Task-Oriented Semantic Communication with Semantic Reconstruction: An Extended Rate-Distortion Theory Based Scheme

Fangfang Liu, Member, IEEE, Wanjie Tong, Zhengfen Sun, Yang Yang, Member, IEEE, and Caili Guo, Senior Member, IEEE

Abstract—Compressed semantic representation of source is essentially important to accomplish various artificial intelligence (AI) tasks in task-oriented semantic communication (TOSC). In this paper, by extending the rate-distortion theory to multiple tasks, we propose a TOSC scheme with semantic reconstruction (SR), named as TOSC-SR, in the joint source and channel coding (JSCC) framework. Besides extracting and compressing task semantics, our basic idea here is to reconstruct images with task semantics rather than traditionally in the pixels or features. The main purpose is to share the semantic-reconstructed images among multiple tasks with enhanced generalization ability under certain rate. We formulate the TOSC-SR scheme as a rate-distortion optimization problem, where a novel semantic distortion measurement is defined by mutual information of source, the semantic-reconstructed images, and task labels, pairwise. We derive an analytic solution for the formulated problem, where the self-consistent equations are obtained to determine the optimal mapping of source and the semantic-reconstructed images by taking task labels into account. In the TOSC-SR scheme which is feasible in practice, a relaxed version of loss function is derived based on variational approximation of mutual information. Then we adopt the classification task to train TOSC-SR, and the object detection task to validate the generalization ability. Experimental results show that the proposed TOSC-SR scheme outperforms conventional JPEG, JPEG2000 based communication schemes and deep learning based TOSC with general reconstruction schemes in terms of reconstruction quality, classification and object detection performance at the same source compression ratio and signal-to-noise (SNR) regime. Compared with conventional rate-distortion and information bottleneck guided communications, the proposed scheme shows better multi-task generalization ability.

Index Terms—Semantic communication, semantic reconstruction, rate-distortion theory, mutual information, multi-task generalization ability.

I. INTRODUCTION

SEMANTIC communication is categorized by Shannon and Weaver as a higher level of communication beyond the traditional technical level. The goal of semantic communication is semantic information exchange rather than the transmission of symbols [1], [2]. With the tighter and deeper integration of communications and artificial intelligence (AI), semantic communication is unfolding as a revolutionary paradigm of the next generation wireless networks [3], [4]. In this novel paradigm, not only the wireless transceivers are spanning from normal devices to intelligent entities with edge intelligence capabilities, but also the applications are extended from connecting people to connecting intelligent unmanned systems with cloud intelligence capabilities. Besides human, the focus in semantic communication becomes mainly on empowering the intelligent entities to accomplish various AI tasks within finite spectrum resource, especially visual tasks like pedestrian monitoring, defect detecting, and security surveillance [4]. The framework of semantic communication provides an entirely new way to enable the intelligent entities such as autonomous cars, robots, and machines to collect, process, transmit, compute, and predict tremendous visual data with limited or even without human participation in a more efficient way. In these task-oriented semantic communication systems, how to represent semantic information of source concisely is the essential issue, in order to not only reduce the wireless transmission cost but also improve the performances of multiple visual tasks at the receiver.

To evaluate how good a representation of source is, the distortion measurement is generally defined by the distance between source and its representation according to the rate-distortion theory [5]. In this way, the problem can be characterized as to find the minimum expected distortion subject to a given rate, where rate refers to the size of data representation after compression in order to meet the spectrum requirement. In the image representation example, lossy image compression techniques for wireless transmission have been widely investigated with three categories of distortion measurements: the reconstruction distortion, the feature distortion, and the semantic distortion.

In the reconstruction distortion based communications, images are usually compressed and reconstructed at the receiver in the pixel domain to ensure the quality of visual perception for human. Here the reconstruction distortion is generally specified as the mean square error (MSE) between the original input images and the reconstructed images in pixels. The image compression techniques are one hand based on the handcrafted time-frequency transforms traditionally such as JPEG and JPEG 2000 [6], [7], and the other hand based on the flexible nonlinear transforms provided by deep learning in recent years, including auto-encoder or generative adversarial networks (GANs) [8]–[12]. These methods are mainly committed to achieving a better trade off between the rate, or image representation size, and the average reconstruction distortion of the images.

In the feature distortion based communications, images can be compressed and reconstructed in the feature domain for deep learning tasks rather than for human consumption [13]. The feature distortion is measured by the distance between the features extracted from the input source images and that extracted from the reconstructed images. Under this distortion measurement, the main purpose is to reserve the global features of source in the reconstructed images. Then the downstream tasks can be well accomplished. However, it should be noted that the performance of tasks is not considered in the compression and reconstruction.

In the semantic distortion based communications, images are compressed in the task domain by taking the performance of AI tasks into account. A specific task has been introduced to provide an additional variable to choose the relevant features as the semantic representation of the source [3], [4], [14]–[16]. Different from the image representations which target...
to reconstruct the global information in the pixel or feature domains, the problem has been changed to capture the semantic features about the prediction of the introduced specific task \cite{4}, \cite{16}. The information bottleneck (IB) distortion has been proposed as the semantic distortion about the single specific task, which measures the distance of the semantic information for the task from source to the captured features \cite{14}. Thus, the rate can be reduced significantly by the effectively compressed semantic features. In addition, the joint source-channel coding (JSCC) methods have been introduced to realize robust semantic transmission over the wireless channels \cite{3}, \cite{17}. In the previous work \cite{4}, \cite{16}, \cite{18}, different tasks may lead to different semantic features for the same source images. Under a given rate, the smaller the IB distortion is, the more relevant the semantic features will be to the specific task, which however degrades the generalization ability to perform other AI tasks.

Considering the various tasks in the intelligent entities in practical applications, it is necessary to study a new semantic distortion measurement to guide the more generalized representation of source images. Our basic idea is to reconstruct the messages in the task domain rather than traditionally in the pixel or feature domains, which can enhance the generalization ability of the semantic-reconstructed images for multiple tasks with certain rate and effectively avoid the frequent representation transmissions for each task. The main contributions of this paper are summarized as follows.

- We investigate the extended rate-distortion theory based TOSC-SR scheme, in which a new semantic distortion measurement is defined to guide the semantic representation, transmission, and reconstruction for multiple AI tasks. By formulating TOSC-SR as a rate-distortion optimization problem, we derive its analytic solution from the information theory perspective, obtaining the self-consistent equations for determining the optimal mapping between source and the semantic-reconstructed images by taking task labels into account.
- We construct the TOSCS-SR scheme in the JSCC framework for more reliable transmission. We describe the system architecture with JSCC encoder and decoder, where SR is realized in the JSCC decoder. Moreover, the additive white Gaussian noise (AWGN) channel is involved. We derive a relaxed version of loss function based on variational approximation \cite{19}. The relaxed loss function will be used for practical training. We further analyze the relationship between the loss function and mutual information, which gives the information theory explanation to semantic reconstruction. To estimate the mutual information between high dimensional latent features, we apply a feasible mutual information estimator CLUB.
- To validate the generalization ability of the proposed TOSC-SR scheme on multiple AI tasks, we train TOSC-SR with the classification task first, then verify with the object detection task. The good generalization ability among multiple AI tasks has been proven in the proposed TOSC-SR scheme.
- To compare the performance of intelligent tasks and reconstruction quality of received images with conventional and deep learning based schemes, extensive experiments are carried out under different compression ratios and SNR regimes. Some visualization results of classification and semantic reconstruction are given. Four mutual information estimation results are presented to explain the effectiveness of the proposed scheme on intelligent tasks from the perspective of information theory.

The rest of this paper is organized as follows. Related works are reviewed in Section II. The system model is presented and a corresponding problem is formulated in Section III. Section IV shows the detailed description of the proposed TOSC-SR scheme. The experimental results along with discussions are presented in Section V. Finally, conclusions are drawn in Section VI.

II. RELATED WORK

A. Definition and Measurement of Semantic Information

Since semantic communication was proposed, the definition and measurement of semantic information has always been the theoretical basis and key issue. Despite decades of development, this is still an open problem \cite{20}. In classic information theory, Shannon regarded anything that could reduce uncertainty as information, and obtained entropy as the measure of information according to the statistical characteristics of information \cite{1}. Along with this method analogously, earlier researchers also defined semantic information based on probability, which was mainly designed for textual processing. Different from the statistical probabilities used in classic information theory, Carnap et al. \cite{21} introduced the semantic information theory based on logical probabilities ranging over the contents. Barwise et al. \cite{22} further proposed the principle of scene logic to define semantic information. Floridi et al. \cite{23} introduced the theory of strongly semantic information. Since fuzzy concepts are very common in text communication, and different people have different understandings of fuzzy concepts, Zadeh proposed fuzzy sets and fuzzy events to describe the fuzziness of semantic information, which is measured by membership function \cite{24}, \cite{25}. However, due to the subjective semantic understanding of the same message by different people, it is difficult to obtain the logical probability and membership functions accurately in practice.

Recently, most researchers in semantic communication adopt deep neural networks (DNNs) to define, extract, and measure semantic information \cite{3}, \cite{14}, \cite{20}, \cite{26}. In these works, semantic is generally considered as the features extracted from the source that related to a introduced specific task, which is the essential difference with the earlier definitions. This definition of semantic information can be traced back to the IB theory \cite{14}, \cite{27}, where the mutual information was used to measure semantic information about the task in the source or in the extracted features. It is a very tricky thing to calculate mutual information between high-dimensional variables. Fortunately, many scalable and flexible mutual information estimation methods based on DNNs have emerged in recent studies \cite{28}, \cite{29}. Taking advantage of DNNs, it becomes feasible to obtain the amount of semantic information objectively by means of mutual information between the raw data and the task labels, or between the reconstructed signals and the task labels, which provides the basic foundation for our work.

