Nonlinear Transform Source-Channel Coding for Semantic Communications

Jincheng Dai, Member, IEEE, Sixian Wang, Student Member, IEEE, Kailin Tan, Student Member, IEEE, Zhongwei Si, Member, IEEE, Xiaoqi Qin, Member, IEEE, Kai Niu, Member, IEEE, and Ping Zhang, Fellow, IEEE

Abstract—In this paper, we propose a class of high-efficiency deep joint source-channel coding methods that can closely adapt to the source distribution under the nonlinear transform, it can be collected under the name nonlinear transform source-channel coding (NTSCC). In the considered model, the transmitter first learns a nonlinear analysis transform to map the source data into latent space, then transmits the latent representation to the receiver via deep joint source-channel coding. Our model incorporates the nonlinear transform as a strong prior to effectively extract the source semantic features and provide side information for source-channel coding. Unlike existing conventional deep joint source-channel coding methods, the proposed NTSCC essentially learns both the source latent representation and an entropy model as the prior on the latent representation. Accordingly, novel adaptive rate transmission and hyperprior-aided codec refinement mechanisms are developed to upgrade deep joint source-channel coding. The whole system design is formulated as an optimization problem whose goal is to minimize the end-to-end transmission rate-distortion performance under established perceptual quality metrics. Across test image sources with various resolutions, we find that the proposed NTSCC transmission method generally outperforms both the analog transmission using the standard deep joint source-channel coding and the classical separation-based digital transmission. Notably, the proposed NTSCC method can potentially support future semantic communications due to its content-aware ability and perceptual optimization goal.

Index Terms—Semantic communications, nonlinear transform, joint source-channel coding, rate-distortion, perceptual loss.

I. INTRODUCTION

Semantic communications are recently emerging as a new paradigm driving the in-depth integration of information and communication technology (ICT) advances and artificial intelligence (AI) innovations [1]. This communication paradigm shifting calls for revolutionary theories and methodology innovations. Unlike traditional module-stacking system design philosophy, it is the very time to bridge the two main branches of Shannon information theory [2], i.e., source and channel parts, together for boosting the end-to-end system capabilities. By this means, the channel transmission process can be aware of the source semantic features.

The paradigm aiming at the integrated design of source and channel codes was named joint source-channel coding (JSCC) [3], which is a classical topic in the information theory and coding theory. However, classical JSCC schemes [3]–[6] were based on the statistical probabilities without considering source semantic aspects. By introducing AI, JSCC is expected to evolve to a modern version. On the one hand, source coding can intelligently extract the most valuable information for human understanding in intelligent human-type communications (IHTC) and decision making in intelligent machine-type communications (IMTC) such that the source coding rate can be efficiently reduced. On the other hand, channel coding can precisely identify the critical part of the source coded sequence to achieve semantics-biased unequal protection.

As an modern version, recent deep learning methods for realizing JSCC have stimulated significant interest in both AI and wireless communication communities [7]–[13]. By using artificial neural networks (ANN), the problem of JSCC can be carried out over analog channels, where transmitted signals are formatted as continuous-valued symbols. From Shannon’s perspective [2], deep JSCC can be viewed as a geometric mapping of a vector in the source space onto a lower-dimensional space embedded in the source space and transmitted over the noisy channel. Similar to the landmark works of Gündüz [9]–[11], we also take the image source as a representative in this paper, but our work is extensible for other source modalities. Standard deep learning JSCC schemes operate on a simple principle: an image, modeled as a vector of pixel intensities \( x \in \mathbb{R}^m \), is mapped to a vector of continuous-valued channel input symbols \( s \in \mathbb{R}^k \) via an ANN-based encoding function. We typically have \( k < m \), and \( R = k/m \) is named channel bandwidth ratio (CBR) [10] denoting the coding rate. The decoder attempts to recover the source from the channel corrupted sequence \( s \). This deep JSCC method yields end-to-end transmission performance surpassing classical separation-based JPEG/JPEG2000/BPG source compression combined with ideal channel capacity-achieving code family, especially for sources of small dimensions, e.g., small CIFAR10 image data set [14].

However, one can observe that, in general, as the source dimension increases, e.g., large-scale images, the performance of deep JSCC degrades rapidly, which is even inferior to the classical separation-based coding schemes. In addition, when the CBR \( R \) increases, existing deep JSCC schemes
cannot provide comparable coding gain as that of classical separated coding schemes, i.e., the slope of performance curve slowing down with the increase of the CBR $R$ or the channel signal-to-noise ratio (SNR). This phenomenon stems from the inherent defect of standard deep JSCC that cannot identify the source distribution for realizing a patch-wise variable-length transmission. For example, when the dimension of embedding vector increases, these embeddings corresponding to simple patches will be saturated rapidly, leading to severe channel bandwidth wastes and inferior coding gain. This saturation phenomenon is more likely to appear on large-scale images that need higher dimensional representation.

Furthermore, current deep JSCC methods do not incorporate any hyperprior as side information, a concept widely used in modern image codecs but unexplored in image transmission using deep JSCC with ANNs.

