Compression Ratio Learning and Semantic Communications for Video Imaging

Bowen Zhang, *Graduate Student Member, IEEE*, Zhijin Qin, *Member, IEEE*, and Geoffrey Ye Li, *Fellow, IEEE*

***Abstract*—It is crucial to improve data acquisition and transmission efficiency for mobile robots with limited power, memory, and bandwidth resources. For efficient data acquisition, a novel video compressed-sensing system with spatially-variant compression ratios is designed, which offers high imaging quality with low sampling rates; To improve data transmission efficiency, semantic communication is leveraged to reduce bandwidth requirement, which provides high image recovery quality with low transmission rates. In particular, we focus on the trade-off between rate and quality. To address the challenge, we use neural networks to decide the optimal rate allocation policy for given quality requirements. Due to the non-differentiable issue of rate, we train the networks by policy-gradient-based reinforcement learning. Numerical results show the superiority of the proposed methods over the existing baselines.**

***Index Terms*—Compressed sensing, rate-distortion, reinforcement learning, semantic communications.**

I. Introduction

VIDEO imaging systems have become ubiquitous in robotic systems, enabling mobile robots to perceive the environment, infer their status, and make intelligent decisions. The data volume from a high-resolution imaging system can be pretty large, making it challenging for mobile robots to store sensed data locally or share it wirelessly for human-robot interaction. To address this issue, snapshot compressive imaging (SCI) [1], [2], [3], which captures a set of consecutive video frames with one single exposure, is promising for realizing sparse measurements and low requirements on memory, bandwidth, and power [4]. Due to the limitations of conventional cameras and communication protocols, data acquisition (i.e., sensing) and transmission efficiency in SCI systems are both restricted. To be specific, the existing SCI systems capture an image by exposing photo-sensitive elements for a fixed exposure time, leading to a fixed temporal compression ratio for all pixels. Using a fixed compression ratio in SCI systems severely limits their ability to record natural scenes in a measurement-efficient way, as natural scenes are usually redundant locally. If pixels within SCI systems can be generated under different compression ratios, a video compressive sensing system can maintain high-quality reconstruction and achieve efficient sensing. However, such a sensing system not only has special requirements on the hardware but also requires a pixel-wise compression ratio assignment policy, which depends on both the read/shot noise levels and the object/camera motions.

Fortunately, programmable sensors or focal-plane sensor-processors [5], [6], [7] can vary compression ratios spatially through pixel-level control of the exposure time and readout operations. Recent works in deep optics, on the other hand, demonstrate the superiority of jointly learning optic parameters and image processing methods for applications in high dynamical range (HDR) imaging [8], video compressive sensing [9], [10],[1] and motion deblurring [11]. Inspired by these pioneering works, we propose a SCI system with pixel-wise compression ratios. We focus on minimizing the average compression ratio to reduce the data volume for an imaging quality requirement. The developed system gives rise to a necessary trade-off between the number of measurements and imaging quality at a pixel location. To tackle this, we train a ratio allocation network to decide the per-pixel ratio. Concerning rates being non-differentiable, we design a policy gradient [12], [13] reinforcement learning (RL)-based framework to optimize the ratio allocation network.

In addition to the sensing efficiency, the transmission efficiency of SCI systems can also be improved. If raw sensed data from SCI systems is transmitted, generating videos at the receiver side requires a signal recovery process. The existing communication systems are designed independently from the signal recovery process and treat all data equally. As different data contribute differently to the signal recovery process, bandwidth resources should be spent more on essential data. To address these challenges, semantic communications [14], [15], [16], [17] are promising solutions. The deep optics and video

Manuscript received 2 October 2023; revised 12 March 2024; accepted 7 May 2024. Date of publication 27 May 2024; date of current version 2 September 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2023YFB2904300, and in part by the National Natural Science Foundation of China (NSFC) under Grant 62293484. The guest editor coordinating the review of this manuscript and approving it for publication was Dr. Wenbo Ding. *(Corresponding author: Zhijin Qin.)*

Bowen Zhang and Geoffrey Ye Li are with the Department of Electrical and Electronic Engineering, Imperial College London, SW7 2AZ London, U.K. (e-mail: k.zhang21@imperial.ac.uk; geoffrey.li@imperial.ac.uk).

Zhijin Qin is with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China, and with the State Key Laboratory of Space Network and Communications, Beijing 100084, China, and also with the Beijing National Research Center for Information Science and Technology, Beijing 100084, China (e-mail: qinzhijin@tsinghua.edu.cn).

Digital Object Identifier 10.1109/JSTSP.2024.3405853

[1] In these deep optic-based works on video compressive sensing, the coded aperture or the exposure time for generating a snapshot is optimized but the compression ratio (i.e. the number of measurements over a time window) is fixed. Unlike these works, we focus on adjusting readout frequencies to achieve pixel-wise compression ratio.

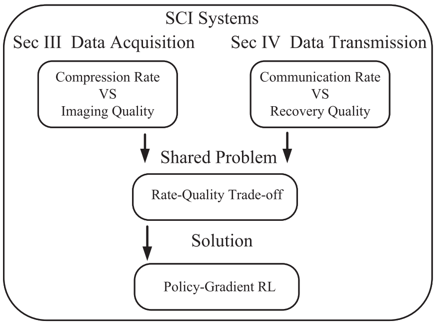


Fig. 1. The main points of this work.

reconstruction networks define special message generation and interpretation processes among transceivers, called semantic encoders and decoders in semantic communications. Semantic communication systems are optimized to ensure the interpretations of the messages through semantic decoders are correct rather than the delivery of raw sensor data itself. To realize semantic communications, one potential way is to introduce task-aware compression [18] for SCI systems and design the compression methods under the guidance of the video reconstruction process. However, the channel coding is still designed separately in this way, while earlier studies on the joint source and channel coding (JSCC) [19] have demonstrated the benefits of co-designing the source and channel coding processes. In this work, we propose an end-to-end semantic communication framework for SCI systems. We focus on minimising the average data transmission rate to reduce the bandwidth for a video quality requirement. The developed system gives rise to an important trade-off between the number of modulated symbols and the video recovery quality for a natural scene. To tackle this, we propose to train a rate allocation network (RAN) to decide the per-scene transmission rate. The RAN is optimized by policy gradient reinforcement learning (RL).

In both data acquisition and transmission, the key is to address the rate-quality trade-off, that is, to characterize the Pareto frontier. As such, we propose a unified framework based on policy-gradient RL. Our contributions can be summarized as follows:

- We introduce a RL-based method for adjusting the parameters in programmable sensors, which differs from the existing methods based on differentiable models or functions [11].
- We build a novel video compressive sensing system with spatially-variant compression ratios, where the ratio allocation policy is learned through an explicit rate-distortion function.
- We introduce a RL-based method for the explicit trade-off between transmission rates and task accuracy in semantic communications, where the rate allocation policy is trained jointly with coding modules.
- We propose a semantic communication system for SCI systems, realizing the co-design of deep optics, data compression, channel coding, and video reconstruction.

For clarity, the primary points of this paper are summarized in Fig. 1.

II. Related Works

In this section, we introduce the related works in deep optical imaging and semantic communications. We also highlight the differences between our work and existing ones.

A. Deep Optics

Recently, many approaches for end-to-end optimization of optics and image processing by machine learning have been developed. For example, Metzler *et. al.* have proposed to jointly learn the point spread function and the reconstruction algorithm to get improved performance in single-shot HDR imaging [8]. Martel *et. al.* have optimized the shuttering functions with the reconstruction algorithm end-to-end for HDR imaging and video compressed sensing [9]. Nguyen *et. al.* have further proposed to learn the spatially-varying exposure time for motion deblurring [11]. In coded apertures (CAs)-based SCI, Vargas *et. al.* have designed time-varying CAs and spatially varying pixel shutters by learning. Their methods outperform the existing SCI systems in compressive light field and hyperspectral imaging [20]. Bacca *et. al.* have introduced different learnable regularizers for CA designs to satisfy special sensing requirements, such as the transmittance constraint, the compression ratio, and the correlation among measurements [21]. Concerning compression ratio learning, Bacca *et. al.* [21] focus on the image-scale compression rate in a multi-shot system. In this work, the adaptive compression ratio is achieved at the pixel level in a SCI system.

Despite the fast development, previous works design differential models and functions for end-to-end optimization, which requires the measurements to change continuously with the sensing parameters. Also, a lot of human effort is required in defining these functions [11], [21]. To address non-differentiable issues and free people from handcrafted designs, we propose a policy-gradient-based RL framework.

B. Semantic Communications

Semantic communications have recently attracted wide attention as an effective data transmission method. The data rate reduction is achieved by jointly optimizing the communication components with the data interpretation processes at the receiver. The interpretation can be either the human's semantic understanding process [22] or the machine's task execution process [23], [24]. The latter case is similar to the earlier task-oriented communications [25], [26] but semantic communications focus more on the tasks with high-dimensional output. In semantic communications, a key component is to adjust the transmission rates with the contents in source data and the interpretation processes. The research in this field can be divided into three categories: entropy-based, sparsity-based, and mask-based. For entropy-based methods, the communication cost is assigned proportionally to the entropy of the source in the feature space [27], where the amount of entropy is adjusted by a rate-distortion function proposed in deep compression methods [28]. The sparsity-based methods adjust the communication costs by imposing some sparse regularizers to the data, as the

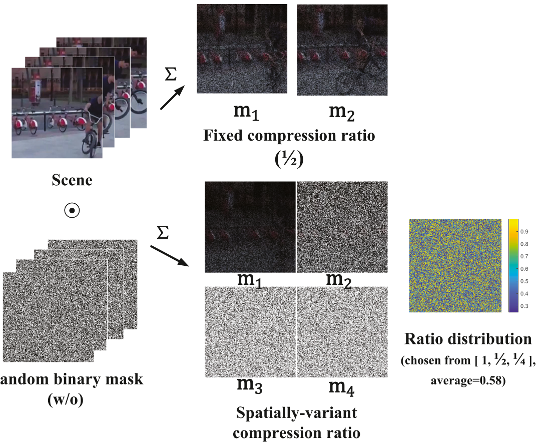


Fig. 2. The illustration of spatially-variant compression ratios in video compressed-sensing.

one based on information bottleneck theory [25]. Mask-based methods directly adjust the data length by deleting unwanted data with a binary mask [15], [29]. These mask-based works use differentiable models to enable the joint training of the mask generation networks and the transceivers. As an improvement, we train the mask generation network by policy-gradient RL. Note that although Q-learning-based RL has been used in [30] to decide the bandwidth for different video frames, a fixed amount of communication costs for different video clips has been assumed in ref [30], while we adjust communication costs for different video clips according to their contents; therefore, we consider a larger action space and our reward is a rate-distortion function.

