Gao, and M. Di Renzo, "Model-driven deep learning based channel estimation and feedback for millimeter-wave massive hybrid MIMO systems," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 8, pp. 2388–2406, Aug. 2021.

[67] J. Deng, *et al.*, "ImageNet: A large-scale hierarchical image database," in *Proc. CVPR 2009* (Miami, FL, USA), Jun. 20-25, 2009, pp. 248–255.



Minghui Wu received the B.S. degree from the Beijing Institute of Technology, Beijing, China, in 2021. He is currently pursuing the Ph.D. degree in the School of Information and Electronics of Beijing Institute of Technology, Beijing, China. His research interests include channel estimation, massive MIMO, deep learning, and hybrid precoding.

Zhen Gao received the B.S. degree in information engineering from the Beijing Institute of Technology, Beijing, China, in 2011, and the Ph.D. degree in communication and signal processing with the Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, China, in 2016. He is currently a Professor with Beijing Institute of Technology. His research interests are in wireless communications, with a focus on multi-carrier modulations, multiple antenna systems, and sparse signal processing.

Dr. Gao was the recipient of IEEE Broadcast Technology Society 2016 Scott Helt Memorial Award (best paper), the recipient of Exemplary Reviewer of IEEE Communications Letters in 2016, the recipient of IET Electronics Letters Premium Award (Best Paper) 2016, and the recipient of Young Elite Scientists Sponsorship Program (2018-2021) by China Association for Science and Technology.

Zhaocheng Wang (Fellow, IEEE) received the B.S., M.S., and Ph.D. degrees from Tsinghua University, in 1991, 1993, and 1996, respectively.

From 1996 to 1997, he was a Post-Doctoral Fellow with Nanyang Technological University, Singapore. From 1997 to 1999, he was a Research Engineer/ Senior Engineer with the OKI Techno Centre (Singapore) Pte. Ltd., Singapore. From 1999 to 2009, he was a Senior Engineer/Principal Engineer with Sony Deutschland GmbH, Germany. Since 2009, he has been a Professor with the Department of Electronic Engineering, Tsinghua University, where he is currently the Director of the Broadband Communication Key Laboratory, Beijing National Research Center for Information Science and Technology (BNRist). He has authored or coauthored two books, which have been selected by IEEE Series on Digital and Mobile Communication and published by Wiley-IEEE Press. He has authored/coauthored more than 200 peer-reviewed journal articles. He holds 60 U.S./EU granted patents (23 of them as the first inventor). His research interests include wireless communications, millimeter wave communications, and optical wireless communications. He is a fellow of the Institution of Engineering and Technology. He was a recipient of the ICC2013 Best Paper Award, the OECC2015 Best Student Paper Award, the 2016 IEEE Scott Helt Memorial Award, the 2016 IET Premium Award, the 2016 National Award for Science and Technology Progress (First Prize), the ICC2017 Best Paper Award, the 2018 IEEE ComSoc Asia-Pacific Outstanding Paper Award, and the 2020 IEEE ComSoc Leonard G. Abraham Prize. He was an Associate Editor of IEEE TRANSACTIONS ON WIRELESS COMMUNICATION from 2011 to 2015 and IEEE COMMUNICATIONS LETTERS from 2013 to 2016. He is currently an Associate Editor of IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE SYSTEM JOURNAL, and IEEE OPEN JOURNAL OF VEHICULAR TECHNOLOGY.

Dusit Niyato (M'09-SM'15-F'17) is a professor in the College of Computing and Data Science, at Nanyang Technological University, Singapore. He received B.Eng. from King Mongkuts Institute of Technology Ladkrabang (KMITL), Thailand and Ph.D. in Electrical and Computer Engineering from the University of Manitoba, Canada. His research interests are in the areas of mobile generative AI, edge intelligence, decentralized machine learning, and incentive mechanism design.

George K. Karagiannidis (Fellow, IEEE) is currently Professor in the Electrical and Computer Engineering Dept. of Aristotle University of Thessaloniki, Greece and Head of Wireless Communications & Information Processing (WCIP) Group. He is also Faculty Fellow in the Artificial Intelligence & Cyber Systems Research Center, Lebanese American University. His research interests are in the areas of Wireless Communications Systems and Networks, Signal processing, Optical Wireless Communications, Wireless Power Transfer and Applications and Communications & Signal Processing for Biomedical Engineering. Dr. Karagiannidis is the Editor-in Chief of IEEE Transactions on Communications and in the past was the Editor-in Chief of IEEE Communications Letters.

Recently, he received three prestigious awards: The 2021 IEEE ComSoc RCC Technical Recognition Award, the 2018 IEEE ComSoc SPCE Technical Recognition Award and the 2022 Humboldt Research Award from Alexander von Humboldt Foundation. Dr. Karagiannidis is one of the highly-cited authors across all areas of Electrical Engineering, recognized from Clarivate Analytics as Web-of-Science Highly-Cited Researcher in the nine consecutive years 2015-2023.

Sheng Chen (IEEE Life Fellow) received his BEng degree from the East China Petroleum Institute, Dongying, China, in 1982, and his PhD degree from the City University, London, in 1986, both in control engineering. In 2005, he was awarded the higher doctoral degree, Doctor of Sciences (DSc), from the University of Southampton, Southampton, UK. From 1986 to 1999, He held research and academic appointments at the Universities of Sheffield, Edinburgh and Portsmouth, all in UK. Since 1999, he has been with the School of Electronics and Computer Science, the University of Southampton, UK, where he holds the post of Professor in Intelligent Systems and Signal Processing. Dr Chen's research interests include adaptive signal processing, wireless communications, modeling and identification of nonlinear systems, neural network and machine learning, evolutionary computation methods and optimization. He has published over 700 research papers. Professor Chen has 20,000+ Web of Science citations with h-index 61 and 39,000+ Google Scholar citations with h-index 83. Dr. Chen is a Fellow of the United Kingdom Royal Academy of Engineering, a Fellow of Asia-Pacific Artificial Intelligence Association and a Fellow of IET. He is one of the original ISI highly cited researchers in engineering (March 2004).