In this paper, we aim to break the above limits by proposing a new joint source-channel coding architecture that integrates the new concept of nonlinear transform coding (NTSCC) and deep JSCC, i.e., nonlinear transform source-channel coding (NTSCC). Particularly, as an emerging class of methods, nonlinear transform coding (NTC) over the past few years has become competitive with the best linear transform codecs for image compression, and outperforms them in terms of rate-distortion (RD) performance under well established perceptual quality metrics, e.g., PSNR, MS-SSIM, LPIPS, etc. In contrast to the linear transform coding (LTC) schemes, NTC can more closely adapt to the source distribution, leading to better compression performance. By integrating NTC into deep JSCC, the proposed NTSCC works on the principle: the source vector $x$ is not directly mapped to the channel input symbols, instead, an alternative (latent) representation of $x$ is found first, a vector $y$ in the latent space, and deep JSCC encoding takes place in this latent representation. By introducing an entropy model on the latent space, NTSCC learns a prior as the side information to approximate the distribution of each patch, which is assumed intractable in practice. Accordingly, the following deep JSCC codec can select an appropriate coding scheme that optimizes the transmission RD performance for each embedding $y_i$. As a result, the proposed NTSCC transmission framework can closely adapt to the source distribution and provide superior coding gain. Notably, the proposed NTSCC method can well support future semantic communications due to its content-aware ability and human perceptual optimization goal.

Specifically, the contributions of this paper can be summarized as follows.

1. **NTSCC Architecture:** We propose a new end-to-end learnable model for high-dimensional source transmission, i.e., NTSCC, that combines the advantages of NTC and deep JSCC. To the best of our knowledge, this is the first work exploiting the nonlinear transform to establish a learnable entropy model for realizing deep JSCC efficiently, where the entropy model on latent code $y$ implicitly represents the source distribution. In our model, the nonlinear analysis transform condenses the source semantic features as a latent representation $y$, thus driving the following source-channel coding.

2. **Adaptive Rate Transmission:** To improve the coding gain of the proposed NTSCC method, we introduce a variable-length transmission mechanism for each embedding vector $y_i$ in the latent code $y$. To this end, a conditional entropy model $P_{y_i|z}$ is performed on each quantized embedding $y_i$ to evaluate the entropy of $y_i$. If the learned entropy model indicates the embedding $y_i$ of high entropy, its corresponding deep JSCC shall be assigned a high coding rate, and vice versa. Accordingly, we develop the Transformer ANN architecture and rate attention mechanism to achieve an adaptive rate allocation for deep JSCC, that can finely tune the coding rate for each embedding $y_i$, thus enable source content-aware transmission.

3. **Hyperprior-aided Codec Refinement:** In the proposed NTSCC method, as the hyperprior on the latent representation, we show that the side information about the entropy model parameters can also be viewed as a prior on the deep JSCC codewords. We exploit this hyperprior to reduce the mismatch between the latent representation marginal distribution for a particular source sample and the marginal for the ensemble of source data the transmission model was designed for. This refinement mechanism only uses a small number of additional bits of information sent from encoder to decoder as signal modifications to achieve much better performance in deep JSCC decoding.

4. **Performance Validation:** We verify the performance of the proposed NTSCC method across test image sources with various resolutions. Results show that the NTSCC method can achieve much better coding gain and RD performance on various established perceptual metrics such as PSNR, MS-SSIM, and LPIPS. Equivalently, achieving the identical end-to-end transmission performance, the proposed NTSCC method can save more than 20% bandwidth cost, compared to both the emerging analog transmission schemes using the standard deep JSCC and the classical separation-based digital transmission schemes.

The remainder of this paper is organized as follows. In the next section [II], we first review the variational perspective on deep JSCC and NTC, and propose the variational model for NTSCC. Then, in section [III], we propose ANN architectures for realizing NTSCC, as well as key methodologies to guide the optimization of the NTSCC model. Section [IV] provides a direct comparison of a number of methods to quantify the performance gain of the proposed method. Finally, section [V] concludes this paper.

**Notational Conventions:** Throughout this paper, lowercase letters (e.g., $x$) denote scalars, bold lowercase letters (e.g., $\mathbf{x}$) denote vectors. In some cases, $x_i$ denotes the elements of $x$, which may also represent a subvector of $x$ as described in the context. Bold uppercase letters (e.g., $\mathbf{X}$) denote matrices, and $\mathbf{I}_m$ denotes an $m$-dimensional identity matrix. $\ln(\cdot)$ denotes the natural logarithm, and $\log(\cdot)$ denotes the logarithm to base 2. $p_x$ denotes a probability density function (pdf) with respect to the continuous-valued random variable $x$, and $P_x$ denotes a probability mass function (pmf) with respect to the discrete-valued random variable $\tilde{x}$. In addition, $\mathbb{E}[\cdot]$ denotes the statistical expectation operation, and $\mathbb{R}$ denotes the real number.
Fig. 1. Left: representation of a deep JSCC encoder combined with the communication channel as a inference model, and corresponding decoder as a generative model. Nodes denote random variables or parameters, and arrows show conditional dependence between them. Right: diagram showing the operational structure of the deep JSCC transmission model. Arrows indicate the data flow, and boxes represent coding functions of data and channel.

set. Finally, \( N(x|\mu, \sigma^2) \triangleq (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \) denotes a Gaussian function, and \( U(a-u, a+u) \) stands for a uniform distribution centered on \( a \) with the range from \( a-u \) to \( a+u \).

II. SYSTEM MODEL AND VARIATIONAL ANALYSIS

Consider the following lossy transmission scenario. Alice is drawing a \( m \)-dimensional vector \( x \) from the source, whose probability is given as \( p_x(x) \). Alice concerns how to map \( x \) to a \( k \)-dimensional vector \( s \), where \( k \) is referred to as the channel bandwidth cost. Then, Alice transmits \( s \) to Bob via a realistic communication channel, who uses the received information \( \hat{s} \) to reconstruct an approximation to \( x \).