III. Video Compressed Sensing With Spatially-Variant Ratios

As shown in Fig. 2, the principle of the proposed video compressed sensing system is to vary compression ratios spatially. In SCI systems with a fixed compression ratio, a fixed number of measurements will be generated for all spatial locations. In the proposed system with spatially-variant ratios, some locations will have more measurements than others. For example, given four video frames, two snapshots (m_1, m_2) will be captured in the current SCI systems with a $1/2$ ratio. By contrast, the proposed systems will generate one result with dense measurements (m_1) and three other results with sparse measurements (m_2, m_3, m_4). Specifically, spatial locations with a $1/4$ ratio will only have measurements on m_1 after one readout operation. Locations with a 1 ratio will have values on all four results after four readout operations. The ratio map should be jointly designed with the optics and video reconstruction algorithms to improve the sensing efficiency. In this section, we will introduce the details of the proposed system, including the forward model and the training losses.

A. System Overview

We show the overall pipeline of the proposed sensing system in Fig. 3. Denote H and W as the height and width of video frames. We first generate a compression ratio map, $\boldsymbol{M} \in \mathcal{R}^{H \times W \times 1}$, from a small trainable matrix using a ratio generation network with one convolution layer, three residual blocks, and one transposed convolution layer. The spatial size of the trainable matrix is set to 1/8 of the ratio map. We then simulate the capture of a scene, $\boldsymbol{S} \in \mathcal{R}^{H \times W \times T}$, with a programmable sensor using $\boldsymbol{M}$ and possibly an extra random binary mask, $\boldsymbol{B} \in \mathcal{R}^{H \times W \times T}$, which is referred to as coded apertures in earlier works [4], [20]. Sensed data, I, stacked with M, is then fed into a video reconstruction network to produce a reconstructed video, $\hat{\boldsymbol{V}} \in \mathcal{R}^{H \times W \times T}$. Two training losses ($\mathcal{L}_1$, $\mathcal{L}_2$) are designed to update the learnable matrix, the ratio generation network, and the video reconstruction network to improve the reconstruction quality while at the same time reduce the average compression ratio.

B. Compression Ratio Generation

We consider five kinds of compression ratios for a T-frame long video clip, i.e., 0, $1/T$, $2/T$, $4/T$, and $8/T$. Each pixel in $\boldsymbol{M}$ takes one of the five discrete values. To decide $\boldsymbol{M}$, we design the feature maps generated by the ratio generation network to have five channels. After applying a Softmax function to the channel dimension of the feature maps, each channel indicates the discrete possibility of taking the corresponding ratio. Denote the feature maps representing the discrete probability distributions as $\boldsymbol{P} \in \mathcal{R}^{H \times W \times 5}$. The final ratio map is sampled using $\boldsymbol{P}$.

C. Sensor Forward Model With Varying Ratios

The learned ratio map will guide the behaviour of the programmable sensor. In our systems, we consider that each measurement is read out after an exposure ends and before a new round of exposure begins. We also assume the generated "signal" electrons are cleared immediately after each readout operation; therefore, the signal will integrate from 0 in each exposure. The exposure time and readout operations under different ratios are summarised in Fig. 4. From the figure, when the ratio is $2/T$, each exposure lasts $T/2$ frame time, generating 2 measurements. The measurement matrix when the ratio is $2/T$, $\boldsymbol{A}_{2/T} \in \mathcal{R}^{2 \times T}$, can be written as,

$$\boldsymbol{A}_{2/T} = \frac{1}{T} \begin{bmatrix} 1 & \cdots & 1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & \cdots & 0 & 0 & 1 & \cdots & 1 & 1 \end{bmatrix}. \tag{1}$$

The first row indicates the first measurement is the integrated signal from frame time 1 to $T/2$ while the second row indicates the second measurement from time $T/2 + 1$ to T. Generally speaking, for a specific ratio, r, the equivalent measurement matrix, $\boldsymbol{A}_r$, is of size (rT, T), with each row representing one measurement base. The i-th ($i = 1, 2, .., rT$) row, $[\boldsymbol{A}_r]_{i:}$, is a sparse vector,

$$[A_r]_{ij} = \begin{cases} 1/T, & \text{if } j \in [(i-1)/r, i/r] \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

With the measurement matrices, the sensing process at the spatial location p with the r_p ratio is,

$$\boldsymbol{I}_p = \mathcal{U}(\boldsymbol{A}_{r_p} \boldsymbol{S}_p), \tag{3}$$

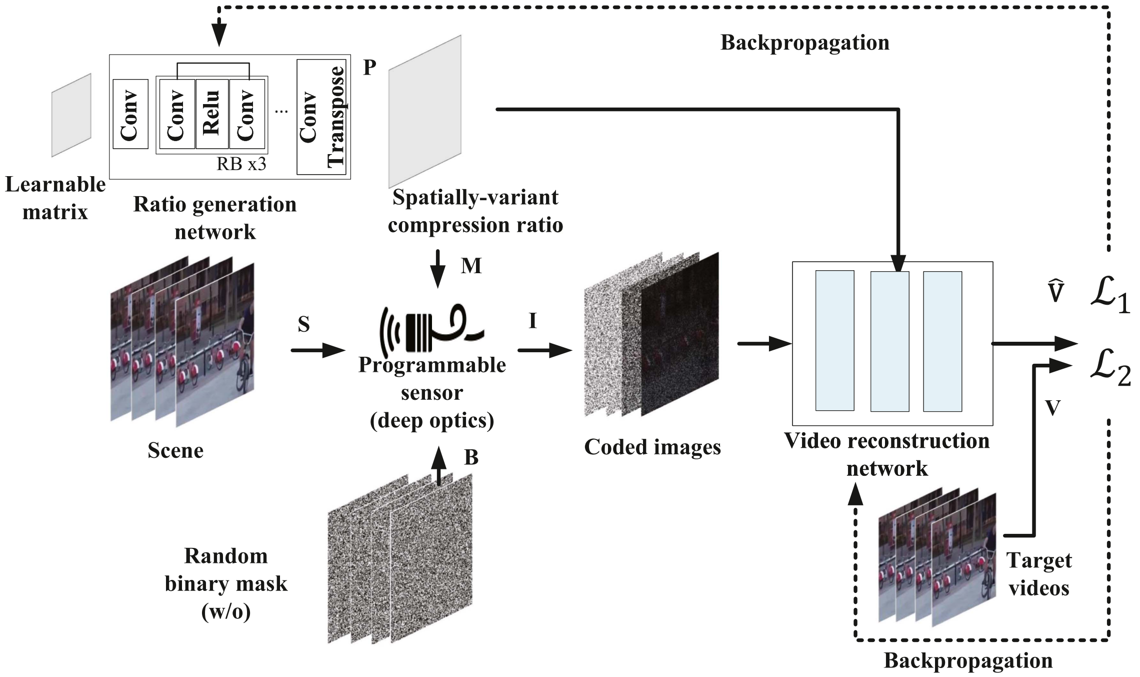


Fig. 3. The overview architecture of the proposed video compressed sensing system with spatially-variant ratios.

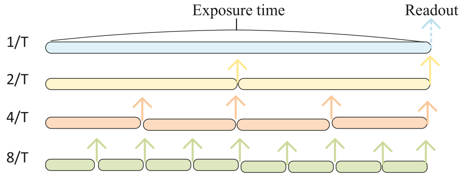


Fig. 4. The exposure time and readout operations in different compression ratios.

where $\boldsymbol{S}_p \in \mathcal{R}^{T\times 1}$ is the signal at the spatial location p, $\boldsymbol{A}_{r_p}\boldsymbol{S}_p$ denotes the integration process of the incident irradiance over time T. In our experiments, $\boldsymbol{S}$ is normalized into [0,1], therefore, $\boldsymbol{A}_{r_p}\boldsymbol{S}_p$ is also in [0,1]. $\mathcal{U}(\cdot)$ represents the camera exposure function capturing the noise effects of the camera [11], and $\boldsymbol{I}_p \in \mathcal{R}^{r_p T\times 1}$ is sensed data at the location p.

If the random binary masks, $\boldsymbol{B} \in \{0,1\}^{H\times W\times T}$, are used to modulate the signal [20], the sensing process can be modelled as,

$$\boldsymbol{I}_p = \mathcal{U}(\boldsymbol{A}_{r_p}(\boldsymbol{S}_p \cdot \boldsymbol{B}_p)), \tag{4}$$

where $\cdot$ represents element-wise multiplication and $\boldsymbol{B}_p \in \mathcal{R}^{T\times 1}$ is the mask at the spatial location p.