B. Frameworks of Semantic Communication Systems

With the effectiveness of DNNs in the design of end-to-end physical layer communication systems, researchers have proposed various semantic communication schemes based on Transformer \cite{3}, \cite{35} and CNN \cite{4}, \cite{13}, \cite{15}–\cite{17}, \cite{20}, \cite{30}–\cite{34}, \cite{36}. Most of these semantic communication systems adopt the JSCC scheme, which has shown great advantages over traditional separated source and channel coding schemes in terms of noise resistance. According to different communication goals, the semantic communication frameworks can be shown in Fig.1: 1) general JSCC communication for image reconstruction in the pixel or feature domain, 2) semantic communication for one specific task, 3) semantic communication for multiple tasks with respective JSCC codecs, and 4) semantic communication for multiple tasks with one generalized JSCC codec, which is proposed in this paper.

As shown in Fig. 1(a), the purpose of this type of JSCC communication is to reconstruct the original image using the extracted features in the pixel or feature domain, so that the recovered signal and the original signal are as consistent as
possible in appearance. Bourtsoulatze et al. [17] designed a JSCC image communication system based on CNN, where peak signal-to-noise ratio (PSNR) is used to measure the accuracy of image recovery at the receiver. By incorporating the channel output feedback into the transmission system, Kurka et al. [30] improved the reconstruction quality at the receiver. In order to make reconstructed images similar to original images not only in pixel-level, but also in feature-level, based on the works of Yang et al. [13], Niu et al. [20] proposed a communication system by jointly optimize the appearance loss and multi-scale discriminator loss. In these researches, though the reconstructed images are more similar with original ones in both appearance and perception, the whole reconstruction training process does not involve the gradient feedback of tasks, so the semantics information relevant to task are still not considered.

As shown in Fig. 1(b), the purpose of this type of semantic communication is to use the extracted semantic information to directly complete the specific task at the receiver, without the need to reconstruct the original signal. Lee et al. [31] designed a joint transmission-recognition scheme for the IoT devices by performing the feature extraction and recognition at the IoT device and server respectively. Along this line, Yang et al. [13] further compressed the transmitted data by introducing the semantic relationship between feature maps and semantic concept. Jankowski et al. [32, 33] proposed a task-based compression scheme for the image retrieval task in the wireless edge scenario. Xie et al. [34] considered a task-oriented multi-user semantic communication system for multimodal data transmission, the proposed system, named MU-DeepSC, is enabled by DNNs to execute the visual question answering (VQA) task. In summary, these non-reconstructed based semantic communication schemes often accompany with specific tasks, which may lose the generalization ability for the other tasks.

As shown in Fig. 1(c), this type of semantic communication system considers multiple tasks at the receiver. By incorporating the semantic information into the codec during image compression, Luo et al. [16] proposed a concept called deep semantic compression (DeepSC) and designed two semantic compression networks by performing the semantic analysis after the feature extractor in the transmitter. In line with this idea, Patwa et al. [4] modified the auto-encoder networks and obtained better multi-task (classification and reconstruction) performance. The schemes proposed by [4], [16] are actually multi-task type semantic communication system, in which the multiple tasks will compete with each other for better performance, resulting in a suboptimal solution. In order to unify the structure of transmitters for different tasks, Xie et al. [35] proposed a Transformer based unique framework. While, multiple transmission of semantic information may be introduced for different tasks, where the limitation of wireless resources should be further considered.

Motivated by these existing semantic communication frameworks, we designed a task-oriented semantic communication system with semantic reconstruction, as shown in Fig. 1(d). Here the input images are reconstructed in the task domain to not only maintain the task performances with the gradient feedback, but also enhance the generalization ability among different tasks. In this way, the transmission cost can be effectively reduced by using the semantic-reconstructed images to accomplish the various tasks at the receiver.

### C. Image Compression and Semantic Representation

As a very common source of information, images are playing an increasingly important role in current communication systems. Different from the natural language, which is human-generated signals that are highly semantic and information-dense [37], images are generally natural signals with quite objective descriptions of things. The rich semantic information is contained in images with heavy spatial redundancy, which leads the image compression a critic problem.

Traditional lossy image compression methods, such as JPEG [6], JPEG2000 [7] and HEVC [38], rely on hand-crafted module design individually. Each module was designed with multiple modes and optimized by the rate-distortion theory to determine the best mode. Recently, a great number of learning-based image compression models utilizing auto-encoder architecture have been proposed, which have achieved great success with promising results [8–12]. Similar to traditional methods, these models were still optimized with the reconstruction distortion,
without considering the contents or semantics of images. Li et al. [39] proposed a content-weighted strategy to allocate more bits to the important parts of an image, which was determined by a simple fully connected network. Considering the region of interest (ROI) of an image, Cai et al. [40] designed an end-to-end optimized ROI image compression scheme, which achieved better visual quality and compression performance than traditional compression methods in ROI. Note that these compression schemes in [39], [40] did not pay attention to the task performance, so the features of image compression are in essence general representations for image reconstruction without semantic.

To extract more semantic information related to the AI tasks during image compression, semantic presentations have been studied extensively. By the effectiveness of DNNs, intelligent entities can perform various AI tasks based on images, such as image classification, image generation, semantic segmentation, object detection, image retrieval and so on [41]. As different tasks require different semantic representations, when it comes to wireless transmission, how to obtain effective semantic representations with limited transmission cost becomes a very worthwhile problem.

D. Rate Distortion and Information Bottleneck

Most of the previous researches on lossy image compression [8], [9], [11] and communication [17], [30] are based on the rate-distortion theory, where the rate is defined as the mutual information between source and its representation, and the distortion is directly defined as the difference between the reconstructed images and the original images in the pixel or feature domain using MSE.

As an extension of the rate-distortion theory, the IB theory was first proposed by Tishby et al. [14] in 1999. The authors formulated a theoretic framework that finds concise representations related to the task variable for the source variable. The IB distortion is defined and extended as the mutual information between representation and task labels. The IB theory has been proven useful for various supervised and unsupervised tasks [27], [44]. Recently, IB has been used to explain the optimization process of DNNs [27]. And the visual experiments in [45] show that the optimization process is indeed corresponding to the optimization objective of IB. Inspired by this, Alemi et al. [19] adopted a variational approximation to the IB, then optimized the supervised learning model constructed by DNN using the relaxed IB trade-off, and obtained better performance than other forms of regularization.

Though some preliminary semantic communication systems have been proposed by researchers, to our best knowledge, there are little theoretical researches in semantic communication systems using the rate-distortion theory as in the traditional lossy compression field. In the semantic communication system proposed in this paper, we further extend the rate-distortion theory by considering the generalization ability among different tasks. Here a new semantic distortion is defined as the trade-off between the semantic reconstructed distortion and the IB distortion oriented towards AI tasks.

III. SYSTEM MODEL AND PROBLEM FORMULATION

In this section, we propose the task-oriented semantic communication with semantic reconstruction (TOSC-SR) scheme using the JS defense framework, giving consideration to both predicting precision and generalization ability among multiple tasks. The TOSC-SR is formulated as a extended rate-distortion problem, and its analytic solution will be derived by solving the equivalent Lagrange multiplier optimization problem.

A. System model

As shown in Fig. 2, the system model of proposed TOSC-SR consists of JS defense encoder, quantizer, phisical channel, JS defense decoder, and AI task networks in the receiver. Let X denote random variable from the source image space. In general, X needs to be compressed to save bandwidth. The relevant variable Y is the desired output of AI task when corresponding X is fed to the task network. JS defense encoder maps the input to semantic representations, which are then quantized by the quantizer to reduce the transmitter cost. After semantic encoding and quantification, the quantized symbols Z will be transmitted to the receiver through physical channel. In the receiver, the JS defense maps the noisy symbols to reconstructed image \( \hat{X} \). At last, AI task network takes \( \hat{X} \) as input and outputs the predicting result \( \hat{Y} \).

Note that the proposed semantic communication framework applies the quantization on the semantic representations. In order to reduce the transmission costs, channel input symbols should be quantified to match the finite precision waveform of cost-sensitive transmitters. On the other hand, quantization helps to calculate mutual information in subsequent analysis. Therefore, we quantify the symbols before sending them into the channel.

In the proposed semantic communication framework, the reconstruction task and other AI tasks both can affect the whole JS defense encoder and decoder pipeline by introducing the corresponding gradient feedback.

B. Semantic Distortion Measurement

In this section, we will show the semantic distortion measurement based on the extended rate-distortion theory.

In order to ensure that the reconstructed images can perform the task better, we should minimize the IB distortion. Denote \( d_{IB}(x, \hat{x}) \) as the relevant information distortion between input image \( x \) and reconstructed image \( \hat{x} \), in the standard information bottleneck method, the IB distortion function between \( x \) and \( \hat{x} \) is represented by the KL divergence \( D_{KL}(p(y|x)p(y|\hat{x})) \) [14], which is given by

\[
d_{IB}(x, \hat{x}) = \sum_y p(y|x) \log \left( \frac{p(y|x)}{p(y|\hat{x})} \right),
\]

and the expectation of IB distortion between \( X \) and \( \hat{X} \) is \( I(X;Y|\hat{X}) \) [27],

\[
D_{IB}(X, \hat{X}) = E[d_{IB}(x, \hat{x})] = \sum_{x} \sum_{\hat{x}} p(x, \hat{x}) p(y|x) \log \left( \frac{p(y|x)}{p(y|\hat{x})} \right).
\]

The following theorem shows that IB distortion has another format, and the different formats are equivalent.

**Theorem 1.** The information bottleneck distortion defined in (2) has another formulation as

\[
D_{IB}(X, \hat{X}) = I(X;Y) - I(\hat{X};Y),
\]

and the different formats are equivalent.

**Proof.** See Appendix A. \(\blacksquare\)

The definition of \( I(\hat{X};Y) \) is used here because this expression more intuitively illustrates the reduction of relevant information. In order to enhance generalization ability among different

*In this work, \( X, Y, Z, \hat{X} \) are random variables, \( x, y, z, \hat{x} \) are single instance of corresponding random variables, \( x, y, z, \hat{x} \) are multiple instances of corresponding random variables.*
AI tasks, we should minimize the reconstruction distortion, which is given by

$$D_{RD}(X, \hat{X}) = \sum_{x \in X} \sum_{\hat{x} \in \hat{X}} p(x, \hat{x})d_{RD}(x, \hat{x}),$$

(4)

here we use the MSE as reconstruction distortion function,

$$d_{RD}(x, \hat{x}) = (x - \hat{x})^2,$$

(5)

considering the trade-off between two distortion, we can define the following semantic distortion measurement,

$$D_S(X, \hat{X}) = D_{RD}(X, \hat{X}) + \beta D_{IB}(X, \hat{X}).$$

(6)

where $\beta$ controls the trade-off between task’s predicting precision and system’s generalization ability.