A. System Model of Deep JSCC

As stated in the introduction part, in deep JSCC [9], the source vector \( x \in \mathbb{R}^m \), is mapped to a vector of continuous-valued channel input symbols \( s \in \mathbb{R}^k \) via an ANN-based encoding function \( s = f_e(x; \phi_f) \), where the encoder was usually parameterized as a convolutional neural network (CNN) with parameters \( \phi_f \). Then, the analog sequence \( s \) is directly sent over the communication channel. The channel introduces random corruptions to the transmitted symbols, denoted as a function \( W(\cdot; \nu) \), the channel parameters are encapsulated in \( \nu \). Accordingly, the received sequence is \( \hat{s} = W(s; \nu) \), whose transition probability is \( p_{\hat{s}|s}(\hat{s}|s) \). In this paper, we consider the most widely used AWGN channel model such that the transfer function is \( \hat{s} = W(s; \sigma_n) = s + n \) where each component of the noise vector \( n \) is independently sampled from a Gaussian distribution, i.e., \( n \sim N(0, \sigma_n^2I_k) \), where \( \sigma_n^2 \) is the average noise power. Other channel models can also be similarly incorporated by changing the channel transition function. The receiver also comprises a parametric function \( \hat{x} = f_d(\hat{s}; \theta_f) \) to recover the corrupted signal \( \hat{s} \) as \( \hat{x} \), where \( f_d \) can also be a format of CNN [9]. The whole operation is shown in the right panel of Fig. 1. The encoder and decoder functions are jointly learned to minimize the average
\[
\min_{\theta_f, \phi_f} \mathbb{E}_{x \sim p_x} \mathbb{E}_{s \sim p_{s|x}} [d(x, \hat{x})],
\]
where \( d(\cdot, \cdot) \) denotes the distortion loss function.

From the variational perspective, deep JSCC can be viewed as a variational autoencoder (VAE) [17, 18]. As shown in the left panel of Fig. 1 the noisy sequence \( \hat{s} \) can be viewed as a sample of latent variables in the generative model. The deep JSCC decoder acts as the generative model ("generating" the reconstructed source from the latent representation) that transforms a latent variable with some predicted latent distribution into a data distribution that is unknown.

B. Variational Modeling of NTC

As indicated in the introduction section, when the dimension of representation \( k \) becomes large, standard deep JSCC cannot provide sufficient coding gain. As an emerging lossy compression model [19–23], NTC imitates the classical source coding procedure, that operates on a simple principle: a source vector \( x \) is not mapped to codeword vector directly, rather, an alternative latent representation of \( x \) is found first, a vector in some other space \( y \), quantization then takes place in this latent representation, yielding discrete-valued vector \( \tilde{y} \). In standard NTC, due to the quantization step, the resulting \( \tilde{y} \) can be compressed using entropy coding methods, e.g., arithmetic coding [24], to create bits streams. Since entropy coding relies on a prior probability of \( \tilde{y} \), the entropy model \( P_2 \) is established in NTC to provide side information.

The encoder in NTC transforms the source image vector \( x \) using a parametric nonlinear analysis transform \( g_a(x; \phi_g) \) into a latent representation \( y \), which is quantized as \( \tilde{y} \). The latent representation \( y \) preserves the source semantic features while its dimension \( n \) is usually much smaller than the source dimension \( m \). The decoder performs the inverse operation that recovers \( \tilde{y} \) from the compressed signal first, a parametric nonlinear synthesis transform \( g_b(\tilde{y}; \theta_g) \) is then performed on \( \tilde{y} \) to recover the source image \( \hat{x} \). Here, an ideal error-free transmission is assumed such that the decoder can losslessly recover \( \tilde{y} \) from entropy decoding. In NTC, \( y_a \) and \( g_a \) are usually parameterized as ANNs of nonlinear property, rather than the conventional linear transforms in LTC, \( \phi_g \) and \( \theta_g \) encapsulate their neural network parameters.

Since the nonlinear transform is not strictly invertible, also the quantization step introduces error, the optimization of NTC is attributed to a compression RD problem [25]. Assuming an efficient entropy coding is used, rate \( R \) as the expected length of the compressed sequence equals to the entropy of \( \tilde{y} \), which is determined by the entropy model \( P_2 \) as
\[
R = \mathbb{E}_{x \sim p_x} [ -\log Q(g_a(x; \phi_g)) ],
\]
where \( Q \) denotes the quantization function. In the context of this paper, without loss generality, \( Q \) employs the uniform scalar quantization \( [\cdot] \) (rounding to integers). Distortion is the expected divergence between \( \tilde{y} \) and \( x \). Clearly, a higher rate allows for a lower distortion, and vice versa. In order to use the gradient descent methods to optimize the NTC model, Ballé et al. have proposed a relaxed method for addressing the zero gradient problem caused by quantization [19]. It uses a proxy "uniformly-noised" representation \( \tilde{y} \) to replace the quantized representation \( \tilde{y} = [y] \) during the model training.