Following [11], we consider two kinds of noises, i.e., shot noise $n_s \in \mathcal{N}(0, \sigma_s)$ and read noise $n_r \in \mathcal{N}(0, \sigma_r)$ in the camera exposure function. The level of shot noise is proportional to the signal electrons' strength. For a signal e in [0,1], $\sigma_s = \sqrt{e}\sigma_{ss}$, where σ_{ss} is independent of signals. By contrast, the read noise level is fixed for a camera, depending on the photon flux. We use the same settings for σ_{ss} and σ_r as [11], that is, $\sigma_{ss} = 4.95 \times 10^{-3}$ and $\sigma_r = 7.25 \times 10^{-3}$. With the definition of noises, $\mathcal{U}(e) = e + n_s + n_r$.

From Fig. 4, increasing r will decrease the exposure time; therefore reducing the signal strength of each measurement. Since the signal-to-noise ratio (SNR) regarding the read noise is e/σ_{ss}^2 and the SNR regarding the shot noise is e^2/σ_r^2, reducing the signal strength also reduces the SNR; therefore, the SNR of measurements decreases when r is increased, which means, although more measurements are available as r increases, the imaging quality may not increase at the same time, as the SNR of each measurement is decreasing. This phenomenon makes it challenging to find the optimal $\boldsymbol{M}$.

D. Video Reconstruction Network

After obtaining I, the next step is to reconstruct the targeted video using a deep neural network. The overall architecture of the proposed video reconstruction network is shown in Fig. 5. Our network consists of three components: an initial reconstruction stage (IR), a fusion network (FN), and a deep reconstruction network (DRN) built based on the single-stage spectral-wise transformers (SST) proposed in [31]. The IR and FN are introduced to mitigate the influence of spatially-variant $\boldsymbol{M}$ and $\boldsymbol{B}$. Specifically, if $\boldsymbol{B}$ is not considered, we get the initial reconstruction, $\hat{\boldsymbol{V}}_{\boldsymbol{0}} \in \mathcal{R}^{H\times W\times T}$, in IR by considering (3),

$$\hat{\boldsymbol{V}}_{0_p} = \boldsymbol{A}_{r_p}^T(\boldsymbol{A}_{r_p}\boldsymbol{A}_{r_p}^T)^{-1}(\boldsymbol{I}_p), \tag{5}$$

where $\hat{\boldsymbol{V}}_{0_p} \in \mathcal{R}^{T\times 1}$ denotes the initial reconstruction results at the spatial location p, $(\boldsymbol{A}_{r_p}\boldsymbol{A}_{r_p}^T)^{-1}$ is fortunately a diagonal matrix as defined in Section III-C. By contrast, if $\boldsymbol{B}$ is considered, we first rewrite (4) as,

$$\boldsymbol{I}_p = \mathcal{U}(\hat{\boldsymbol{A}}_{r_p}\boldsymbol{S}_p), \tag{6}$$

where $\hat{\boldsymbol{A}}_{r_p} = \boldsymbol{A}_{r_p}\text{diag}(\boldsymbol{B}_p)$ and $\boldsymbol{B}$ denotes the $T \times T$ diagonal matrix constructed from $\boldsymbol{B}_p$. In this case, some rows of $\hat{\boldsymbol{A}}_{r_p}$

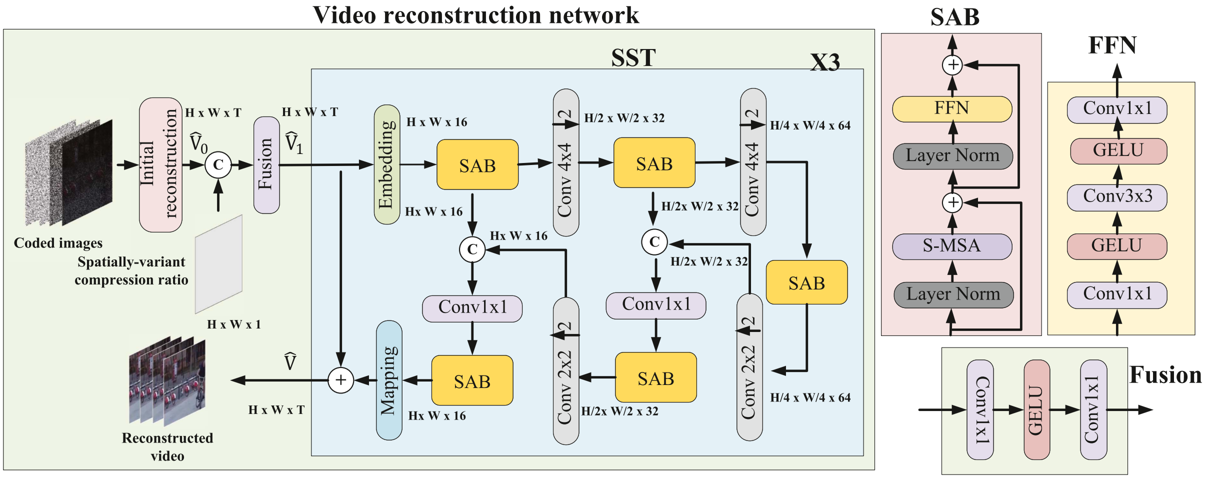


Fig. 5. The architectural details of video reconstruction network.

may be all zero, making $\hat{\boldsymbol{A}}_{r_p}\hat{\boldsymbol{A}}_{r_p}^T$ non-invertible. Since the row vectors of $\hat{\boldsymbol{A}}_{r_p}$ are orthogonal to each other, we reconstruct $\hat{\boldsymbol{V}}_{\mathbf{0}}$ by using the non-zero row vectors of $\hat{\boldsymbol{A}}_{r_p}$ independently,

$$\hat{\boldsymbol{V}}_{0_p} = \sum_{\substack{i=1 \\ \left[\hat{\boldsymbol{A}}_{r_p}\right]_{i:} \neq \mathbf{0}_{1\times T}}}^{rT} \left[\hat{\boldsymbol{A}}_{r_p}\right]_{i:}^T \left(\left[\hat{\boldsymbol{A}}_{r_p}\right]_{i:} \left[\hat{\boldsymbol{A}}_{r_p}\right]_{i:}^T\right)^{-1} [\boldsymbol{I}_p]_{i:} + \mathbf{0}_{T\times 1}. \tag{7}$$

Following the coarse-to-fine criteria, we further use a shallow FN to deal with the spatially-variant coded exposures and ratios, which takes $\hat{\boldsymbol{V}}_{\mathbf{0_p}}$, $\boldsymbol{M}$, and possibly $\boldsymbol{B}$ as inputs and outputs the second-level reconstruction results, $\hat{\boldsymbol{V}}_{\mathbf{1}} \in \mathcal{R}^{H\times W\times T}$. The architecture of FN is shown in Fig. 5.

Next, the DRN takes $\hat{\boldsymbol{V}}_{\mathbf{1}}$ as the input and reconstructs the final videos, $\hat{\boldsymbol{V}}$, by cascading three SSTs [31]. Each SST follows the design of U-net [32] with an encoder, a bottleneck, and a decoder. The embedding and mapping blocks are convolutional layers (conv) with 3 × 3 kernels. The feature maps in the encoder sequentially pass through one spectral-wise attention block (SAB), one conv with stride 2 (for downsampling), one SAB, and one conv with stride 2. The bottleneck is one SAB. The decoder has a symmetrical architecture to the encoder. Following the spirit of U-Net, the skip connections are used for feature aggregation between the encoder and decoder to alleviate the information loss from the downsampling operation. The basic unit of SST is SAB, whose architecture is also shown in Fig. 5. It has one feed-forward network (FFN), one spectral-wise multi-head self-attention (S-MSA), and two layer-normalization. Unlike the original MSA that calculates the self-attention along the spatial dimension, S-MSA regards each feature map as a token and calculates the self-attention along the channel dimension, making it computationally effective. More details of S-MSA are explained in [31].[2]

E. Training Losses

In our system, the optimal ratio map should have a smaller average ratio to reduce data volume while at the same time offer high video reconstruction quality; therefore, the problem can be formulated as a rate-quality (or rate-distortion) trade-off. As shown in Fig. 3, for the explicit trade-off, we introduce a $\mathcal{L}_1$ based on the rate-distortion theory. A policy-gradient RL framework is used as "rate" is discrete and non-differentiable. Besides, the video generation network should ensure the generated videos are close to the true video (i.e., true label). A $\mathcal{L}_2$ based on supervised learning is designed to tackle this. To sum up, we use $\mathcal{L}_1$ and $\mathcal{L}_2$ to train the ratio generation parts and the video reconstruction network, respectively.

Specifically, the learnable matrix and the ratio generation network are trained by RL. In our system, each spatial location in P is regarded as an agent, and its action space is the available compression ratio. We define the reward of each location, $Q_p(r_p)$, under action r_p according to the rate-distortion trade-off theory,

$$Q_p(r_p) = \log(1/||\boldsymbol{V}_p - \hat{\boldsymbol{V}}_p||^2) - \lambda r_p T, \tag{8}$$

where $\boldsymbol{V}$ denotes target videos, $||\boldsymbol{V}_p - \hat{\boldsymbol{V}}_p||^2$ denotes the mean-squared error (MSE) (i.e., distortion) of the reconstructed videos at spatial location p, r_pT denotes the number of measurements employed at location p, which can also be viewed as the compression rate, λ is an introduced parameter for the rate-distortion trade-off. Increasing λ will penalize more on the compression rate, leading to a smaller average compression ratio. With (8), we can find the rate achieving the best trade-off between the compression rate and the video reconstruction quality. Although $||\boldsymbol{V}_p - \hat{\boldsymbol{V}}_p||^2$ can only approximately evaluate the effect of action r_p on the $\boldsymbol{V}_{\boldsymbol{p}}$ as the DRN will aggregate information from neighbouring pixels, it is still the most direct way to evaluate the action r_p.