C. Problem Formulation and Analytic Solution

Based on the semantic distortion defined in (6), the TOSCSR scheme can be formulated as an extended rate-distortion optimization problem, which is optimized for the trade-off between concise semantic representation and multi-task generalization ability. The analytic solution is obtained by solving the equivalent lagrange function and self-consistent equations are derived by using the alternate iteration Blahut–Arimoto algorithm [14].

If minimizing semantic distortion $D_S$ is our only optimized objective, the best solution would be an identity map of original data $(Z = X)$, but for communication systems, it is obviously very wasteful to transmit raw data directly without compression. Similar to rate distortion optimization, we apply a constraint on the rate $I(X; \hat{X})$, which is given by

$$I(X; \hat{X}) = \sum_{x \in X} \sum_{\hat{x} \in \hat{X}} p(x, \hat{x}) \log \left( \frac{p(x, \hat{x})}{p(x)p(\hat{x})} \right),$$

(7)

and the constraint on $I(X; \hat{X})$ is

$$I(X; \hat{X}) \leq I_C,$$

(8)

then we can get the following optimization problem,

$$\min_{p(\hat{x}|x); I(X; \hat{X}) \leq I_C} D_S(X, \hat{X}),$$

(9)

combined with equation (6) and considering the normalization of conditional probability $p(\hat{x}|x)$, optimization problem (9) can be converted to

$$\min_{p(\hat{x}|x); I(X; \hat{X}) \leq I_C} D_{RD}(X, \hat{X}) + \beta D_{IB}(X, \hat{X}),$$

(10)

according to equation (3), since $I(X; Y)$ is a constant with respect to dataset, optimization problem (10) can be converted to

$$\min_{p(\hat{x}|x); I(X; \hat{X}) \leq I_C} D_{RD}(X, \hat{X}) - \beta I(\hat{X}; Y).$$

(11)

This optimization problem can be solved by the Lagrange multiplier method, and the only variable is $p(\hat{x}|x)$. The complete process for solving this problem is shown in Appendix B. Here we directly give the analytic solution.

**Theorem 2.** The optimal mapping from source images set $X$ to semantic-reconstructed images set $\hat{X}$ satisfy the following equation,

$$p(\hat{x}|x) = \frac{p(\hat{x})e^{-\beta d_S(x, \hat{x})}}{\mu(x)},$$

(12)

$$p(\hat{x}) = \sum_{x} p(x)p(\hat{x}|x),$$

(13)

$$p(y|\hat{x}) = \sum_{x} p(y|x)p(x|\hat{x}),$$

(14)

where $\mu(x)$ is given by

$$\mu(x) = \sum_{\hat{x}} p(\hat{x})e^{-\beta d_S(x, \hat{x})}.$$

(15)

**Proof.** See Appendix B.

To find the optimal distributions $p(\hat{x}|x)$, $p(\hat{x})$, and $p(y|\hat{x})$, it is natural to think of using alternate iteration Blahut–Arimoto algorithm [5], according to the traditional rate distortion function. First we pick a $\lambda$ and a $\beta$, and initialize the distribution $p(\hat{x}|x)$ and $p(y|\hat{x})$, then minimize the semantic distortion to find the distribution $p(\hat{x}|x)$. The self consistent equations (12), (13), and (14) are satisfied simultaneously at the minima of the functional $\mathcal{L}(p(\hat{x}|x))$. The minimization is done independently over the convex sets of the normalized distributions, $p(\hat{x}|x)$ and $p(y|\hat{x})$. Namely

$$\min_{p(\hat{x}|x)p(\hat{x})p(y|\hat{x})} \min_{p(\hat{x}|x)p(\hat{x})p(y|\hat{x})} \mathcal{L}(p(\hat{x}|x); p(\hat{x}); p(y|\hat{x})).$$

(16)

This minimization is performed by the converging alternating iterations. Denote $k$ as the iteration step, we have the self-consistent equations as

$$p_k(\hat{x}|x) = \sum_{\hat{x}} p(\hat{x})e^{-\beta d_S(x, \hat{x})},$$

(17)

$$p_{k+1}(\hat{x}) = \sum_{x} p(x)p_k(\hat{x}|x),$$

$$p_{k+1}(y|\hat{x}) = \sum_{x} p(y|x)p_k(\hat{x}|x)p(x).$$

It should be noted that $d_S(x|\hat{x})$ will directly affect the mapping $p(\hat{x}|x)$. As $d_S(x|\hat{x}) = d_{RD}(x, \hat{x}) + \beta d_{IB}(x, \hat{x})$ is related to $p(\hat{x}|x)$ according to equation (11), which indicates that the AI task’s predicting performance will be influenced by the trade-off between pixel-level distortion and AI task-level distortion. When $p(y|\hat{x})$ is fixed, we will go back to the rate distortion case with fixed distortion $d_{RD}(x, \hat{x})$ and $d_{IB}(x, \hat{x})$. By iteratively solving the above self-consistent equations, we can obtain the optimal mapping $p(\hat{x}|x)$ and achieve the optimal trade-off between rate and semantic distortion.
A. System Architecture

The proposed system is shown in the Fig. 3. The overall architecture consists of three parts, i.e. transmitter, channel and receiver. Transmitter is composed of JSCC encoder and quantifier. The general AWGN channel is adopted here. Receiver is mainly composed of JSCC decoder, ResNet-18 classification network, and RFB based object detection network [46]. It is important to note that we adopt the classification task as our basic AI task producing the training feedback on the semantic communication pipeline. The whole architecture is an end-to-end deep neural network. Since the subsequent AI task of the receiver will have a positive impact on the performance of the whole network, the images reconstructed by the receiver contain more task-relevant information and have a certain generalization ability among different AI tasks. The implementation details of each part are described below.

1) JSCC Encoder and Decoder: The JSCC encoder and decoder used in our system are fully convolutional networks. The encoder network consists of three convolution layers and three residual blocks. Except for the last convolution layer, which is followed by a sigmoid activation function so that encoder’s output can be in range of [0, 1], previous convolution layers and residual blocks are followed by relu activation function. Each residual block has two convolution layers. The whole encoder achieves eight times downsampling on the spatial dimension of the image. Specifically, the input image \( x \) is first convolved with 128 filters with size \( 8 \times 8 \) and stride 4 and followed by one residual block. The feature maps are then convolved with 256 filters with \( 4 \times 4 \) and stride 2 and followed by two residual blocks. Finally, feature maps are convolved with \( N \) filters with \( 3 \times 3 \) and stride 1 to yield semantic encoder output \( E(\ast) \). It should be noted that we set different \( K \) for different levels of compression.

IV. PROPOSED TOSC-SR SCHEME

Based on the theoretical analysis of the previous section, here, we will design a concrete TOSC-SR system. First, we introduce the components of proposed system architecture in detail. Then, the practical loss function is given by relaxing the mutual information optimization objective for ease of computation during the training process. Meanwhile, we point out that the loss function has close relationships with mutual information. Next, multi-task generalization validation is adopted to make the proposed system applicable to different AI tasks. Finally, corresponding training algorithms are presented.

D. Discussion

Observing the self-consistent equations \[17\], we notice that the analytic solution needs the specific dimension and value space of \( \hat{x} \). For low dimensional variables, the achievable value space is limit and the search of optimal mapping \( p(\hat{x}|x) \) is computationally acceptable. In this case, we can obtain the theoretical lower bound of semantic distortion at a certain rate. While for high-dimensional data as images, the proper dimension and value space for its representation are difficult to determine, which prevents the normal process of analytic solution.

According to \[8\]–\[12\], deep learning provides an excellent solution for seeking such a transformation from source to its representation. Observing the optimization problems \[11\], the essence of which is finding the smallest semantic distortion of conditional probability \( p(\hat{x}|x) \) under the condition of certain compression. If the optimization objective can be written as an expression and the gradient can be calculated by back propagation, then the conditional probability \( p(\hat{x}|x) \) can be solved by using DNNs. In the next section we will build a feasible solution for TOSC-SR scheme based on DNNs.

Fig. 3. The architecture of the TOSC-SR system.
The network of decoder $D(x)$ is almost symmetric to that of the encoder except some difference in the numbers of convolutional layers and filters in each layer. In addition, we apply depth-to-space method to implement upsampling.

2) Quantizer: Quantization is necessary for a practical communication system, especially for the cost-sensitive edge devices, because the continuous-like channel input symbols need high resolution waveforms to represent, which is unaffordable for most intelligent entities. While, once the quantization is introduced, it will be the only non-differentiable operation in the whole end-to-end semantic image communication network. The derivative of the rounding function is zero everywhere, rendering gradient descent ineffective. In order to initiate the training process, this problem should be solved. Some smooth approximations are proposed in related works. Balle et al. [8] introduced a task-relevant distortion, because the continuous-like channel input symbols need some smooth approximations. Thesis et al. [9] proposed a proxy function to replace the non-differentiable rounding in the back propagation. The form of the proxy function is simple, which is $z_{kij} = e_{kij}$. So $\frac{dz_{kij}}{de_{kij}} = 1$. Toderici et al. [10] used a stochastic binarization function as $z_{kij} = -1$ when $e_{kij} < 0$, and $z_{kij} = 1$ otherwise. Here we notice that either using a proxy function, or adding an uniform noise, their derivatives of corresponding rounding functions are both 1 in the back propagation.