The optimizing problem of NTC can also be formulated as a VAE model as shown in Fig. 2, a probabilistic generative model stands for the synthesis transform, and an approximate inference model corresponds to the analysis transform. Like that in deep JSCC, the goal of inference model is also creating
a parametric variational density \( q_{\tilde{y}|x} \) to approximate the true posterior \( p_{\tilde{y}|x} \), which is assumed intractable, by minimizing their Kullback-Leibler (KL) divergence over the source distribution \( p_x \), i.e.,

\[
\begin{align*}
\min_{\phi_y, \theta_y} & \quad \mathbb{E}_{x \sim p_x} D_{KL}[q_{\tilde{y}|x} \| p_{\tilde{y}|x}] = \min_{\phi_y, \theta_y} \mathbb{E}_{x \sim p_x} \mathbb{E}_{\tilde{y} \sim q_{\tilde{y}|x}} \\
& \quad \log q_{\tilde{y}|x}(\tilde{y}|x) - \log p_{\tilde{y}|x}(\tilde{y}|x) + \text{const},
\end{align*}
\]

(3)

The minimization of KL divergence is equivalent to optimizing the NTC model for compression performance. As shown in [19], the first term in (3) is computing the transition probability from the source \( x \) to the proxy latent representation \( \tilde{y} \) as

\[
q_{\tilde{y}|x}(\tilde{y}|x) = \prod_i \mathcal{U}(\tilde{y}_i | y_i - \frac{1}{2}, y_i + \frac{1}{2}) \quad \text{with} \quad y = g_a(x; \phi_g),
\]

(4)

where \( \mathcal{U} \) denotes a uniform distribution centered on \( y_i \). Since the uniform distribution width is constant, the first term is also constant which can be technically dropped. The last term can also be similarly dropped. The third term representing the log likelihood can also be modeled by measuring the squared error between \( \hat{x} \) and \( x \), we have \( p_{z|x}(x|\hat{y}) = \mathcal{N}(\hat{x}|x, (2\sigma_g)^{-1}I_4) \) with \( \hat{x} = g_s(\hat{y}; \theta_s) \) where the squared error is weighted by the hyperparameter \( \tau_g \). The second term reflects the cross-entropy between the marginal \( q_{\tilde{y}}(\tilde{y}) = \mathbb{E}_{x \sim p_x}[q_{\tilde{y}|x}(\tilde{y}|x)] \) and the prior \( p_{\tilde{y}}(\tilde{y}) \). It represents the cost encoding \( \tilde{y} \) is constrained by the entropy model \( p_{\tilde{y}} \). In [19], Ballé et al. modeled the prior using a non-parametric fully-factorized density model as

\[
p_{\tilde{y}}(\tilde{y}) = \prod_i \left( p_{\tilde{y}_i|\psi(i)}(\tilde{y}_i|\psi(i)) \ast \mathcal{U}(-\frac{1}{2}, \frac{1}{2}) \right)(\tilde{y}_i),
\]

(5)

where \( \psi(i) \) encapsulates all the parameters of \( p_{\hat{y}_i|\psi(i)} \), the convolutional operation “\( \ast \)” with a standard uniform distribution is used to better match the prior to the marginal. This model is referred to as a factorized-prior model [19].

However, in general cases, there may still exist clear spatial dependencies among the latent representation \( \tilde{y} \), in which case the performance of the factorized-prior model degrades. To tackle this, Ballé et al. introduced an additional set of latent variables \( \tilde{z} \) to represent the dependencies of \( \tilde{y} \) in the same way that the original source \( x \) is transformed to the latent representation \( \tilde{y} \) [20]. Here, each \( \tilde{y}_i \) is variationally modeled as a Gaussian with mean \( \mu_i \) and standard deviation \( \sigma_i \), where the two parameters are predicted by applying a parametric synthesis transform \( h_a \) on \( \tilde{z} \) as

\[
p_{\tilde{y}|\tilde{z}|\tilde{z}}(\tilde{y}, \tilde{z}) = \prod_i \mathcal{N}(\mu_i, \sigma_i^2) \ast \mathcal{U}(-\frac{1}{2}, \frac{1}{2}) \quad (\tilde{y}_i),
\]

with \( (\mu_i, \sigma_i) = h_a(\tilde{z}; \theta_h) \).

The corresponding analysis transform \( h_a \) is stacked on top of \( \tilde{y} \) creating a joint factorized variational posterior as

\[
q_{\tilde{y}, \tilde{z}|x}(\tilde{y}, \tilde{z}|x) = \prod_i \mathcal{U}(\tilde{y}_i | y_i - \frac{1}{2}, y_i + \frac{1}{2}) \cdot \prod_j \mathcal{U}(\tilde{z}_j | z_j - \frac{1}{2}, z_j + \frac{1}{2}) \quad \text{with} \quad y = g_a(x; \phi_g), z = h_a(y; \phi_h),
\]

(7)

Since we do not have prior beliefs about the hyperprior \( \tilde{z} \), it can be modeled as non-parametric fully factorized density like [5], i.e.,

\[
p_{z|x}(z|x) = \prod_j \left( p_{z_j|\psi(j)}(z_j|\psi(j)) \ast \mathcal{U}(-\frac{1}{2}, \frac{1}{2}) \right)(\tilde{z}_j),
\]