At the same time, the expected reward, J_p, is the value function for spatial location p, where the expectation is w.r.t. r_p with probability P_p. Note that, Min$\mathcal{L}_1 = $ Max $\sum_p J_p$. Following [13], we can approximate the gradient of the J_p to parameters

[2] As the network design of DRN is not the paper's main focus, we omit some details here for page limits. Interesting readers can refer to the original paper for more details.

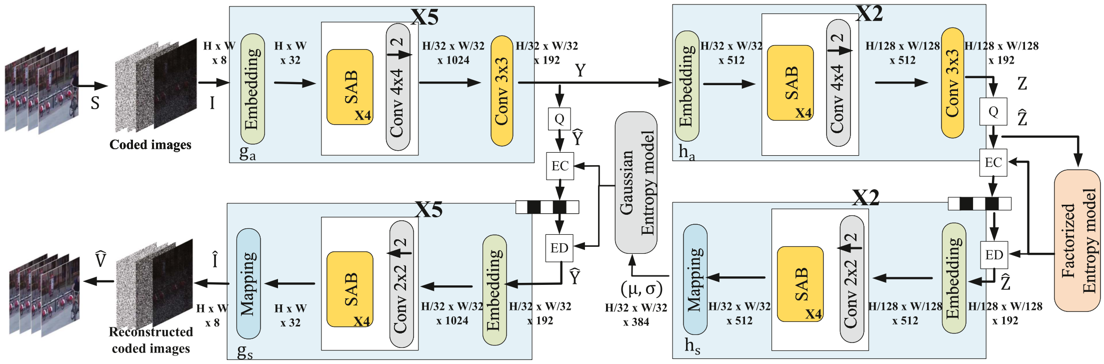


Fig. 6. The overall framework of task-aware compression methods.

θ in ratio generation parts with samples generated from P_p,

$$\nabla_\theta J_p = \mathbb{E}_{r_p} \nabla_\theta \log P_p(r_p) \times Q_p(r_p), \tag{9}$$

where $P_p(r_p)$ denotes the probability of the chosen action r_p from the distribution P_p. Note that the update of θ is based on the average value of $\nabla_\theta J_p$ from different p. Earlier works have proved the global convergence of policy gradient RL in multi-agency situations [33].

On the other hand, the video reconstruction network is trained in a supervised way based on the MSE between V and $\hat{V}$,

$$\mathcal{L}_2 = ||V - \hat{V}||^2, \tag{10}$$

IV. Efficient Data Transmission in SCI Systems

As discussed above, the existing communication systems focusing on reproducing sensed data are sub-optimal. It is better to design the compression methods concerning the video reconstruction network (i.e., task-aware compression [18]). After compressing the original sensor data into compact bit streams, the transmitter will add parity bits and modulate the streams for robust transmission over unreliable channels using off-the-shelf channel coding, such as low-density parity-check (LDPC) code. The communication cost is the number of modulated symbols after channel coding and modulation.

Nevertheless, there is an increasing belief in the communication community that the classic framework based on the Shannon separation theory needs to be upgraded for joint designs [26]. For semantic communication systems, jointly optimizing the channel coding and modulation with the other components may lead to better video reconstruction quality with fewer communication resources.

In this section, we will design semantic communication frameworks for the proposed video compressed-sensing systems based on task-aware compression and semantic communications with joint designs. Note that RL-based approaches are not used in task-ware compression.

A. Task-Aware Compression

1) Architecture: The overall architecture based on deep compression methods is shown in Fig. 6. It is mainly based on the architecture used in [18] by substituting the basic feature extraction unit from convolutional blocks to SABs. Some other changes are also made for simplicity.[3] For sensed data $\boldsymbol{I}$, we first reshape its dimension to $H \times W \times 8$, as each spatial location has at most eight measurements. Zero-padding is used for locations with fewer measurements. An encoder, $g_a(\cdot)$, then takes the reshaped sensed data as inputs and generates a latent representation $\boldsymbol{Y} \in \mathcal{R}^{\frac{H}{32} \times \frac{W}{32} \times 192}$, which is given by, $\boldsymbol{Y} = g_a(I)$. The features inside the encoder subsequently go through one embedding layer (conv3x3), five SAB×4-downsampling×1 pairs, and one mapping layer (conv3x3). Each downsampling operation will decrease the spatial size of the features by $\frac{1}{4}$ but double the channel dimension. $\boldsymbol{Y}$ is then quantized to $\hat{\boldsymbol{Y}}$ by a quantizer. Next, a hyper-encoder, $h_a(\cdot)$, takes $\boldsymbol{Y}$ as input and generates an image-specific side information $\boldsymbol{Z} \in \mathcal{R}^{\frac{H}{128} \times \frac{W}{128} \times 192}$ via $\boldsymbol{Z} = h_a(\boldsymbol{Y})$. The channel dimension of the features remains constant in h_a. Then, the quantized side information, $\hat{\boldsymbol{Z}} = Q(\boldsymbol{Z})$, is saved as a lossless bitstream through a factorized entropy model and entropy coding [28]. After that, $\hat{\boldsymbol{Z}}$ is forwarded to the hyper-decoder, $h_s(\cdot)$, to draw the parameters (μ, σ) of a Gaussian entropy model [28], which approximates the distribution of $\hat{\boldsymbol{Y}}$ and is used to save $\hat{\boldsymbol{Y}}$ as a lossless bitstream. For reconstructing $\boldsymbol{I}$, a decoder, $g_s(\cdot)$, operates on $\hat{\boldsymbol{Y}}$ and generates $\boldsymbol{I}$ by $\boldsymbol{I} = g_s(\hat{\boldsymbol{Y}})$. At last, the reconstructed $\hat{\boldsymbol{I}}$ is used for video reconstruction. Note that $g_s(\cdot)$ and $g_a(\cdot)$, $h_s(\cdot)$ and $h_a(\cdot)$ have symmetric architectures, respectively. More details of deep compression methods are explained in [18], [28]. [4]

2) Training Losses: The goal of task-aware compression is to minimize the length of the bitstreams required to transmit $\boldsymbol{I}$ without increasing the distortion between $\boldsymbol{V}$ and $\hat{\boldsymbol{V}}$, which equals the trade-off between the source coding rate (i.e. proportional to the communication cost) and the distortion. This objective raises an optimization problem of minimizing $||\boldsymbol{V} - \hat{\boldsymbol{V}}||^2 + \beta(-\log_2 P_{\hat{\boldsymbol{Y}}} - \log_2 P_{\hat{\boldsymbol{Z}}})$, where $||\boldsymbol{V} - \hat{\boldsymbol{V}}||^{\boldsymbol{2}}$ denotes the distortion of reconstructed videos, $-\log_2 P_{\hat{\boldsymbol{Y}}} - \log_2 P_{\hat{\boldsymbol{Z}}}$ represents

[3] As the learned compression ratio is universal to all scenes, the spatial importance of I should also be independent of the contents of I; therefore, the side links conditioned on a quality map used in [18] can be safely removed. Also, instead of estimating the quality map by a task loss function, we directly change the distortion metric in the training loss of [18] to task loss, which should be more precise in describing the spatial importance.

[4] As the deep compression is a well-studied field, we omit some details here for page limits. Interesting readers can refer to the original papers for more details.

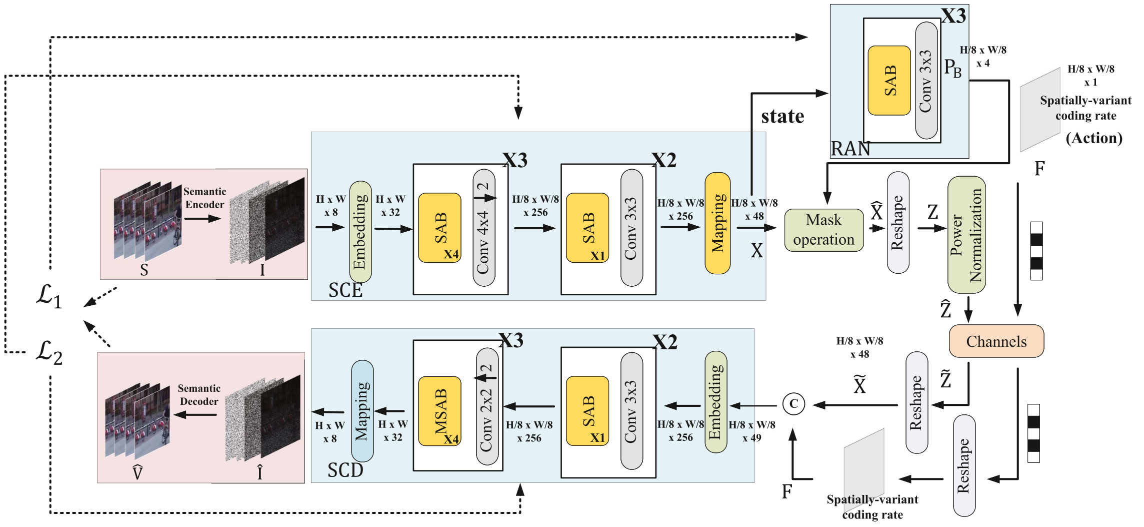


Fig. 7. The overall framework of semantic communications with joint designs.

the entropy of $\hat{\boldsymbol{Y}}$ and $\hat{\boldsymbol{Z}}$, which equals to the number of bits used to encode $\hat{\boldsymbol{Y}}$ and $\hat{\boldsymbol{Z}}$ by entropy coding, and β is an introduced trade-off parameter between coding rate and video distortion. Increasing β will reduce the average bit length (i.e. the communication cost) but increase the distortion of reconstructed videos.