In the proposed system, binary quantization is implemented by using a simple proxy function. Because the JSCC encoder’s last convolution layer is activated by sigmoid function, the values of $e = E(x)$’s should be in range of $[0,1]$. In the forward propagation, the binarizer is defined as

$$z_{kij} = \begin{cases} 1, & \text{if } e_{kij} > 0.5, \\ 0, & \text{if } e_{kij} \leq 0.5, \end{cases}$$

(18)

in the back propagation, the proxy function of binarizer is defined as

$$\tilde{z}_{kij} = \begin{cases} 1, & \text{if } e_{kij} > 1, \\ e_{kij}, & \text{if } 0 \leq e_{kij} \leq 1, \\ 0, & \text{if } e_{kij} < 0, \end{cases}$$

(19)

so the derivatives of the proxy function is,

$$\frac{d\tilde{z}_{kij}}{de_{kij}} = \begin{cases} 1, & \text{if } 0 \leq e_{kij} \leq 1, \\ 0, & \text{otherwise}, \end{cases}$$

(20)

after quantization, as the channel input symbols has been limited by 0 or 1, we can use bits to represent the amount of data input to the channel, and the bpp can be used to represent the compression ratio,

$$bpp = \frac{K \times 16 \times 16}{128 \times 128} = \frac{K}{64}.$$
notice that \( I(X;Y) - H(Y) \) is a constant, which is independent of our optimization procedure and can be ignored.

Above all, we know that,

\[
\min D_{RD}(X, \hat{X}) + \beta D_{1B}(X, \hat{X}),
\]

(28)
can be relaxed to minimizing its upper bound,

\[
\min D_{RD}(X, \hat{X}) - \beta E_{p(y|x)}[\log q(y|\hat{x})],
\]

(29)
here we adopt the mean square error function as our \( D_{RD} \),

\[
D_{RD}(X, \hat{X}) = \text{MSE}(X; \hat{X}) = \sum_{x \in X} \sum_{y \in X} ||x - \hat{x}||^2.
\]

(30)
we notice out that the negative expectation term \(-E_{p(y|x)}[\log q(y|\hat{x})] \) is actually the usual cross-entropy loss \( CE(Y, \hat{Y}) \) in supervised learning.

Put everything together to get the following objective function which we try to minimize:

\[
\mathcal{L}_R = \text{MSE}(X; \hat{X}) + \beta CE(Y, \hat{Y})
\]

(31)
with this loss function, we can train the proposed TOSC-SR system. Before doing this, we will further analyze this loss function from information theoretic by establishing the relationship with mutual information.

2) Mutual Information Estimation with CLUB: Mutual information (MI) is a fundamental quantity for measuring the relationship between random variables. Minimizing or maximizing MI has gained considerable interests in a wide range of deep learning tasks. Similar to the idea in \([27]\), which suggests using MI as the measurement of the quality of DNN. In our proposed TOSC-SR scheme, MI also plays an important role in the following two aspects.

First, we consider the information relevant to various target tasks as semantic information. MI provides a natural quantification of the relationship between random variables. Minimizing or maximizing MI has gained considerable interests in a wide range of deep learning tasks. Similar to the idea in \([27]\), which suggests using MI as the measurement of the quality of DNN. In our proposed TOSC-SR scheme, MI also plays an important role in the following two aspects.

Second, the optimization objective applied in our system, i.e. \([22]\), actually has close relationship with MI. The practical loss function adopted in the experiment, i.e. \([31]\), is a relaxed version of \([22]\). As shown in \([22]\) and \([23]\), the optimization objective consists of rate \( R \), reconstruction distortion \( D_{RD} \) and single task relevant distortion \( D_{1B} \). Next we will prove that all of the three parts have close relationship with MI.

Rate \( R = I(X; \hat{X}) \), which is the MI between \( X \) and \( \hat{X} \). On the one hand, this MI determine the minimal number of bits per pixel, and this is a theoretical lower bound. On the other hand, it also indicates the information about input \( X \) contained in the code \( \hat{X} \). In general, the more this information is, the better reconstruction quality will be achieved. According to the data processing inequality,

\[
I(X; \hat{X}) \leq I(X; Z)
\]

(32)
\( D_{RD} \) is the MSE between input \( X \) and reconstruction output \( \hat{X} \), as shown in equation \([30]\). MSE is the direct measurement of the difference between two variables, while the MI quantify the similarity between two variables. These two metrics appear to be moving in opposite directions. In practice, \( MSE(X; \hat{X}) \) is indeed related to \( I(X; \hat{X}) \). According to the definition of MI,

\[
I(X; Z) = H(X) - H(X|Z),
\]

(33)
as \( H(X) \) is a constant when dataset \( X \) is selected, we mainly focus on \( H(X|Z) \).

\[
H(X|Z) = - \sum_x p(x) \sum_z p(z|x) \log p(x|z))
\]

\[
= \mathbb{E}_{x \sim p(x)} \left[ \mathbb{E}_{z \sim p(z|x)} [-\ln p(x|z)] \right],
\]

(34)
where \( p(x|z) \) is the semantic decoder, if we assume it follows the multivariate gaussian distribution

\[
p(x|z) = N(x; \mu, \sigma^2 I_D),
\]

(35)
where \( \mu = D_{\phi}(z) = \hat{x} \). \( D \) is the total dimension of \( x \). So

\[
p(x|z) = \frac{1}{\sqrt{2\pi}\sigma^2} \exp \left( -\frac{1}{2} \left\| \frac{x - \hat{x}}{\sigma^2} \right\|^2 \right),
\]

(36)
then,

\[
- \log p(x|z) = \frac{1}{2} \left\| \frac{x - \hat{x}}{\sigma^2} \right\|^2 + \frac{D}{2} \ln (2\pi) + \frac{1}{2} \sum_{i=1}^D \ln \sigma_i^2,
\]

(37)
where the first term is the scaled MSE. Combine the equations \([33]\), \([34]\) and \([37]\), we can find that if we use MSE as the loss function, minimizing \( MSE(X; \hat{X}) \) means maximizing \( I(X; Z) \). This is intuitive and reasonable that the more information about \( X \) is contained in \( Z \), the better construction quality of \( \hat{X} \) is.

Single task relevant distortion \( D_{1B} \) is defined by the reduction of MI, as shown in \([5]\). Minimizing \( D_{1B} \) means maximizing \( I(\hat{X}; Y) \). According to the data processing inequality,

\[
I(\hat{X}; Y) \geq I(Y; Y).
\]

(38)
Above all, \( Z \) and \( \hat{Y} \) are two critical variables in the whole end-to-end communication pipeline. Meanwhile, \( Z \) and \( \hat{Y} \) are bottleneck variables (which have the minimal number of dimension) of JSCC codec and classifier respectively. So computing MI between these two variables is necessary.

Despite having excellent properties, the mutual information between high-dimensional variables has been quite difficult to calculate. Fortunately, there has emerged some sample-based MI estimation methods in recent years.

Belghazi et al. \([28]\) introduces a Mutual Information Neural Estimator (MINE), which treats MI as the Kullback-Leibler (KL) divergence between the joint and marginal distributions, and converts it into the dual representation as

\[
I_{\text{MINE}} := \mathbb{E}_{p(x,y)} [T_\theta(x,y)] - \log \mathbb{E}_{p(x)p(y)} \left[ e^{T_\theta(x,y)} \right],
\]

(39)
where \( T_\theta(x,y) \) is the function parametrized by a neural network, which takes samples \( x, y \) as input. Along with this method, the estimation of mutual information between high dimensional continuous random variables can be achieved by gradient descent over neural networks.

Recently, Cheng et al. \([29]\) introduce a Contrastive Log-ratio Upper Bound (CLUB). Specifically, CLUB bridges mutual information estimation with contrastive learning where MI is estimated by the difference of conditional probabilities between positive and negative sample pairs, which is

\[
I_{\text{CLUB}} := \mathbb{E}_{p(x,y)} \left[ \log p(y|x) \right] \mathbb{E}_{p(x)} \mathbb{E}_{p(y)} \left[ \log p(y|x) \right],
\]

(40)
where \( p(y|x) \) is the conditional distribution of \( y \) given \( x \). When \( p(y|x) \) is not provided, we can use a variational distribution \( q(y|x) \) to approximate \( p(y|x) \), and \( q(y|x) \) is usually implemented by the neural network.

MINE estimates the MI through the lower bound. CLUB estimates the MI through the upper bound. In our application, it doesn’t matter whether the bounds are upper or lower. We are more concerned about the accuracy of the estimate. According to the comparison experiments in \([29]\), CLUB achieves the best accuracy.

C. Training Algorithms

In the proposed TOSC-SR scheme, the main training process consists of two phases: 1) pre-train the image classifier and JSCC codec separately, 2) jointly train the JSCC codec and the classifier. Algorithm \([13]\) describe the main training process.
Algorithm 1 TOSC-SR training algorithm.

Input: The background knowledge $\mathcal{K}$, i.e., dataset.

1. Load the pre-trained feature extraction layer parameters and freeze the parameters of $F_x$.
2. while Stop criterion is not met do
   3. Forward: $X \rightarrow \hat{X} = D_\phi(E_\theta(X))$
   4. Loss: $L_c = MSE(X, \hat{X}) + \beta CE(Y, \hat{Y})$
   5. Back-propagation: $L_c \rightarrow \frac{\partial L_c}{\partial W_{c_{ca}}}$
   6. Update parameters: $W_{c_{ca}} := W_{c_{ca}} - lr \frac{\partial L_c}{\partial W_{c_{ca}}}$
end while

Output: The parameterized network $E^*_\theta$ and $D^*_\phi$.

Algorithm 2 Pre-train the classifier.

Input: The dataset $\mathcal{K}$.

1. Initialization: Load the pre-trained feature extraction layer parameters of ResNet18 network, modify the output classes number of fully connection layer, then we can get the classification network’s parameters $W_{cla}$.
2. while Stop criterion is not met do
   3. Forward: $X \rightarrow \hat{Y} = F_x(X)$
   4. Loss: $L_{cla} = CE(Y, \hat{Y})$
   5. Back-propagation: $L_{cla} \rightarrow \frac{\partial L_{cla}}{\partial W_{cla}}$
   6. Update parameters: $W_{cla} := W_{cla} - lr \frac{\partial L_{cla}}{\partial W_{cla}}$
end while

Output: The parameterized network $F_x$.

Algorithm 3 Pre-train the JSCC codec.