(8)

where \( \psi(j) \) encapsulates all the parameters of \( p_{z_j|\psi(j)} \). The optimization goal (3) works out to be changed as:

\[
\min_{\phi_y, \theta_y, \phi_x, \theta_x, \phi_h, \theta_h} \mathbb{E}_{x \sim p_x} D_{KL}[q_{\tilde{y}, \tilde{z}|x} \| p_{\tilde{y}, \tilde{z}|x}] = \min_{\phi_y, \theta_y, \phi_x, \theta_x, \phi_h, \theta_h} \mathbb{E}_{x \sim p_x} \mathbb{E}_{\tilde{y}, \tilde{z} \sim q_{\tilde{y}, \tilde{z}|x}} \left[ \log q_{\tilde{y}, \tilde{z}|x}(\tilde{y}, \tilde{z}|x) - \log p_{\tilde{y}, \tilde{z}}(\tilde{y}, \tilde{z}) - \log p_x(x) \right]
\]

\[
- \text{const} + \log p_x(x|\tilde{y}, \tilde{z}) + \text{const},
\]

(9)

where the third term can be viewed as the side information widely used in traditional transform coding schemes. The right panel of Fig. 2 depicts the procedure of how the model is used for data compression. Following the variational analysis, the loss function for training the NTC model is

\[
L = \mathbb{E}_{x \sim p_x(x)} \left[ \lambda \left( -\log p_{\tilde{y}|\tilde{z}}(\tilde{y}|\tilde{z}) - \log p_x(x) \right) + d(x, \tilde{x}) \right]
\]

(10)

where \( \tilde{y} = g_a(x; \phi_g) + o, \tilde{x} = h_a(y; \phi_h) + o, \tilde{x} = g_s(\tilde{y}; \theta_s) \), with \( o \) denotes uniformly sampling one random quantization offset per latent dimension.

### C. Variational Modeling of the Proposed NTSCC

We integrate both the advantages of NTC and classical deep JSCC, that is collected under the name nonlinear transform source-channel coding (NTSCC). In the transmitter, the analysis transform in NTC is used as a type of “pre-coding” before the encoding of deep JSCC, which extracts the source semantic features as a latent representation. Deep JSCC then operates on this latent space. The bottom panel of Fig. 3 illustrates how the NTSCC model is used for data transmission. The analysis transform module subjects the input source vector
Let’s take a close look at each term of the last line in (11). First, the variational inference model is computing the analysis transform and the deep JSCC encoding of the source vector x, and adding the channel noise n, thus:

\[ q_{\tilde{z}|x}(\tilde{z}|x) = \prod_i \mathcal{N}(\tilde{s}_i|s_i, \sigma_i^2) \cdot \prod_j \mathcal{U}(\tilde{z}_j|z_j - \frac{1}{2}, z_j + \frac{1}{2}) \]

with \( y = g_a(x; \phi_g), s = f_e(y; \phi_f), z = h_a(y; \phi_h) \).

As discussed before, the first term in the KL divergence can be technically dropped from the loss function.

The second term is identical to the cross-entropy between the marginal \( q_{\tilde{z}}(\tilde{z}) = \mathbb{E}_{x \sim p_x} \mathbb{E}_{\tilde{z} \sim q_{\tilde{z}|x}}[q_{\tilde{z}|x}(\tilde{z}|x)] \) and the prior \( p_{\tilde{z}}(\tilde{z}) \). It stands for the cost of encoding the side information in the inference model assuming \( p_{\tilde{z}} \) as the entropy model. In practice, since we do not have prior beliefs about the hyperprior \( \tilde{z} \), it can also be modeled as non-parametric fully factorized density as \( \mathcal{N}(\mu, \sigma^2) \).

The third term corresponds to the cross-entropy encoding \( \hat{s} \) that denotes the transmission rate of source message. As shown in Fig.3, we derive \( p_{\hat{y}|\tilde{z}} \) with the help of the intermediate proxy variable \( y \), i.e., the latent presentation of \( x \). Each element \( y \) is variationally modeled as a Gaussian with mean \( \mu_i \), and standard deviation \( \sigma_i \), where the two parameters are predicted by applying a parametric synthesis transform \( h_a \) on \( \tilde{z} \) as

\[ p_{y|\tilde{z}}(y|\tilde{z}) = \prod_i \mathcal{N}(y_i|\mu_i, \sigma_i^2) \] with \( (\mu, \sigma) = h_a(\tilde{z}; \theta_h) \).

The density \( p_{y|x} \) can be transformed to a new density \( p_{y|x} \) by using the deep JSCC encoder function as \( p_{y|x} = f_e(p_{y|x}; \phi_f) \).
that employs the formula of functional distribution of random variable \( [Z] \). Under the AWGN channel, the received signal \( \hat{s} = s + n \) with \( n \sim \mathcal{N}(0, \sigma_n^2 I_k) \). We can thus calculate the density \( p_{y|z} \) as

\[
p_{y|z}(\hat{s}|\hat{z}) = \prod_i (p_{y_i|z}(s_i|\hat{z}) \ast \mathcal{N}(0, \sigma_n^2)) (s_i),
\]

where “\( \ast \)” denotes the convolutional operation. Since the latent representation \( y \) is directly fed into the deep JSCC encoder without quantization, \( \mathcal{L} \) indeed represents a differential entropy, as opposed to the discrete entropy used for the rate constraint in NTC. Note that \( p_{y|z}(\hat{s}|\hat{z}) \) is originally determined by \( p_{y|z}(y|\hat{z}) \), thus, to ensure a stable model training, like that in NTC, we employ a proxy “uniformly-noised” representation \( \hat{y} \) that is variationally modeled using \( \hat{z} \) as \( [Z] \) to alternate as the transmission rate constraint term in practical implementations of NTSCC, which is marked as dashed lines in the inference model of Fig. [5]. During the model testing, the conditional entropy model \( p_{y|z}(y|\hat{z}) \) established by taking discrete values from the learned entropy model \( p_{y|z}(\hat{s}|\hat{z}) \), i.e., \( p_{y|z}(y|\hat{z}) = p_{y|z}(\hat{y}|\hat{z}) \) with \( \hat{y} = [y] \) and \( \hat{z} = [z] \). Therefore, the transmission rate is constrained proportionally to \(- log p_{y|z}(\hat{y}|\hat{z})\).