3) Communication Costs: After obtaining the data compression systems with different bit lengths by changing β, we can then estimate the required number of modulated symbols to transmit these bit streams under different channel conditions. Specifically, the number of modulated symbols (in complex numbers) used for transmitting a bitstream depends on the implemented channel coding rate r_c (*e.g.* 1/3, 1/2, 2/3) and modulation order r_m (*e.g.* 4, 16, 64). Suppose the length of the bitstream is l_b, the length of modulated symbols can be calculated as $l_s = \frac{l_b}{r_c \log_2(r_M)}$. Another way to evaluate the communication costs in additive white Gaussian noise (AWGN) channel is to use the Shannon capacity theorem, $l_s = \frac{l_b}{\log_2(1+snr)}$ if the channel signal-to-noise (SNR) ratio is snr.

B. Semantic Communications With Joint Design

Although the aforementioned deep compression method enables the co-design of deep optics, reconstruction algorithms, and source coding, the widely-used entropy coding method [18], [28] prohibits the co-design of communication-specific components with deep compression methods. When entropy coding methods are used, each bit from entropy coding methods needs to be transmitted error-freely; otherwise, error propagation will happen in the entropy decoding process. This property leaves the communication systems with no choice but to treat each bit equally and carefully. On the other hand, semantic communications with joint designs allow the effect of channel noise to be considered by end-to-end training.

1) Architecture: As shown in Fig. 7, the framework consists of three components: semantic coders, a semantic-channel encoder (SCE) and decoder (SCD), and a rate allocation network (RAN). Semantic coders define a unique message generation and interpretation method between transceivers based on a shared knowledge base. Semantic-channel coders directly learn the end-to-end mappings between semantic messages and modulated symbols. It is also called the joint source and channel coding in communication systems. The RAN is responsible for controlling the transmission rates. The rate denotes the number of modulated symbols required for each S. Different from deep compression methods that focus on the source coding rate (i.e. bit length) only, the transmission rates in semantic communications depend on the source coding rate, channel coding rate, and modulation order.

Specifically, the deep optic methods are special semantic encoders, which encode a natural scene, $\boldsymbol{S}$, into sensed data, $\boldsymbol{I}$, in a predefined way. The video reconstruction networks, which decode the target video, $\boldsymbol{V}$, from, $\boldsymbol{I}$, are special semantic decoders. Sensed data, $\boldsymbol{I}$, is the semantic message of a scene to be shared between transmitters and receivers. Note that conventional communication systems emphasise the accurate transmission of $\boldsymbol{I}$. With the semantic decoder (defining $\boldsymbol{I} \rightarrow \boldsymbol{V}$), semantic communication systems aim to maximize the quality of $\boldsymbol{V}$ under limited communication costs.

In our framework, the transmitter will use a SCE to encode $\boldsymbol{I}$ into a predefined maximum number of modulated symbols (evaluated in real number), $\boldsymbol{X} \in \mathcal{R}^{\frac{H}{8} \times \frac{W}{8} \times 48}$, from which the rate control techniques will select some symbols to transmit through noisy channels. [5] We can model this process as,

$$\boldsymbol{X} = \mathrm{SCE}(\boldsymbol{I}), \tag{11}$$

The SCE is composed of an embedding layer, three consecutive SAB×4-downsampling×1 pairs, two stacked SAB×1-Conv×1 pairs, and a mapping layer. The size of feature maps after each operation is shown in Fig. 7.

[5] Otherwise, if all the messages are transmitted with the same number of modulated symbols, it disobeys the Shannon source coding theory saying that the minimal possible expected length of codewords should be a function of the entropy of the input word.

To adjust the communication costs according to the semantic contents of in the $\boldsymbol{I}$, the RAN takes $\boldsymbol{X}$ as inputs and generates a spatially-variant coding rate map $\boldsymbol{F} \in \mathcal{R}^{\frac{H}{8} \times \frac{W}{8} \times 1}$. Each element in the spatial location of $\boldsymbol{F}$ indicates how many symbols out of the 48 symbols in X with the same spatial location will be kept for transmission. Specifically, $\boldsymbol{F}$ takes four discrete values $f \in \{1, 2, 3, 4\}$, and only the first $12f$ symbols of $\boldsymbol{X}$ in the channel dimension will be transmitted. The generation of $\boldsymbol{F}$ is similar to the generation of $\boldsymbol{M}$ in Section III-B. Specifically, the RAN generates a feature map $\boldsymbol{P_F} \in \mathcal{R}^{\frac{H}{8} \times \frac{W}{8} \times 4}$ representing the discreet probability distribution of taking each value from $\{1, 2, 3, 4\}$ and a sampling operation is conducted to generate $\boldsymbol{F}$. With $\boldsymbol{F}$, a mask operation is conducted on X to delete unnecessary symbols. The remaining symbols $\hat{\boldsymbol{X}}$ are then reshaped into a vector $\boldsymbol{Z} \in \mathcal{R}^{n_z \times 1}$, where n_z denotes the number of remaining symbols. Next, power normalization is applied to $\boldsymbol{Z}$ to satisfy power constraints,

$$\hat{\boldsymbol{Z}} = \sqrt{n_z}\boldsymbol{Z}/||\boldsymbol{Z}||^2. \tag{12}$$

Finally, $\hat{Z}$ is directly transmitted through wireless environments. The effect of channel noise on $\hat{Z}$ can be represented as,

$$\tilde{\boldsymbol{Z}} = h\hat{\boldsymbol{Z}} + \boldsymbol{n}, \tag{13}$$

where $h \in \mathcal{CN}(0, 1)$ is multiplicative noise and $\boldsymbol{n} \in \mathcal{N}(0_{n_z \times 1}, \sigma_n I_{n_z \times n_z})$ is additive noise. In the AWGN channel, $h = 1$ and the value of σ_n depends on the current channel SNR, which is defined as $10\log_{10}(1/\sigma_n^2)$ dB. Simultaneously, each element of $\boldsymbol{F}$ can be quantized using 2 bits, and the bitstream from $\boldsymbol{F}$ is of length $HW/32$ bits and also transmitted. This is the side information of how many symbols are deleted for each spatial location, which will be used at the receiver side to reshape the flattened $\tilde{Z}$ back to the desired shape. Different from the transmission of $\hat{Z}$, the bitstream of F is transmitted in an error-free way, as in the deep compression methods, because a minor error in $\boldsymbol{F}$ will make it impossible to reshape $\tilde{\boldsymbol{Z}}$ correctly.

At the receiver side, $\tilde{\boldsymbol{X}}$ and $\boldsymbol{F}$ are reshaped from $\tilde{\boldsymbol{Z}}$ and bitstreams, respectively. After concatenation, they are fed into a SCD with a symmetric architecture to the SCE. This process can be modelled as,

$$\tilde{\boldsymbol{I}} = \mathrm{SCD}(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{F}}), \tag{14}$$

The recovered sensor data $\tilde{\boldsymbol{I}}$ is used to generate videos using the video generation network.

2) Training Losses: In our system, the optimal F should have a lower average value to reduce the bandwidth requirement while at the same time offer high video reconstruction quality; therefore, the problem can be formulated as a rate-quality (or rate-distortion) trade-off. Different from Section III-E, each scene has a different rate map depending on its content. Similarly, we use $\mathcal{L}_1$ and $\mathcal{L}_2$ to train the RAN and SCE/SCD, respectively. Specifically, the RAN is trained based on policy gradients RL, where $\boldsymbol{X}$ is regarded as the state, $\boldsymbol{F}$ describes the actions, and each spatial location of $\boldsymbol{F}$ is an agent. Note that the spatial size of $\hat{\boldsymbol{V}}$ is 8 times larger than that of $\boldsymbol{F}$ and $\boldsymbol{X}$ so the action taken at each spatial location of $\boldsymbol{F}$ will have a strong effect on the reconstruction quality of an 8×8 area of $\hat{\boldsymbol{V}}$. Considering this, we first define $\boldsymbol{U} = (\boldsymbol{V} - \hat{\boldsymbol{V}})^2$ and apply an 8×8 average pooling to $\boldsymbol{U}$, obtaining $\tilde{\boldsymbol{U}} \in \mathcal{R}^{\frac{H}{8} \times \frac{W}{8} \times 1}$. We then define the reward of location p, $Q_F^p(f_p)$, under action f_p as,

$$Q_F^p(f_p) = \log(1/\tilde{U}_p) - \mu f_p, \tag{15}$$

where $\log(1/\tilde{U}_p)$ denotes the video reconstruction quality at spatial location p, f_p is the transmission rate at p, μ is an introduced parameter for rate-quality trade-off. Increasing μ will penalize more on the transmission rate, leading to fewer modulated symbols to be transmitted. We adjust the average communication costs by tuning μ.

At the same time, the expected reward, J_F^p, is the value function for spatial location p, $\mathrm{Min}\mathcal{L}_1 = \mathrm{Max}\sum_p J_F^p$, and the gradient of J_F^p to the parameters δ in RAN can be approximated as,

$$\nabla_\delta J_F^p = \mathbb{E}_f \nabla_\delta \log P_F^p(f_p) \times Q_F^p(f_p), \tag{16}$$

where $P_F^p(f_p)$ is the probability of taking action f_p at the spatial location p. Note that the update of δ is based on the average value of $\nabla_\delta J_F^p$ from different locations.

On the other hand, the SCE and SCD are trained in a supervised way based on the MSE between $\boldsymbol{V}$ and $\hat{\boldsymbol{V}}$,

$$\mathcal{L}_2 = ||\boldsymbol{V} - \hat{\boldsymbol{V}}||^2, \tag{17}$$

3) Communication Costs: The communication costs consist of two parts: the transmission of $\hat{Z}$ and F. As $\hat{Z}$ is of shape $n_z \times 1$, the number of complex modulated symbols used to transmit $\hat{Z}$ can be denoted as $l_s^{(1)} = n_z/2$. On the other hand, the bitstream length from F, l_b, is $\frac{HW}{32}$. If the channel coding rate is r_c and the modulation order is r_m, the length of the modulated symbols for F can be calculated as $l_s^{(2)} = \frac{l_b}{r_c \log_2(r_M)}$. The total communication costs are $l_s = l_s^{(1)} + l_s^{(2)}$.