Input: The dataset $\mathcal{K}$.

1. Initialization: Initialize the auto-encoder network’s parameters $W_{cae}$.
2. while Stop criterion is not met do
   3. Forward: $X \rightarrow \hat{X} = D_\phi(E_\theta(X))$
   4. Loss: $L_{cae} = MSE(X, \hat{X})$
   5. Back-propagation: $L_{cae} \rightarrow \frac{\partial L_{cae}}{\partial W_{cae}}$
   6. Update parameters: $W_{cae} := W_{cae} - lr \frac{\partial L_{cae}}{\partial W_{cae}}$
end while

Output: The parameterized network $E_\theta$ and $D_\phi$.

Algorithm 4 Describes the steps of estimating mutual information provided by CLUB method.

Input: Dataset $\mathcal{K}$, fine-tuned auto-encoder $E'_\theta$, $D'_\phi$.

1. Initialization: Initialize a variational network $q_\xi(z|x)$ with parameters $\xi$.
2. Collect variables $X$ and $Z$: $X \rightarrow Z = E_\theta(X)$
3. while Stop criterion is not met do
   4. Sampling: Sample $(x_i, y_i)^N_{i=1}$ from $p(x, z)$;
   5. Forward: $X \rightarrow \mu, \sigma^2 \rightarrow \hat{Z}$;
   6. Loss: $L_{CLUB} = MSE(Z, \hat{Z})$
   7. Back-propagation: $L_{CLUB} \rightarrow \frac{\partial L_{CLUB}}{\partial W_{CLUB}}$
   8. Update parameters: $W_{CLUB} := W_{CLUB} - lr \frac{\partial L_{CLUB}}{\partial W_{CLUB}}$
end while

Output: The estimating mutual information $I_{CLUB}$.

Algorithm 5 Generalization ability validation: from classification task to object detection task.

Input: The dataset $\mathcal{K}$, fine-tuned auto-encoder $E'_\theta$, $D'_\phi$ pretrained object detection network $T_\psi$.

1. Test with normal auto-encoder: $X \rightarrow \hat{X} = D_\phi(E_\theta(X))$
2. Test with fine-tuned auto-encoder: $X \rightarrow \hat{X}^* = D'_\phi(E'_\theta(X))$
3. Compare the mAP performance of $\hat{Y}$ and $\hat{Y}^*$

Output: The generalization ability.

V. EXPERIMENTS

In this section, we provide extensive experiments to show the effectiveness of proposed task-oriented semantic communication with semantic reconstruction (TOSC-SR) system.
These experiments consist of four parts. First, we compare the proposed method with traditional and DNN-based methods in terms of the AI tasks performance and reconstruction quality under different bpp and SNR regimes. Second, four specific mutual information are estimated to validate the assertion that, the loss function we actually used has close relationships with mutual information. On the other hand, the changes of the these mutual information can also be used to explain the changes of performance in the first part of experiments. Then, in order to validate the generalization among different AI tasks of the proposed TOSC-SR scheme, we test the generalization ability on object detection task. Finally, the impacts of parameter \( \beta \) are explored , which shows that the trade-off between \( D_{RD} \) and \( D_{IB} \) is necessary.

### A. Simulation settings

1) **Datasets:** In our experiments, STL10 dataset is adopted for classification, and Pascal VOC 2007 dataset is adopted for object detection.

The STL-10 dataset is an image recognition dataset for developing deep learning algorithms. Specifically, there are 100000 unlabeled images for unsupervised learning. The image size is \( 96 \times 96 \). For supervised learning, it has 10 classes, each class has 500 training images and 800 test images with labels. Since our method is semi-supervised, we mainly utilize the labeled images for training and testing.

The Pascal VOC 2007 dataset provides standardised image data sets for object class recognition. The dataset includes 9,963 images split into train/test sets which separately include 5,011 and 4,952 images. We use our pretrained image communication system to transmitting the test set and use the mature detection models to measure the mAP.

2) **Network and Hyper-parameters:** The specific settings about various networks can be found in Table I. Note that the JSCC encoder performs downsampling while implementing the convolutional operation by setting the stride to different number, in the decoder, upsampling is performing by depth-to-space operation. Besides, in the forward propagation, the rounding operation is executed in the binary quantizer, where \([·]\) is defined as rounding. In addition, The network structures of classifier and objection detection are default settings of ResNet-18 and RFBNet, exception the fully connected layer in the ResNet-18, which is modified to 10 dimensions to adapt the practical task. Finally, the MI model is a variational network, which is to approximate the mappings between two variables.

The training parameters in algorithm 1 to 4 are shown in Table II. Note that TOSC-SR is trained based on the separate classifier and JSCC codec, so the learning rate is lower. The batch size of MI model is set to 200 due to this model need rich data to reduce the estimation error. The trade-off factor \( \beta \) and SNR only exist in our proposed TOSC-SR.

#### B. Comparison Experiments

1) **Performance versus bpp:** Fig.\( 4 \) shows the comparisons of different methods with respect to the performance of classification and reconstruction under different bpp with SNR = \( \infty \), \( \beta = 0.01 \). In order to avoid the model’s fluctuation, we took the average of the final five classification accuracy rates, PSNRs and SSIMs of validation set. The classification accuracy
for original images is 94.61%. In the following analysis, we jointly consider the classification accuracy and reconstruction quality.

We first compare the proposed TOSC-SR with traditional methods (JPEG and JPEG2000). Fig. 4(a) shows that the TOSC-SR outperforms the JPEG and JPEG2000 CS in terms of classification accuracy under all bpp settings, especially in the low bpp. Fig. 4(b) and (c) show that the TOSC-SR is superior or competitive to traditional methods in terms of the reconstruction quality, which is consistent with the conclusions of most DNN-based image compression methods [8]–[10]. Essentially, better reconstruction quality leads to better classification accuracy.

Then compare the proposed TOSC-SR with SPICS [4]. Fig. 4 shows that TOSC-SR outperforms SPICS both in terms of classification and reconstruction performance under the same CAE networks and hyperparameters. These improvements should be attributed to the joint consideration of RD distortion and IB distortion. SPICS is actually a multi-task learning model, which results that the performance in different tasks may compete due to the different learning objectives. While thanks to the cascade structure in Fig. 3, our proposed TOSC-SR don’t suffer from this problem. In general, the task performance is almost positively related to the quality of reconstruction. Besides, cascade structure is more flexible because of the networks of JSCC codec and task are decouple. So we can take advantage of various mature pre-trained task networks, i.e., making full use of transfer learning by fine-tuning these pre-trained networks, which is impossible for the structure in [4].

Observing the comparisons between TOSC-SR and two baselines, Fig. 4 shows that IB baseline attained the best classification performance, while the quality of reconstruction drops a lot compared with RD baseline and TOSC-SR, which will weaken the generalization ability among different tasks, i.e., these reconstructed images are unable to perform other tasks well. This will be further discussed in the later experiments. On the other hand, observing the RD baseline. Though it achieves the best quality of reconstruction, the classification performance is inferior to TOSC-SR, which support our hypothesis that perceptual-significant visual features may not be the most suitable for AI task. The TOSC-SR achieved a ideal trade-off between predictive precision and generalization ability compared to RD baseline by improving the reconstructed image’s predictive precision in classification task at the expense of a small amount of perception quality.

![Fig. 4. Comparisons of different methods with respect to the performance of classification and reconstruction under different bpp with SNR = ∞, β = 0.01. The performance of classification is assessed by classification accuracy, and the performance of reconstruction is assessed by PSNR and SSIM.](image)

2) Visualization of semantic reconstruction: Fig. 5 shows the visualization results of reconstructed images using RD Baseline and TOSC-SR schemes respectively and the performance comparison of classification tasks using them. Intuitively, there are some differences in the images reconstructed by the two schemes. In general, the images reconstructed by the proposed TOSC-SR method are more orderly and the contour of the target object is more distinct. These may be the reasons for the improved accuracy of the final classification. Specific feature information is difficult to see directly from the appearance of the picture. In later experiments, we will try to explain this phenomenon from information theory perspective by estimating the corresponding mutual information.

3) Performance versus SNR: Fig. 6 shows the comparisons of different methods with respect to the performance of classification and reconstruction under different SNR with bpp = 2, β = 0.01. Here we compare the TOSC-SR with traditional methods and two baselines.

Observing the performance of JPEG CS and JPEG2000 CS, there solutions appers abrupt performance degradations in certain SNR, known as the "cliff effect" in digital communication. The "cliff effect" means that the performance falls off steeply until to the worst levels when the channel condition is lower than a certain threshold. As shown in the [17], DNN-
based JSCC scheme can overcome the "cliff effect". Here, our proposed TOSC-SR also don’t suffer from it, specifically, as the SNR decreases, the classification and construction performances of TOSC-SR decrease slowly.

Comparing TOSC-SR with RD and IB baselines. Fig.6(a) shows that the TOSC-SR is superior to RD baseline but inferior to IB baseline in terms of classification performance. And the IB baseline is exceeded by the TOSC-SR in the high SNRs. On the contrary, Fig.6(b) and (c) show that the TOSC-SR is superior to IB baseline by a significant margin but slightly inferior to RD baseline in terms of reconstruction performance. This phenomenon implies that, once again, the TOSC-SR achieved a ideal trade-off between predictive precision and generalization ability.

C. Mutual Information Estimation Results

Why is it that images with best visual perception are not the most suitable for classification task? This can be to explained from the perspective of information theory.

According to the previous analysis, random variables Z and \(\hat{Y}\) are two critical variables in the whole end-to-end communication pipeline. In this section we mainly estimate these MI values among TOSC-SR, RD baseline and IB baseline. In line with information bottleneck method, we estimated the mutual information between intermediate representations \(Z, \hat{Y}\) and source inputs X and the mutual information between intermediate representations \(Z, \hat{Y}\) and desired task output Y, namely \(I(X;Z), I(Z;Y), I(X;\hat{Y}),\) and \(I(\hat{Y};Y)\). The physical meanings of these four MI are described below.