The fourth term represents the log likelihood to recovering \( x \), the output of the synthesis transform \( g_s \), which is weighted by the output of the deep JSCC decoder \( f_d \). Here, the change of the fourth term from (b) to (c) in \( [1] \) follows the law of Jensen inequality, which indicates an upper bound on the KL divergence. The densities \( p_{y|x} \) and \( p_{y|x, z} \) can be assumed as

\[
p_{y|x} (x|\hat{y}) = \mathcal{N}(x|\hat{x}, (2\tau_g)^{-1} I_m),
\]

and

\[
p_{y|x, z} (y|\hat{s}, \hat{z}) \geq p_{y|x} (y|\hat{s}) \cdot p_{y|x} (y|\hat{z})
= \prod_i \mathcal{N}(y_i|\hat{y}_i, (2\tau_g)^{-1}) \cdot \prod_i \mathcal{N}(y_i|\hat{y}_i, \bar{\sigma}_i^2),
\]

where \( \hat{y} = f_d(\hat{s}; \theta_f), (\hat{\mu}, \bar{\sigma}) = h_s(\hat{s}; \theta_h) \),

which measure the squared difference. Different from conventional schemes, \( [1] \) indicates that the deep JSCC decoder can use both the received signal \( \hat{s} \) and the side information \( \hat{z} \) to estimate the latent representation \( y \).

From the above analysis, we also summarize the operations in the receiver, it first recovers the hyperprior \( \hat{z} \) from the transmitted side information. It then exploits \( h_s \) to recover \( \hat{\mu} \) and \( \bar{\sigma} \), which provide a prior probability helping to recover \( \hat{y} = \max_y [p_{y|x} (y|\hat{s}) \cdot p_{y|x} (y|\hat{z})] \) that can be computed as \( [1] \). In practice, the decoder for recovering \( \hat{y} \) can be summarized as a new parametric function \( \hat{y} = f_d(\hat{s}, \hat{\mu}, \bar{\sigma}; \theta_f) \) with both \( \hat{s} \) and \( \hat{\mu}, \bar{\sigma} \) as inputs, where \( \theta_f \) encapsulates the ANN parameters constituting \( f_d \). During model training, \( \hat{z} \) is replaced by the “uniformly-noised” proxy \( \hat{Z} \) to generate \( \hat{\mu}, \bar{\sigma} \). It then feeds \( \hat{y} \) into \( g_s \) to reconstruct the source \( x \). The whole transceiver operation diagram is illustrated in Fig. [3].

III. ARCHITECTURE AND IMPLEMENTATIONS

In this section, we discuss details about the NTSCC architecture and key methodologies. Following the aforementioned VAE model, the optimizing of NTSCC can also be attributed as a transmission RD optimization problem, i.e.,

\[
L = \mathbb{E}_{x \sim p_{x}(x)} \left[ \lambda \left( \frac{-\eta \log p_{\hat{y}|z}(\hat{y}|\hat{z}) - \log p_{\hat{y}|z}(\hat{y}|\hat{z})}{C_z} \right) + d(x, \hat{x}_{NTSCC}) \right]
\]

with \( y = g_a(x; \phi_g), \hat{y} = [y], z = h_a(y; \phi_h), \hat{z} = [z] \),

\[
\hat{s} = W(f_c(y; \phi_f; \nu), (\hat{\mu}, \bar{\sigma}) = h_s(\hat{s}; \theta_h), \hat{y} = f_d^*(\hat{s}, \hat{\mu}, \bar{\sigma}; \theta_f), \hat{x} = g_s(\hat{y}; \theta_g),
\]

which where the parameter \( \nu \) controls the scaling relation between the entropy of latent representation \( \hat{y} \) and its analog transmission channel bandwidth cost \( k_y \), \( C_z \) denotes the digital channel capacity to transmit the quantized hyperprior \( \hat{z} \), thus the digital channel bandwidth cost \( k_z \) can be computed to transmit the side information. The Lagrange multiplier \( \lambda \) on the total transmission channel bandwidth term determines the trade-off between the transmission rate \( k = k_y + k_z \) and the end-to-end distortion \( d \). Moreover, from the above variational modeling analysis, we find that both the conditional entropy model \( p_{y|z}(\hat{y}|\hat{z}) \) and the hyperprior model \( P_{\hat{y}|z}(\hat{y}|\hat{z}) \) can be factorized. In order to use the gradient descent methods to optimize the NTSCC model, we relax the quantized variables as “uniformly-noised” proxies like that in NTC, therefore, the loss function for model training can be written as

\[
L = \mathbb{E}_{x \sim p_{x}(x)} \left[ \lambda \sum_i \left( \frac{-\eta \log p_{\hat{y}_i|z}(\hat{y}_i|\hat{z})}{k_{y_i}} \right) + \sum_j \left( \frac{1}{C_z} \log p_{\hat{z}_j|\psi(j)}(\hat{z}_j|\psi(j)) \right) + d(x, \hat{x}_{NTSCC}) \right]
\]