V. Experiments

In this section, we first demonstrate the superiority of the proposed video compressed sensing system in terms of sampling rate. Based on the developed sensors, we then evaluate the performance of different communication systems regarding communication costs. The PyTorch source code to reproduce all experiments is at https://github.com/Bowen-zhang96/CRL-SemCom-VidCI.

A. Video Imaging With Spatially-Variant Compression Ratios

The following experiments are conducted to evaluate the effectiveness of the method proposed in Section III.

1) Dataset: Following [11], we use the Need for Speed (NfS) dataset [34] to train the network and evaluate its performance. The NfS dataset is collected with significant camera motions and is suitable for representing the scene captured by moving robots' onboard cameras. The dataset consists of 100 videos obtained from the internet, of which 80 are used for training and 20 for testing. Each video is captured at 240 frames per second (fps) with a 1280 × 720 resolution that we centre crop to 256 × 256.

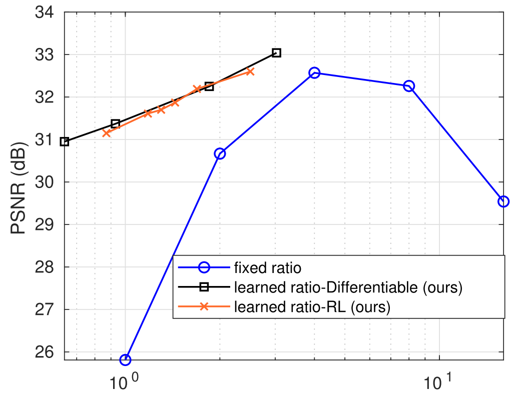


Fig. 8. The performance comparison between learned spatially-variant and fixed ratios methods in video imaging systems without coded aperture.

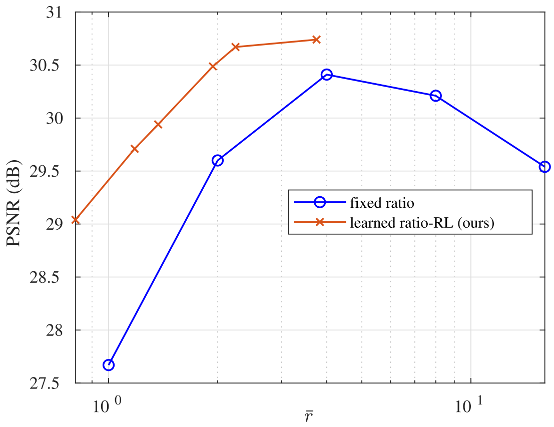
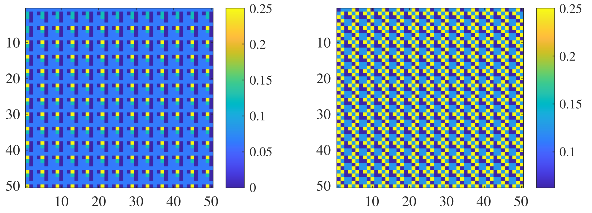


Fig. 9. The performance comparison between learned spatially-variant and fixed ratios methods in video imaging systems with coded aperture.



(a) $\bar{r} = 0.998$ (b) $\bar{r} = 2.31$

Fig. 10. The learned ratio maps in a 50×50 area of an image when B is used.

For each video, we select 80 random 16-frame-long segments within the video; Therefore, $T = 16$ in our experiments. The images are turned into grey images and normalized to be within [0,1]. Our inputs and outputs to the end-to-end model are both 16-frame video segments.

2) Implementation Details: Our model is implemented in PyTorch. The ratio generation parts are trained with the SGD optimizer (momentum $= 0.9$) at a learning rate of 5×10^{-3}. The video reconstruction network is trained with the Adam optimizer at a learning rate of 5×10^{-5}. The training process is as follows: We first fix the ratio at $8/T$ for all spatial locations and train the video reconstruction network for 100 epochs. After that, we gradually increase λ in (8) from 5×10^{-3} to 0.5 and jointly train the ratio generation parts and video reconstruction network. For each λ, we train for about 100 epochs. For baselines with a fixed compression ratio, we set the ratio from $1/T$ to $16/T$ and train the video reconstruction network for 100 epochs in each fixed ratio. Furthermore, we consider both cases where the binary mask, B, is used or not. When B is used, it is randomly initialized and fixed during the experiments.

3) Differentiable-Function-Based Realization: To better understand the performance of the proposed RL-based framework, we also considered learning pixel-wise compression ratio by defining differentiable functions. However, the differentiable models/functions used in existing methods [11], [21] cannot be directly used in our sensor framework as the considered problem is significantly different; therefore, we propose a differentiable-function-based realization by ourselves, where the function is designed in a heuristic way. The implementation details are given in the supplementary material, which can also be found in https://github.com/Bowen-zhang96/CRL-SemCom-VidCI. We omit the details here for the page limit.

4) Results Without Binary Mask: We first compare our methods with its fixed-ratio version when B is not used. We use the peak signal-to-noise ratio (PSNR), $10\log_{10}(1/\mathrm{MSE}(\boldsymbol{V}, \hat{\boldsymbol{V}}))$, as the performance metric, where $\mathrm{MSE}(\boldsymbol{V}, \hat{\boldsymbol{V}}) = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}(V_{ij} - \hat{V}_{ij})^2$. The results are shown in Fig. 8, where $\bar{r} = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W} M_{i,j}$ denotes the average compression ratio. As shown in Fig. 8, the imaging quality increases along with the growth of $\bar{r}$ for both methods until $\bar{r}$ reaches $0.25 (= 4/T)$, where the effects of shot noises and read noises surpass the growth of the number of measurements. From the figure, the proposed method with learned ratios has a significant performance gain over the method with a fixed ratio. For example, when $\bar{r} = 0.0625 (= 1/T)$, the learned-ratio method has nearly 5 dB gain over the fixed-ratio method, demonstrating the superiority of learning a compression ratio map in reducing the data volume. Furthermore, we find learning pixel-wise compression ratios by introducing differentiable models or RL lead to similar performance in the considered settings. This further validates the effectiveness of the proposed method. Note that, compared with differentiable function-based realization, RL-based methods are promising to free people from heuristic designs for end-to-end optimization.

5) Results With Binary Mask: We then consider B and show the performance comparison in Fig. 9. Due to the usage of binary masks, the system performance of the fixed-ratio method increases when $\bar{r} = 0.0625 (= 1/T)$, showing the positive effect of coded apertures. However, as using B will decrease the signal strength, the performance of the fixed-ratio method decreases as $\bar{r}$ increases when compared to the same method without B in Fig. 8. From the figure, the proposed method with a learned ratio still has a steady performance gain over the fixed-ratio method, further proving the effectiveness of compression ratio learning.

6) Learned Compression Ratio Maps: We show the learned ratio maps when B is used in Fig. 10. As discussed above, the maps have only five ratio choices, and their colours are shown in the colour bar. The figure shows that the learned

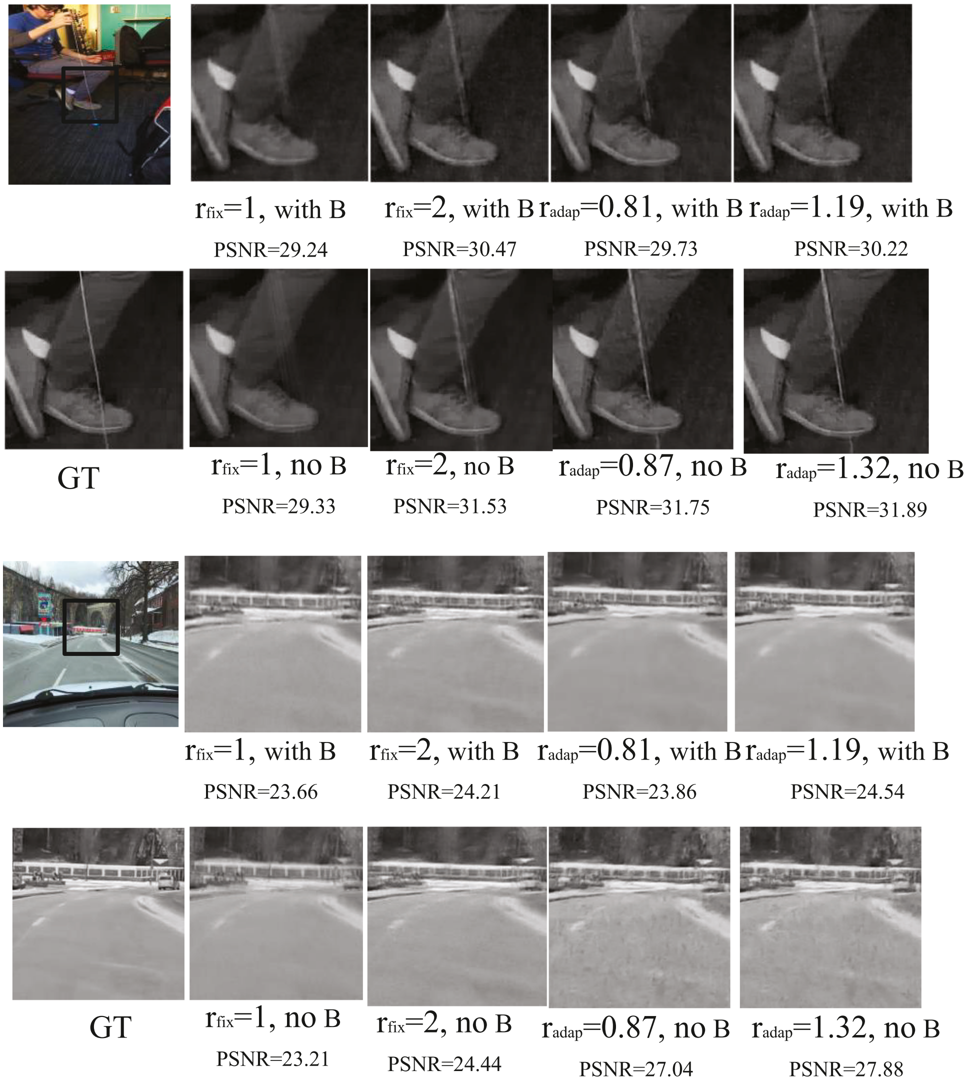


Fig. 11. Examples of restored video frames from different sensing methods with different compression ratios.

ratio maps have fixed patterns for a local area and are a mix of low and high ratios. The patterns are slightly different for different $\bar{r}$.