\(I(X;Z)\), the mutual information of the JSCC codec’s information bottleneck variable Z and the input X, represents the amount of information relevant to X contained in Z. This value shows that, on the one hand, how much information relevant to X is retained after encoding. On the other hand, it reflects the compression of Z with respect to X. The smaller this MI value, the greater compression degree.

\(I(Z;Y)\), the mutual information of the auto-encoder’s information bottleneck variable Z about the desired output Y, represents the amount of information about Y contained in Z, which reflects the potential of Z to perform the task. The greater this MI value, the more information about the label Y captured by Z.

\(I(X;\hat{Y})\), the mutual information of classifier’s bottleneck variable \(\hat{Y}\) and input X, represents the amount of information about X contained in \(\hat{Y}\).

\(I(\hat{Y};Y)\), the mutual information of the classifier’s information bottleneck variable \(\hat{Y}\) and desired output Y, represents the amount of information relevant to Y contained in \(\hat{Y}\), which reflects the potential of \(\hat{Y}\) to perform the task. The larger the value is, the more information about Y is captured by \(\hat{Y}\).

Fig.(7) shows the comparisons of different methods in terms of four mutual information. Since the CLUB method produces different values within a certain range each estimation, we estimate mutual information multiple times and compute the maximal, minimal and mean values. Then we draw the region bounded by the maximal and minimal values in light colors and plot the corresponding mean values in dark colors. These estimating results yield many findings listed as following.

- The loss functions we actually used in different methods have close relationships with mutual information. From Fig.(7) (a) we can find that, RD baseline, adopted the MSE-only objective, produced the maximal \(I(X;Z)\), since minimizing MSE\((X; \hat{X})\) is equivalent to maximizing \(I(X;Z)\). TOSC-SR, optimized for MSE and \(D_{IB}\) trade-off objective, exhibited a slightly smaller \(I(X;Z)\), since the influence of \(D_{IB}\). IB baseline showed the minimal \(I(X;Z)\) because the \(D_{IB}\) is the only objective, and minimizing \(D_{IB}\) means maximizing \(I(\hat{Y};Y)\), which is different from \(I(X;Z)\). On the contrary, from Fig.(7) (d) we can find that, RD baseline produced the minimal \(I(\hat{Y};Y)\), IB baseline produced the maximal values, and the values of TOSC-SR is slightly lower than IB baseline. This phenomenon is also resulted by the different settings of optimization objectives.

- The changes of mutual information result in the changes of performance of classification and reconstruction. Observing the changes of mutual information in Fig.(7) (a) (d) and performance of different tasks, we can find that, they have the consistent trend. The maximal \(I(X;Z)\) and minimal \(I(\hat{Y};Y)\) of RD baseline correspond to the best reconstruction performance and worst classification performance. The minimal \(I(X;Z)\) and maximal \(I(\hat{Y};Y)\) of IB baseline correspond to the worst reconstruction performance and best classification performance. Finally, the median MI values of TOSC-SR correspond to medium performance of different tasks. These results show that...
proper trade-off between different optimization objectives can reach both satisfactory performance of various tasks.

- Fig. 7(b) shows the comparisons of different methods in terms of \(I(Z; Y)\). \(I(Z; Y)\) is not included in our optimization objective. From the results we can find that the three methods have almost the same values of \(I(Z; Y)\), and the values is located around 3.51. Note that \(I(Z; Y) = H(Y) - H(Y|Z) \leq H(Y) = \log_2(10) = 3.3219\). So the estimations of \(I(Z; Y)\) have saturated. It is bigger than \(\log_2 10\) because the CLUB produced the upper bound of mutual information. Since the most of information contained in \(Z\) is relevant to \(X\), though \(I(Z; Y)\) is large enough, the redundant information about \(X\) will affect the classification performance.

- Fig. 7(c) shows the comparisons of different methods in terms of \(I(X; \hat{Y})\). \(I(X; \hat{Y})\) is also not included in our optimization objective. And \(\hat{Y}\) is the bottleneck variable of classification network. Through the classification network, \(\hat{Y}\) tried to discard as much information relevant to \(X\) as possible, and retain as much information relevant to \(Y\) as possible. So \(I(X; \hat{Y})\) in different methods have the similar variation trends.

If the distorted description is merely \(D_{RD}\), which is actually MSE here, then this pixel-by-pixel approach is equivalent to preserving as much feature information as possible of the original image without discriminating and any trade-off. However, we believe that in the case of limited bit rate, if the image is still recovered with MSE as the target, it is not beneficial to the execution of downstream AI tasks. In specific, the network is no longer able to recover the original image well in the low bit rate. If you insist on recovering the image pixel-by-pixel, it will become less recognizable.

If the description of distortion is only \(D_{IB}\), which is, the loss of mutual information \(I(\hat{X}; Y)\) is taken as the distortion in the way of information bottleneck, this is equivalent to a certain selection of recovered features. In this way, the recovered \(\hat{X}\) will retain as much as possible of the feature information relevant to corresponding task. Although this is good for the task, the structural information of the original image is completely destroyed, making it impossible to perform other tasks, especially those tasks need the structural information, and in fact, it is almost impossible to reconstruct the a meaningful image for human eyes.

According to our method, the loss of two kinds of distortion, MSE and mutual information \(I(\hat{X}; Y)\), are jointly considered, so that the task-related feature information can be selectively recovered, and the generalization ability can be retained. In the case of limited bit rate, it not only ensures the performance of completing the task, but also enables the network to reconstruct the image with slightly reduced quality. At this time, the reconstructed image has already contained more feature information related to the task, which is called semantic reconstruction.
D. Generalization Performance

Fig. 8 shows the generalization performance of different methods from classification to object detection. Note that RFBNet’s mAP on the original Pascal VOC 2007 test dataset is 80.67%. Observing Fig. 8 (a) we can find that, in contrast to the classification task case, here TOSC-SR shows the best performance among these three methods. IB baseline is even inferior to RD baseline in terms of mAP. Fig. 8 (b) and (c) show the similar relationships as in previous basic experiments in terms of reconstruction quality.

According the above results we can believe that, the trade-off factor $\beta$ plays a critical role in ensuring good generalization ability. Though IB baseline exhibited the better classification performance than TOSC-SR in the previous experiments, the following experiment results will show that, the images reconstructed by IB baseline has poor generalization on a different task, which means that when the JSCC codec is transferred to another system with any other tasks, its performance will drop quickly.

On the other hand, the performance improvement of object detection is not as large as that of classification case, which increases with the decrease of bpp. This is because what we did was just a simple generalization ability test, and the model trained by STL10 classification was directly applied to the object detection task. At this time, there was no joint training between the JSCC codec network and the object detection network, and there may exist a mismatching between the two networks (The distribution of the data reconstructed by the JSCC codec has changed, while the object detection network is a fixed model after the training of the original image data set). If we combine the JSCC codec and the object detection network for fine-tuning training, the performance may be further improved. That’s our future work.

When the JSCC codec trained by the classification task are applied to the scene of object detection, the performance is also improved steadily. This is surprising and remarkable, and it can be explained. When we are training the basic semantic image transmission model, while using the distortion and MSE weighted and classification task, but the classification task itself is a more basic intelligence task, there is reason to believe that this is because the classification distortion makes the model in the reconstructed image and keep the semantic features of the image information as much as possible, which leads to the improvement of object detection performance.

Consequently, if the target task in the receiver is determined, we’d better use the IB baseline. But in practice, the receiver has to handle multiple tasks, in order to ensure that the reconstructed images can perform these tasks well simultaneously, TOSC-SR is the better choice.

E. Impact of Trade-off Factor $\beta$

Essentially, the only difference between TOSC-SR, RD baseline and IB baseline is $\beta$, which is the trade-off factor between $D_{RD}$ and $D_{IB}$. In the previous experiments, $\beta$ in TOSC-SR is fixed to 0.01 for proper performances. What’s more, we have intuitively seen the powerful effect of $\beta$, and Fig. 8 shows that, IB baseline is inferior to TOSC-SR in terms of both performance of AI task and reconstruction, which shows that IB baseline has poor generalization performance. From RD baseline to IB base line, $\beta$ is monotonically increasing. In this section, we explore the dynamics of performance as $\beta$ increases from 0 to $\infty$.

Fig. 9 shows the performance’s dynamics of source and target tasks in terms of trade-off factor $\beta$. We first observe the source task’s dynamics. Fig. 9(a) shows that the classification accuracy gets improved as $\beta$ increases, and the growth rate is getting slower and slower. Meanwhile, this promotion is limited by the case of $\beta = \infty$, from our training curves, the accuracy has reached the plateau. Fig. 9(b) and (c) show that the quality of reconstruction is getting worse and worse as $\beta$ increases, and it’s going down faster. Meanwhile, this deterioration is unlimited, from the training curves, the perceptual quality will go further down as the training iterations increase. Unlike the classification task, Fig. 9(a) shows that the mAP of object detection task exhibits a trend of increase first and then decrease. Fig. 9(b) and (c) show that both tasks perform similar dynamics of reconstruction performance.

There are several interesting phenomenon need to explain. First, source task and target task show different dynamics in terms of precision, the reason is that we directly transferred the system trained on classification task to object detection task without further training the JSCC codec for object detection. When $\beta$ increases at first, our JSCC codec can extract and remain the general semantic features, which is beneficial to various AI tasks. But as $\beta$ goes up beyond a threshold, the JSCC codec will extract and remain the features specific to classification tasks, which results the degradation of performance of...
target task, i.e. mAP. Second, the reconstruction performance of Pascal VOC2007 is always inferior to STL10. This is also because we train the TOSC-SR with STL10 dataset, and these two datasets differ greatly in both image size and style. We believe the performance will get improved if the Pascal VOC dataset is used to train the TOSC-SR.

VI. Conclusion

In this paper, we proposed a TOSC-SR system based on the extended rate-distortion theory, which is jointly optimized for the predicting precision and generalization ability among different AI tasks of the reconstructed images. Specifically, we extended the rate-distortion theory by taking the trade-off between reconstruction distortion and task-related distortion as the semantic distortion, and further designed the semantic communication system based on the widely used JSCC codec architecture followed by the task network. We also noted that, the semantic distortion we define here is closely related to mutual information, which means that our semantic communication system can be explained from the information theoretic view. The simulation results have shown that the proposed method outperforms the traditional and DNN-based methods in both classification and object detection tasks under different bpp and SNR. And the mutual information estimation results show that our TOSC-SR indeed extract more semantic information. Therefore our proposed method can be useful in IoT communication with various AI tasks need to perform.