with

\[
\hat{y}_i = y_i + \alpha_i, \hat{z}_j = z_j + \alpha_j, y = g_a(x; \phi_g), z = h_a(y; \phi_h),
\]

and

\[
p_{\hat{y}_i|z}(\hat{y}_i|\hat{z}) = \mathcal{N}(\hat{y}_i|\hat{\mu}_i, \bar{\sigma}_i^2) \ast \mathcal{U}(-\frac{1}{2}, \frac{1}{2}) (\hat{y}_i)
\]

and

\[
p_{\hat{z}_j|\psi(j)}(\hat{z}_j|\psi(j)) = \left( p_{\hat{z}_j|\psi(j)}(\hat{z}_j|\psi(j)) \ast \mathcal{U}(-\frac{1}{2}, \frac{1}{2}) \right) (\hat{z}_j)
\]

Correspondingly, during model testing, the conditional entropy model \( P_{\hat{y}_i|z} \) is established by taking discrete values from the learned model \( p_{\hat{y}_i|z} \) as

\[
P_{\hat{y}_i|z}(\hat{y}_i|\hat{z}) = p_{\hat{y}_i|z}(\hat{y}_i|\hat{z}),
\]

and \( P_{\hat{z}_j|\psi(j)} \) is similarly computed as

\[
P_{\hat{z}_j|\psi(j)}(\hat{z}_j|\psi(j)) = p_{\hat{z}_j|\psi(j)}(\hat{z}_j|\psi(j)).
\]
communication decoder refinement hyperprior-aided adaptive rate transmission

Fig. 4. Illustration of an NTSCC codec architecture using learned FA.

A. The Overall Architecture of NTSCC

The learned NTSCC model in [18] indicates that the pmf \( P_{\hat{y}|z}(\hat{y}|\tilde{z}) \) for predicting the entropy of latent representation is factorized over each dimension without relying on preceding dimensions. Conditioning on the hyperprior vector \( \tilde{z} \) typically requires transmitting the vector as side information. The whole structure corresponds to a learned forward adaptation (FA) of the density model [15]. To seek computational parallelism, in this paper, we focus on the FA mode, a better performance backward adaptation (BA) mode with higher processing latency will be discussed in the future [27].

Fig. 4 illustrates the overall architecture of NTSCC using learned FA. \( x \) is the source vector at the transmitter, and \( \hat{x} \) denotes the recovered vector at the receiver. \( y \) is the semantic latent representation tensor that is obtained by performing an ANN-based transform function \( g_a \) on \( x \), and \( \tilde{y} \) is the uniformly quantized counterparts of \( y \). Also, \( z \) is the latent representation of \( y \) computed using an ANN-based transform \( h_a \), denoting the side information, whose uniformly quantized version is \( \tilde{z} \). While the entropy model of \( \tilde{y} \) on \( \tilde{z} \) is predetermined, the factorized entropy model of \( \tilde{y} \) is assumed to be conditionally independent Gaussian with the mean tensor \( \tilde{\mu} \) and the standard deviation tensor \( \tilde{\sigma} \) as [19]. The two tensors are obtained by performing the ANN-based function \( h_s \) on \( \tilde{z} \). The resulting entropy terms \( -\log P_{\tilde{y}|\tilde{z}}(\tilde{y}|\tilde{z}) \) are employed for determining the channel bandwidth cost of each dimension \( y_i \), so as to achieve adaptive rate transmission. Bob begins with channel decoding (CD) and entropy decoding (ED) to recover the side information \( \tilde{z} \), and then uses it to decode \( \tilde{y} \). Alice should know the entropy model on \( \tilde{z} \) to entropy encode (EE) and channel encode (CE) it, that is modeled as a non-parametric density conditioned on \( \psi \) as [20].

A Special Note: As shown in Fig. 4 the side information \( \tilde{z} \) is not necessary for Bob. As analyzed in previous section, when the hyperprior \( \tilde{z} \) is transmitted through the side information link, to decode \( \tilde{y} \), Bob can jointly use channel received \( \hat{s} \) and side information \( \tilde{\mu}, \tilde{\sigma} \) as the inputs of decoder function \( f_{\hat{d}} \) for getting better performance. In this case, to ensure reliable transmission of the side information, channel coding (CE and CD) should adopt advanced capacity-approaching codes, such as low-density parity-check (LDPC) codes [28] or polar codes [29]. On the other hand, when \( \tilde{z} \) is only used by Alice to allocate channel bandwidth cost of each \( y_i \), some performance degradation will be observed in the deep JSCC decoder \( f_{\hat{d}} \) while the cost for transmitting \( \tilde{z} \) can be omitted. In subsequent sections, unless specified clearly, NTSCC refers to the model where the hyperprior \( \tilde{z} \) is transmitted as the side information to refine the decoder. Detailed performance comparison about whether to transmit the side information \( \tilde{z} \) will be given in the subsequent ablation study.

Next, we discuss the procedure of NTSCC model training as shown in Algorithm 1. Here, some tricks should be noted to ensure fast and stable training. First, before the training of NTSCC model, the parameters of \( g_a, g_s, h_a, h_s \) should be initialized by pre-training the corresponding NTC model as [19], where no transmitting error is considered such that NTC only executes the data compression task. During the training of NTSCC model, we modify the loss function derived in [18] by adding the NTC distortion terms, in this way, the convergence of NTSCC model training can be more stable and we can maintain the performance of NTC component for compression.