7) Visual Results: Fig. 11 shows the restored video frames from different methods with different average compression ratios. Each example has two rows. The first rows are the results when $\boldsymbol{B}$ is used, while the second is when $\boldsymbol{B}$ is not used. $r_{fix} = 1$ or $r_{fix} = 2$ denote the fixed-ratio method with $\bar{r} = 1$ and $\bar{r} = 2$, respectively. $r_{adap} = 0.81, 1.19,$ or 1.32 denotes learned-ratio method with $\bar{r} = 0.81, 1.19,$ or 1.32, respectively. From the figure, the frames from the learned-ratio method have more texture details and less artefact.

B. Semantic Communications for Programmable Sensors

In this subsection, we will first evaluate semantic communication frameworks based on task-aware compression and joint design, respectively. After that, we will compare semantic communications with the existing transmission methods.

In the following experiments, we first restore sensor-related network parameters from the pre-trained models in the previous subsection and fix them when training the semantic communication frameworks. We use the sensor network with $r_{avg} = 0.156$ in Fig. 8. The experiments are called fixed-sensing experiments. Next, we jointly train the sensing network with the communication parts, which are called joint-sensing experiments.

1) Channel Condition: We assume the sensor data is transmitted through the AWGN channel with SNR $= 10$ dB.

2) Implementation Details of Different Semantic Communication Frameworks: We now describe the implementations of different semantic communication frameworks in more detail.

- **Task-aware compression plus capacity-achieving channel coding (Compr+Cap):** We first convert the sensor data into bitstreams and then assume the transmission of the bitstreams can reach Shannon channel capacity. In the considered channel condition, $\frac{l_s}{l_b} = 0.289$. Note that it is hard to achieve Shannon capacity in real systems, so its performance can only be regarded as an ideal reference for compression methods. We train the compression network under different β to adjust the average communication costs. During training, we first set $\beta = 10^{-7}$ and gradually decrease β to increase the bitstream length. The network is first trained with the Adam optimizer at a learning rate of 5×10^{-5} for about 50 epochs and then at 5×10^{-6} for another 50 epochs under each β.

- **Task-aware compression plus LDPC plus QAM (Compr+LDPC):** The bitstreams from task-aware compression methods are transmitted through LDPC channel coding and quadrature amplitude modulation (QAM) modulation. Following [15], we should use 16QAM and 2/3 LDPC for AWGN channels with SNR $= 10$ dB. In this case, $\frac{l_s}{l_b} = 0.376$.
- **Semantic communications without RAN (SemCom+noRAN):** In earlier task-oriented communications [26], the RAN is not implemented and all source data is transmitted with the same communication costs. We also delete the RAN in the proposed framework and demonstrate its performance as a reference. To adjust the communication costs, we change the channel dimension of X in (11) from 16 to 48. The network is trained for about 50 epochs at a learning rate of 5×10^{-5} and then for 50 epochs at 5×10^{-6} under each dimension of X.
- **Semantic communications (SemCom):** This is the proposed end-to-end semantic communication framework with rate control. We train the network under different μ in (15) to adjust the average transmission rate. The training consists of two steps: we first set $f = 4$ for all locations and train the SCE/SCD for about 80 epochs. After that, we set $\mu = 1e - 3$ and gradually increase it to decrease communication costs. The RAN is trained with the SGD optimizer, and the other parts are trained with the Adam optimizer at a learning rate of 5×10^{-5} for 50 epochs and 5×10^{-6} for another 50 epochs under each μ. For joint-sensing experiments, we transmit the first $6f$ symbols of X rather than $12f$ for different f. Also, exploration strategies are used when sampling f. Specifically, the sampling probability is set to $0.6P_F + 0.4\hat{p}$, where we set $\hat{p}(1)$=$\hat{p}(2)$=$\hat{p}(3)$=$\hat{p}(4)$=0.25 to encourage the RAN to explore more on other actions so that the SCE/SCD can perform well on all actions.

3) Comparison of Semantic Communications in Fixed-Sensing Experiments: We show the performance comparison of different semantic communication methods in the fixed-sensing experiment in Fig. 12, where $\bar{l}_s$ denotes the average number of modulated symbols l_s used for the video clips in the test dataset. From the figure, the 'Compr+LDPC' performs slightly better than 'SemCom+noRAN'. At the same time, the proposed 'SemCom' method outperforms these methods to a relatively large extent, showing the advantage of jointly designing the channel coding and modulation. It also demonstrates the effectiveness of directly implementing the rate-distortion tradeoff on the modulated symbols through the proposed RAN. Furthermore, the proposed 'SemCom' has a similar performance to 'Compr+Cap', showing that semantic communications with joint designs are promising ways to approach Shannon capacity.

4) Comparison of Semantic Communications in Joint-Sensing Experiments: The performance comparison of different semantic communication methods in the joint-sensing experiment is shown in Fig. 13. From the figure, SemCom performs significantly better than 'Compr+LDPC' and even surpasses the 'Compr+Cap' in large $\bar{l}_s$ cases, further proving the benefits of joint designs. However, we cannot conclude that semantic communications can surpass Shannon capacity, as the performance of 'Compr+Cap' depends on our implementation of task-aware compression methods.

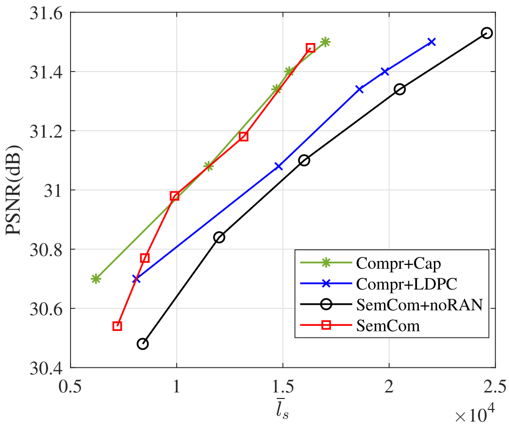


Fig. 12. The performance comparison among different semantic communication frameworks in fixed-sensing experiments.

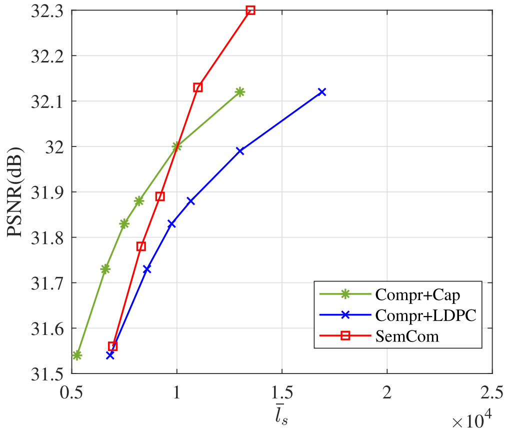


Fig. 13. The performance comparison among semantic communication frameworks in joint-sensing experiments.

5) Implementation of Conventional Communication Methods: Here, we explain how sensed data is transmitted in conventional communication systems and introduce their implementation details.

- **Transmit sensed data by joint source and channel coding (Sensordata+JSCC):** In conventional communication systems, the most straightforward way is to transmit raw sensed data directly, regardless of its usage. We should apply source and channel coding to raw sensed data to simulate this process. Recent works on deep JSCC [19] have shown that training a network for joint source and channel coding can perform better than using standardized image compression methods and channel coding. Therefore, we followed its design and built a deep network similar to SCE/SCD to transmit the raw sensor. The network

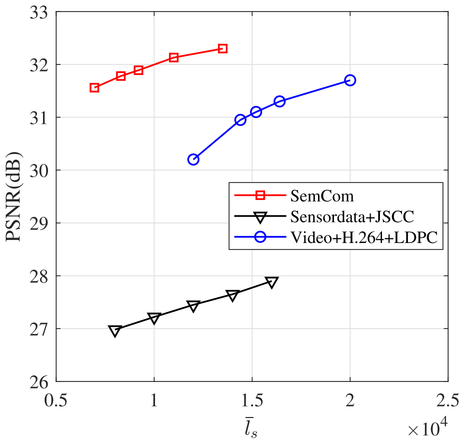


Fig. 14. The performance comparison among different transmission methods.

is optimized by the mean-squared error (MSE) between original and reconstructed sensed data. The reconstructed sensed data is then used to generate the target video.

- **Transmit reconstructed video by H.264 video coding plus LDPC plus QAM (Video+H.264+LDPC):** Another choice is to reconstruct the video first via a locally deployed video reconstruction network and then transmit the reconstructed video. In this way, the computational-costive reconstruction network needs to be run at the transmitter. We use H.264 [35] for video source coding, LDPC for channel coding, and QAM for modulation to transmit the video.