APPENDIX A

PROOF OF THEOREM 1

Denote \( d_{IB}(x, \hat{x}) \) as the relevant information distortion between \( x \) and \( \hat{x} \),
\[
d_{IB}(x, \hat{x}) = \sum_y p(y|x) \log \left( \frac{p(y|x)}{p(y|\hat{x})} \right). \tag{41}
\]

Denote \( D_{IB}(X, \hat{X}) \) as the expectation of relevant information distortion between \( X \) and \( \hat{X} \),
\[
D_{IB}(X, \hat{X}) = E[d_{IB}(x, \hat{x})] = \sum_{\hat{x}} \sum_x \sum_y p(x, \hat{x}) p(y|x) \log \left( \frac{p(y|x)}{p(y|\hat{x})} \right). \tag{42}
\]

We first prove conditional mutual information \( I(X; Y | \hat{X}) \) does describe the relevant information distortion. According to the markov chain \( Y \rightarrow X \rightarrow \hat{X} \),
\[
p(y|x, \hat{x}) = p(y|x), \tag{43}
\]
according to the definition of conditional mutual information, and implement equation \( 43 \)
\[
I(X; Y | \hat{X}) = \sum_{\hat{x}} \sum_x \sum_y p(x, \hat{x}) p(y|x, \hat{x}) \log \left( \frac{p(x, y|\hat{x})}{p(x|\hat{x}) p(y)} \right)
= \sum_{\hat{x}} \sum_x \sum_y p(\hat{x}) p(x|\hat{x}) p(y|x, \hat{x}) \log \left( \frac{p(x, y|\hat{x})}{p(x|\hat{x}) p(y)} \right)
= \sum_{\hat{x}} \sum_x \sum_y p(\hat{x}) p(x|\hat{x}) p(y|x) \log \left( \frac{p(y|x)}{p(y)} \right)
\tag{44}
\]

Then we prove
\[
I(X; Y) - I(\hat{X}; Y) = I(X; Y | \hat{X}), \tag{45}
\]
According to the definition of mutual information,
\[
I(X; Y) - I(\hat{X}, Y) = \sum_{x, y} p(x, y) \log \left( \frac{p(x, y)}{p(x)p(y)} \right)
- \sum_{\hat{x}} \sum_x \sum_y p(\hat{x}, y) \log \left( \frac{p(\hat{x}, y)}{p(\hat{x})p(y)} \right)
= \sum_{\hat{x}} \sum_x \sum_y p(\hat{x}, x, y) \log \left( \frac{p(x, y)}{p(x)p(y)} \right)
- \sum_{\hat{x}} \sum_x \sum_y p(\hat{x}, x, y) \log \left( \frac{p(\hat{x}, y)}{p(\hat{x})p(y)} \right) \tag{46}
\]
\[
= \sum_{\hat{x}} \sum_x \sum_y p(\hat{x}, x, y) \log \left( \frac{p(y|x)}{p(y)} \right)
- \sum_{\hat{x}} \sum_x \sum_y p(\hat{x}, x, y) \log \left( \frac{p(y|\hat{x})}{p(y)} \right)
= D_{IB} = E[d_{IB}(x, \hat{x})].
\]
Rewrite the optimization problem (11) as follows:

\[
\min_{p(\hat{x}|x): I(x, \hat{x}) \leq I_C} D_{RD}(X, \hat{X}) - \beta I(\hat{X}; Y). \tag{47}
\]

Using the Lagrange multiplier method by first formulating the Lagrange functional,

\[
L(p(\hat{x}|x)) = \alpha I(x, \hat{x}) + D_{RD}(X, \hat{X}) - \beta I(\hat{X}; Y) \\
+ \sum_x r(x) \sum_{\hat{x}} p(\hat{x}|x)
\]

\[
= \alpha \sum_x \sum_{\hat{x}} p(x)p(\hat{x}|x) \log \frac{p(\hat{x}, x)}{p(\hat{x})} \\
+ \sum_x \sum_{\hat{x}} p(x)p(\hat{x}|x)d_{RD}(x, \hat{x})
\]

\[-\beta \sum_y \sum_{\hat{x}} p(y)p(\hat{x}|y) \log \frac{p(\hat{x}|y)}{p(\hat{x})} \\
+ \sum_x r(x) \sum_{\hat{x}} p(\hat{x}|x),
\]

where \(\alpha\) is the Lagrange multiplier attached to the constrained \(I(x, \hat{x})\), \(r(x)\) is the Lagrange multiplier attached to the normalization of the mapping \(p(\hat{x}|x)\) corresponding to \(x\).

First we note that the distribution \(p(y|x)\) is given as

\[
p(y|x) = \frac{1}{p(\hat{x})} \sum_{\hat{x}} p(y|x)p(\hat{x}|x)p(x).
\tag{49}
\]

According to the Markov chain \(Y \leftrightarrow X \leftrightarrow \hat{X}\), we get \(p(\hat{x}|y) = p(\hat{x}|x)\) and \(p(y|x, \hat{x}) = p(y|x)\), then

\[
p(\hat{x}) = \sum_x p(x)p(\hat{x}|x), \tag{50}
\]

\[
p(\hat{x}|y) = \sum_{\hat{x}} p(\hat{x}, x|y) = \sum_{\hat{x}} p(\hat{x}|x, y)p(x|y) = \sum_{\hat{x}} p(\hat{x}|x)p(x|y). \tag{51}
\]

from equations (49) and (50), we can get the derivatives w.r.t. \(p(\hat{x}|x)\)

\[
\frac{\partial p(\hat{x})}{\partial p(\hat{x}|x)} = p(x), \tag{52}
\]

\[
\frac{\partial p(\hat{x}|y)}{\partial p(\hat{x}|x)} = p(x|y). \tag{53}
\]

Taking derivatives of (11) with respect to \(p(\hat{x}|x)\) for given \(x\) and \(\hat{x}\)

\[
\frac{\partial L}{\partial p(\hat{x}|x)} = \alpha \log \frac{p(\hat{x}|x)}{p(\hat{x})} + \alpha \log \frac{p(\hat{x}|x)}{p(\hat{x})} - \frac{1}{p(\hat{x}|x)} \\
- \alpha \sum_x p(x) \log \frac{p(\hat{x}|x)}{p(\hat{x})} + \alpha \log \frac{p(\hat{x}|x)}{p(\hat{x})}
\]

\[-\beta \sum_y \sum_{\hat{x}} p(y)p(\hat{x}|y) \log \frac{p(\hat{x}|y)}{p(\hat{x})} + \beta \sum_y \sum_{\hat{x}} p(y)p(\hat{x}|y) \log \frac{p(\hat{x}|y)}{p(\hat{x})} + r(x), \tag{54}
\]

substitute (52) and (53) into (54)

\[
\frac{\partial L}{\partial p(\hat{x}|x)} = \alpha \log \frac{p(\hat{x}|x)}{p(\hat{x})} + \alpha \log \frac{p(\hat{x}|x)}{p(\hat{x})} - \frac{1}{p(\hat{x}|x)} \\
- \alpha \sum_x p(x) \log \frac{p(\hat{x}|x)}{p(\hat{x})} + \alpha \log \frac{p(\hat{x}|x)}{p(\hat{x})}
\]

\[-\beta \sum_y \sum_{\hat{x}} p(y)p(\hat{x}|y) \log \frac{p(\hat{x}|y)}{p(\hat{x})} + \beta \sum_y \sum_{\hat{x}} p(y)p(\hat{x}|y) \log \frac{p(\hat{x}|y)}{p(\hat{x})} + r(x), \tag{55}
\]

rearranging this equation

\[
\frac{\partial L}{\partial p(\hat{x}|x)} = \alpha \log \frac{p(\hat{x}|x)}{p(\hat{x})} + \alpha \log \frac{p(\hat{x}|x)}{p(\hat{x})} - \frac{1}{p(\hat{x}|x)} \\
- \beta \sum_y \sum_{\hat{x}} p(y)p(\hat{x}|y) \log \frac{p(\hat{x}|y)}{p(\hat{x})} + \beta \sum_y \sum_{\hat{x}} p(y)p(\hat{x}|y) \log \frac{p(\hat{x}|y)}{p(\hat{x})} + r(x), \tag{56}
\]

Notice that \(\sum_x p(y|x) \log \frac{p(y|x)}{p(y)} = I(x; Y)\) is a function of \(x\) only (independent of \(\hat{x}\), and thus can be absorbed into the multiplier \(r(x)\)). Introducing

\[
\log \mu(x) = \frac{1}{\lambda} \left( r(x) - \beta \sum_y p(y|x) \log \frac{p(y|x)}{p(y)} \right), \tag{57}
\]

finally, (56) can be simplified to

\[
\frac{\partial L}{\partial p(\hat{x}|x)} = \alpha \log \frac{p(\hat{x}|x)}{p(\hat{x})} + \alpha \log \frac{p(\hat{x}|x)}{p(\hat{x})} - \frac{1}{p(\hat{x}|x)} \\
- \beta \sum_y \sum_{\hat{x}} p(y|x) \log \frac{p(y|x)}{p(\hat{x})} + \beta \sum_y \sum_{\hat{x}} p(y|x) \log \frac{p(y|x)}{p(\hat{x})} + r(x) \tag{58}
\]

\[
= \alpha \log \frac{p(\hat{x}|x)}{p(\hat{x})} + \alpha \log \frac{p(\hat{x}|x)}{p(\hat{x})} - \frac{1}{p(\hat{x}|x)} \\
- \beta \sum_y \sum_{\hat{x}} p(y|x) \log \frac{p(y|x)}{p(\hat{x})} + \beta \sum_y \sum_{\hat{x}} p(y|x) \log \frac{p(y|x)}{p(\hat{x})} + r(x) \tag{59}
\]

we know that

\[
d_{W}(x, \hat{x}) = D_{KL}[p(y|x)||p(y|x)] = \sum_y p(y|x) \log \frac{p(y|x)}{p(y|x)}.
\]

\[
d_{W}(x, \hat{x}) = D_{KL}[p(y|x)||p(y|x)] = \sum_y p(y|x) \log \frac{p(y|x)}{p(y|x)}.
\]

\[
d_{W}(x, \hat{x}) = D_{KL}[p(y|x)||p(y|x)] = \sum_y p(y|x) \log \frac{p(y|x)}{p(y|x)}.
\]

\[
d_{W}(x, \hat{x}) = D_{KL}[p(y|x)||p(y|x)] = \sum_y p(y|x) \log \frac{p(y|x)}{p(y|x)}.
\]
and set (58) to zero,

\[ p(\hat{x}|x) = \frac{p(\hat{x})e^{-\lambda d_s(x, \hat{x})}}{\mu(x)}, \quad (60) \]

and

\[ d_s(x, \hat{x}) = d_{RD}(x, \hat{x}) + \beta d_{IB}(x, \hat{x}), \quad (61) \]