6) Comparison of Semantic Communications With Conventional Communication Systems: The performance comparison between semantic and conventional communication systems is shown in Fig. 14. From the figure, Sensordata+JSCC performs the worst because important data for video reconstruction networks does not get targeted protection during transmission. Video+H.264+LDPC performs better than Sensordata+JSCC as the videos have been reconstructed at the transmitter side. However, this method is still not as effective as SemCom, which shows we can achieve efficient transmission of sensor data without running a time-consuming and resource-costly reconstruction algorithm at the transmitter side.

VI. Conclusion

In this work, we propose a policy-gradient RL-based framework for rate-quality trade-off in SCI systems. The proposed framework can address the non-differentiable problem of "rate" and enable end-to-end optimization of rate allocation and video recovery networks. We first apply our framework in a novel video imaging system to learn the optimal per-pixel compression rate. The experiments show that learning ratios can significantly improve sensing efficiency. Next, we apply the framework in semantic communications to learn the optimal per-video transmission rate. The experiments demonstrate that the proposed framework can significantly improve transmission efficiency. We conclude that the proposed RL-based framework is universal and can be extended to many other rate-distortion problems in real-world applications.

References

[1] P. Llull et al., "Coded aperture compressive temporal imaging," *Opt. Exp.*, vol. 21, no. 9, pp. 10526–10545, 2013.

[2] A. Wagadarikar, R. John, R. Willett, and D. Brady, "Single disperser design for coded aperture snapshot spectral imaging," *Appl. Opt.*, vol. 47, no. 10, pp. B44–B51, 2008.

[3] A. A. Wagadarikar, N. P. Pitsianis, X. Sun, and D. J. Brady, "Video rate spectral imaging using a coded aperture snapshot spectral imager," *Opt. Exp.*, vol. 17, no. 8, pp. 6368–6388, 2009.

[4] X. Yuan, Y. Liu, J. Suo, and Q. Dai, "Plug-and-play algorithms for large-scale snapshot compressive imaging," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1447–1457.

[5] S. J. Carey, A. Lopich, D. R. Barr, B. Wang, and P. Dudek, "A 100,000 FPS vision sensor with embedded 535GOPS/W 256 × 256 SIMD processor array," in *Proc. Symp. VLSI Circuits*, 2013, pp. C182–C183.

[6] J. Chen, S. J. Carey, and P. Dudek, "Feature extraction using a portable vision system," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst., Workshop Vis.-Based Agile Auton. Navigation UAVs*, 2017, pp. 3–4.

[7] M. Z. Wong, B. Guillard, R. Murai, S. Saeedi, and P. H. Kelly, "AnalogNet: Convolutional neural network inference on analog focal plane sensor processors," 2020, *arXiv:2006.01765*.

[8] C. A. Metzler, H. Ikoma, Y. Peng, and G. Wetzstein, "Deep optics for single-shot high-dynamic-range imaging," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1375–1385.

[9] J. N. P. Martel, L. K. Müller, S. J. Carey, P. Dudek, and G. Wetzstein, "Neural sensors: Learning pixel exposures for HDR imaging and video compressive sensing with programmable sensors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 7, pp. 1642–1653, Jul. 2020.

[10] M. Iliadis, L. Spinoulas, and A. K. Katsaggelos, "Deepbinarymask: Learning a binary mask for video compressive sensing," *Digit. Signal Process.*, vol. 96, 2020, Art. no. 102591.

[11] C. M. Nguyen, J. N. P. Martel, and G. Wetzstein, "Learning spatially varying pixel exposures for motion deblurring," in *Proc. IEEE Int. Conf. Comput. Photography*, 2022, pp. 1–11.

[12] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Proc. Adv. Neural Inf. Process. Syst.*, 1999, pp. 1–7, vol. 12.

[13] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy, "Improved image captioning via policy gradient optimization of spider," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 873–881.

[14] Z. Qin, X. Tao, W. Tong, J. Lu, and G. Y. Li, "Semantic communications: Principles and challenges," 2021, *arXiv:2201.01389*.

[15] B. Zhang, Z. Qin, and G. Y. Li, "Semantic communications with variable-length coding for extended reality," *IEEE J. Sel. Topics Signal Process.*, vol. 17, no. 5, pp. 1038–1051, Sep. 2023.

[16] Q. Lan et al., "What is semantic communication? A view on conveying meaning in the era of machine intelligence," *J. Commun. Inf. Netw.*, vol. 6, no. 4, pp. 336–371, 2021.

[17] Z. Qin, J. Ying, D. Yang, H. Wang, and X. Tao, "Computing networks enabled semantic communications," *IEEE Netw.*, vol. 38, no. 2, pp. 122–131, Mar. 2024.

[18] M. Song, J. Choi, and B. Han, "Variable-rate deep image compression through spatially-adaptive feature transform," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 2380–2389.

[19] 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.

[20] E. Vargas, J. N. Martel, G. Wetzstein, and H. Arguello, "Time-multiplexed coded aperture imaging: Learned coded aperture and pixel exposures for compressive imaging systems," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 2692–2702.

[21] J. Bacca, T. Gelvez-Barrera, and H. Arguello, "Deep coded aperture design: An end-to-end approach for computational imaging tasks," *IEEE Trans. Comput. Imag.*, vol. 7, pp. 1148–1160, 2021.

[22] H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, "Deep learning enabled semantic communication systems," *IEEE Trans. Signal Process.*, vol. 69, pp. 2663–2675, 2021.

[23] H. Xie, Z. Qin, X. Tao, and K. B. Letaief, "Task-oriented multi-user semantic communications," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 9, pp. 2584–2597, Sep. 2022.

[24] H. Xie, Z. Qin, and G. Y. Li, "Task-oriented multi-user semantic communications for VQA," *IEEE Wirel. Commun. Lett.*, vol. 11, no. 3, pp. 553–557, Mar. 2022.

[25] J. Shao, Y. Mao, and J. Zhang, "Learning task-oriented communication for edge inference: An information bottleneck approach," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 197–211, Jan. 2022.

[26] M. Jankowski, D. Gündüz, and K. Mikolajczyk, "Wireless image retrieval at the edge," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 1, pp. 89–100, Jan. 2021.

[27] 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.

[28] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimized image compression," in *Proc. Int. Conf. Learn. Representation*, 2017, pp. 1–27.

[29] M. Yang and H.-S. Kim, "Deep joint source-channel coding for wireless image transmission with adaptive rate control," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process*, 2022, pp. 5193–5197.

[30] S. Wang et al., "Wireless deep video semantic transmission," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 214–229, Jan. 2023.

[31] Y. Cai et al., "Multi-stage spectral-wise transformer for efficient spectral reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 19–20.

[32] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Proc. Med. Image Comput. Comput.-Assist. Interv.*, 2015, pp. 234–241.

[33] S. Leonardos, W. Overman, I. Panageas, and G. Piliouras, "Global convergence of multi-agent policy gradient in Markov potential games," in *Proc. Int. Conf. Learn. Representation*, 2021, pp. 1–29.

[34] H. Kiani Galoogahi, A. Fagg, C. Huang, D. Ramanan, and S. Lucey, "Need for speed: A benchmark for higher frame rate object tracking," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1125–1134.

[35] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H. 264/AVC video coding standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 560–576, Jul. 2003.

Bowen Zhang (Graduate Student Member, IEEE) received the B.Sc. degree in telecommunication engineering and the M.Sc. degree in telecommunication and information systems from Beijing Jiaotong University, Beijing, China, in 2018 and 2021, respectively. Since 2021, he has been working toward the Ph.D. degree with the Department of Electrical and Electronic Engineering, Imperial College London, London, U.K., under the supervision of Prof. Geoffrey Ye Li. His research interests include compressed sensing, transformers, integrated sensing and communications, and semantic communications.

Zhijin Qin (Member, IEEE) is currently an Associate Professor with the Department of Electronic Engineering, Tsinghua University, Beijing, China. From 2016 to 2022, she was with Imperial College London, London, U.K., Lancaster University, Lancaster, U.K., and the Queen Mary University of London, London. Her research interests include semantic communications and sparse signal processing. She was a recipient of several awards, such as the 2018 IEEE Signal Processing Society Young Author Best Paper Award, 2022 IEEE Communications Society Fred W. Ellersick Prize, 2023 IEEE ICC Best Paper Award, and 2023 IEEE Signal Processing Society Best Paper Award. She has been an Area Editor of IEEE Journal on Selected Areas in Communications (JSAC) Series on Machine Learning for Communications and Networking and a Guest Editor for IEEE JSAC Special Issue on Semantic Communications. She was the Symposium Co-Chair for various flagship conferences, such as IEEE GLOBECOM 2020/ 2021/2024. She is an Area Editor of IEEE Communications Letters, an Associate Editor for IEEE Transactions on Communications and IEEE Transactions on Cognitive Communications and Networking.

Geoffrey Ye Li (Fellow, IEEE) is currently a Chair Professor with Imperial College London, London, U.K. Before joining Imperial in 2020, he was a Professor with the Georgia Institute of Technology, Atlanta, GA, USA, for 20 years and a Principal Technical Staff Member with AT&T Labs – Research (previous Bell Labs), Murray Hill, NJ, USA, for five years. He made fundamental contributions to orthogonal frequency division multiplexing for wireless communications, established a framework on resource cooperation in wireless networks, and introduced deep learning to communications. In these areas, he has authored or coauthored around 700 journal and conference papers in addition to more than 40 granted patents. His publications have been cited more than 68,000 times with an H-index of 119. He has been listed as a Highly Cited Researcher by Clarivate/Web of Science almost every year. Dr. Li was elected to IEEE Fellow and IET Fellow for his contributions to signal processing for wireless communications. He was the recipient of 2024 IEEE Eric E. Sumner Award, 2019 IEEE ComSoc Edwin Howard Armstrong Achievement Award, and several other awards from IEEE Signal Processing, Vehicular Technology, and Communications Societies.