Since \( \sum_{x} p(\hat{x}|x) = 1 \), substitute it to (60)

\[ \mu(x) = \sum_{x} p(\hat{x})e^{-\lambda d_s(x, \hat{x})}. \quad (62) \]

**APPENDIX C**

**EXAMPLES OF SEMANTIC RECONSTRUCTION WITH DIFFERENT \( \beta \)**

Fig. 10 shows the examples of reconstructed image at different \( \beta \) settings. As we can see, with the increase of \( \beta \), although the reconstructed image has a certain degree of distortion in appearance, the semantic information of the overall contour and spatial combination mode are still retained, and the retained semantic information is often enough to complete the subsequent AI tasks.

**REFERENCES**

[1] C. E. Shannon, “A mathematical theory of communication,” The Bell system technical journal, vol. 27, no. 3, pp. 379–423, 1948.

[2] E. C. Strinati and S. Barbarossa, “6G networks: Beyond shannon towards semantic and goal-oriented communications,” Comm. Com. Inf. Sc., vol. 190, p. 1079–190, 2021.

[3] 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.

[4] N. Patwa, N. Ahuja, S. Somayazulu, O. Tickoo, S. Varadarajan, and S. Koolagudi, “Semantic-preserving image compression,” in IEEE Int. Conf. Image Process. (ICIP), Abu Dhabi, United Arab Emirates, Oct. 2020, pp. 1281–1285.

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

[6] G. K. Wallace, “The JPEG still picture compression standard,” IEEE transactions on consumer electronics, vol. 38, no. 1, pp. xvi–xxiv, 1992.

[7] M. Rabbani and R. Joshi, “An overview of the JPEG 2000 still image compression standard,” Signal processing: Image communication, vol. 17, no. 1, pp. 3–48, 2002.

[8] J. Ballé, V. Laparra, and E. P. Simoncelli, “End-to-end optimized image compression,” in Proc. Int. Conf. Learn. Representations (ICLR), Sain Juan, Puerto Rico, May 2016.

[9] L. Theis, W. Shi, A. Cunningham, and F. Huszár, “Lossy image compression with compressive autoencoders,” in Proc. Int. Conf. Learn. Representations (ICLR), Vancouver, Canada, Apr. 2018.

[10] G. Toderici, S. M. O’Malley, S. J. Hwang, D. Vincent, D. Minnen, S. Baluja, M. Covell, and R. Sukthankar, “Variable rate image compression with recurrent neural networks,” in Proc. Int. Conf. Learn. Representations (ICLR), Toulon, France, Apr. 2017.

[11] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, “Variational image compression with a scale hyperprior,” in Proc. Int. Conf. Learn. Representations (ICLR), Vancouver, Canada, Apr. 2018.

[12] Z. Cheng, H. Sun, M. Takeuchi, and J. Katto, “Learned image compression with discretized gaussian mixture likelihoods and attention modules,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit (CVPR), Seattle, United States, Jun. 2020, pp. 7939–7948.

[13] Z. Yang, Y. Wang, C. Xu, D. Pu, C. Xu, C. Xu, and Q. Tian, “Discernible image compression,” in Proc. 28th ACM Int. Conf. Multimedia (ACMMM), Westminster, United States, Oct. 2020, pp. 1561–1569.

[14] N. Tishby, F. C. Pereira, and W. Bialek, “The information bottleneck method,” in Proc. 37th Annual Allerton Conf. Commun. Control Comput., Allerton House, Monticello, Illinois, Sep. 1999.

[15] Y. Yang, C. Guo, F. Liu, C. Liu, L. Sun, Q. Sun, and J. Chen, “Semantic communications with ai tasks,” 2021, arXiv:2109.14170. [Online]. Availabe: https://arxiv.org/abs/2109.14170.

[16] S. Luo, Y. Yang, Y. Yin, C. Shen, Y. Zhao, and M. Song, “DeepSIC: Deep semantic image compression,” in Proc. Int. Conf. on Neural Inf. Process. (ICONIP), Siem Reap, Cambodia, Dec. 2018, pp. 96–106.

[17] E. Bourtsoulatze, D. B. Kurka, and D. Güntüz, “Deep joint source-
channel coding for wireless image transmission,” IEEE Trans. on Cogn. Commun. Netw., vol. 5, no. 3, pp. 567–579, 2019.

[18] Z. Qin, X. Tao, J. Lu, and G. Y. Li, “Semantic communications: Principles and challenges,” 2021, arXiv:2201.01389. [Online]. Availabe: https://arxiv.org/abs/2201.01389.

[19] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, “Deep variational information bottleneck,” Apr. 2017.

[20] P. Zhang, W. Xu, H. Gao, K. Niu, X. Xu, X. Qin, C. Yuan, Z. Qin, H. Zhao, J. Wei et al., “Toward wisdom-evolutionary and primitive-concise 6G: A new paradigm of semantic communication networks,” Eng., vol. 8, pp. 60–73, 2022.

[21] R. Carnap, Y. Bar-Hillel et al., “An outline of a theory of semantic information,” 1952.

[22] J. Barwise and J. Perry, “Situations and attitudes,” J. Philos., vol. 78, no. 11, pp. 668–691, 1981.

[23] L. Floridi, “Outline of a theory of strongly semantic information,” Minds and machines, vol. 14, no. 2, pp. 197–221, 2004.

[24] J. Chai, H. Zeng, A. Li, and E. W. Ngai, “Deep learning in computer vision: A critical review of emerging techniques and application scenarios,” Mach. Learn. Appl., vol. 6, p. 100134, 2021.

[25] L. A. Zadeh, “Probability measures of fuzzy events,” J. Math. Anal., vol. 23, no. 2, pp. 421–427, 1968.

[26] Z. Weng, Z. Qin, and G. Y. Li, “Semantic communications for speech recognition,” 2021, arXiv:2107.11190. [Online]. Availabe: https://arxiv.org/abs/2107.11190.

[27] N. Tishby and N. Zaslavsky, “Deep learning and the information bottleneck,” Apr. 2017. 

[28] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, J. Courville, and D. Hjelm, “Mutual information neural estimation,” in Proc. Adv. Neural Inf. Process. Syst. (NIPS), 2015, pp. 1–5.

[29] P. Cheng, W. Hao, S. Dai, J. Liu, Z. Gan, and L. Carin, “CLUB: A contrastive log-ratio upper bound of mutual information,” in Proc. Int. Conf. Mach. Learning (ICML), Stockholm, Sweden, Jun. 2018, pp. 531–540.

[30] H. Xie, Z. Qin, and G. Y. Li, “Task-oriented semantic communication systems for speech recognition,” 2021, arXiv:2112.10255. [Online]. Availabe: https://arxiv.org/abs/2112.10255.

[31] Z. Weng and Z. Qin, “Semantic communication systems for speech transmission,” IEEE J. Sel. Areas Commun., vol. 39, no. 8, pp. 2344–2444, 2021.

[32] J. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, “Masked autoencoders are scalable vision learners,” 2021, arXiv:2111.06377. [Online]. Availabe: https://arxiv.org/abs/2111.06377.

[33] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, “Overview of the high efficiency video coding (HEVC) standard,” IEEE Trans. Circuits Syst. Video Technol., vol. 22, no. 12, pp. 1649–1668, 2012.

[34] M. Li, W. Zhao, S. Gu, D. Zhao, and D. Zhang, “Learning convolutional networks for content-weighted image compression,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Salt Lake City, USA, Jun. 2018, pp. 3214–3223.

[35] C. Cai, L. Chen, X. Zhang, and Z. Gao, “End-to-end optimized ROI image compression,” IEEE Trans. Image Process., vol. 29, pp. 3442–3457, 2019.

[36] R. Carnap, Y. Bar-Hillel et al., “An outline of a theory of semantic information,” 1952.

[37] I. Sandryhaila and J. Mouron, “Laplacian matrix as a graph filter,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Salt Lake City, USA, Jun. 2018, pp. 3214–3223.

[38] J. Chai, H. Zeng, A. Li, and E. W. Ngai, “Deep learning in computer vision: A critical review of emerging techniques and application scenarios,” Mach. Learn. Appl., vol. 6, p. 100134, 2021.

[39] J. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, “Masked autoencoders are scalable vision learners,” 2021, arXiv:2111.06377. [Online]. Availabe: https://arxiv.org/abs/2111.06377.

[40] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, “Overview of the high efficiency video coding (HEVC) standard,” IEEE Trans. Circuits Syst. Video Technol., vol. 22, no. 12, pp. 1649–1668, 2012.

[41] M. Li, W. Zhao, S. Gu, D. Zhao, and D. Zhang, “Learning convolutional networks for content-weighted image compression,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Salt Lake City, USA, Jun. 2018, pp. 3214–3223.

[42] C. Cai, L. Chen, X. Zhang, and Z. Gao, “End-to-end optimized ROI image compression,” IEEE Trans. Image Process., vol. 29, pp. 3442–3457, 2019.