**Algorithm 1: Training the NTSCC model**

**Input:** Training data \( x \), the Lagrange multiplier \( \lambda \) on the rate term, the scaling factor \( \eta \) from entropy to channel bandwidth cost, learning rate \( \gamma \);

// model pre-training
1. Initialize parameters \( (\phi_g^{(0)}, \phi_h^{(0)}, \theta_y^{(0)}, \theta_h^{(0)}) \) by pre-training the corresponding NTC model as that in [15];
2. Randomly initialize parameters \( (\phi_f^{(0)}, \theta_f^{(0)}) \);
3. Encapsulate \( w^{(0)} = (\phi_g^{(0)}, \phi_h^{(0)}, \phi_f^{(0)}, \theta_y^{(0)}, \theta_h^{(0)}, \theta_f^{(0)}) \);
4. for \( t = 1, 2, \cdots, T \) do
5. Sample \( x \sim P_x \);
6. Calculate the RD loss function \( L(w^{(t-1)}, \lambda, \eta) = d(x, \hat{x}_{NTSC}) + d(x, \hat{x}_{NTC}) + \lambda(k_a + k_z) \);
7. Calculate the gradients \( \nabla_{w^{(t-1)}} L(w^{(t-1)}, \lambda, \eta) \);
8. Update \( w^{(t)} = w^{(t-1)} - \gamma \nabla_{w^{(t-1)}} L(w^{(t-1)}, \lambda, \eta) \);
9. return Trained parameters \( w^{(T)} = (\phi_g^{(T)}, \phi_h^{(T)}, \phi_f^{(T)}, \theta_y^{(T)}, \theta_h^{(T)}, \theta_f^{(T)}) \).
task. Moreover, due to the introduction of rate adaptation in NTSCC, the channel bandwidth cost of each embedding $y_i$ is different, thus, multi-head ANN structures shall be used to implement deep JSCC codec $f_e$ and $f_e^*$. In each round of model training, only parts of $\phi_f$ and $\theta_f$, will be updated depending on the selected transmitting rates. Details will be given later.

B. Modular Implementation Details

In this part, we present the implementation details of each module in NTSCC. As aforementioned, the proposed NTSCC model mainly consists of a nonlinear transform step and a rate-adaptive deep JSCC step. The key of NTSCC implementation is designing dynamic and efficient ANN structures that can learn patch-wise representations and use the side information provided by hyperprior to flexibly determine the transmission bandwidth cost of each patch. To this end, the encoder function $f_e$ should incorporate these external parameters, such as the transmission rate of each embedding $y_i$ that is determined by the learned entropy model $-\log P_{\hat{y}_i|y_i}$ and the predetermined scaling factor $\eta$. We inform JSCC codec ANNs of this global or contextual information using “indicators”. These indicators are generally referred to as a group of learnable parameters, each one is associated with a specific external state. In this work, we employ vision Transformers as the ANN backbone [30], that integrates tokens or positional embedding as indicators.

Next, we describe the implementation details of each module in NTSCC, where the source is assumed as image.

1) Nonlinear Transform Modules $g_a$, $g_s$, $h_a$, and $h_s$. The proposed network architecture of the analysis transform $g_a$ is illustrated in Fig. 5. An RGB image source $x \in \mathbb{R}^{h \times w \times 3}$ is first split into $l_1 = \frac{h}{2} \times \frac{w}{2}$ non-overlapping patches and each patch is of $2 \times 2 \times 3$ dimensions. Each patch embedding can be viewed as a “token”, we thus have a sequence of tokens $(x_1, \ldots, x_{l_1})$ by putting these tokens in the order from top left to bottom right. After the patch partition, patch tokens are fed into a linear fully-connected (FC) layer for obtaining the embedding with $c$ feature dimensions.

Then, $N_1$ Transformer blocks are applied on these $l_1$ patch embedding tokens [30]. As shown in Fig. 5, a Transformer block is a sequence-to-sequence function that consists of a multi-head self-attention layer and a feed-forward layer with both having skip connection and layer normalization (LayerNorm) operations [31], [32]. The operations in each layer of the Transformer block are described as

$$O_1 = X + \text{MHSA}(X)$$

and

$$O_2 = O_1 + \text{MLP}(	ext{LayerNorm}(O_1)),$$

where $O_1$ and $O_2$ denote the output from the self-attention layer and the feed-forward layer, respectively. The function of MHSA denotes the multi-head self-attention with learnable relative positional bias [33]. The function of MLP consists of a hidden layer and an embedding layer with the GELU activation [34].

After that, the analysis transform $g_a$ merges neighboring embeddings and reduces the concatenated $4c$-dimensional embeddings to $c$-dimensional by a linear FC layer, which achieves down-sampling of resolution by two times. Then, it feeds the $l_2 = \frac{h}{4} \times \frac{w}{4}$ higher-resolution patch embedding tokens to another $N_2$ Transformer blocks. As shown in Fig. 5 we encapsulate a down-sampling layer and the following Transformer blocks as one stage, the aforementioned procedure has
included two stages. For small images, e.g., the CIFAR10 dataset of $32 \times 32$ pixels per image [14], $g_y$ with two stages is enough, thus the analysis transform module outputs the latent representation $y \in \mathbb{R}^{\frac{1}{4} \times \frac{1}{4} \times c}$. For large images, e.g., the CLIC2021 image dataset with size up to 2048 $\times$ 1890 pixels [35], one needs more stages to obtain higher-resolution patch embeddings, in this paper, we adopt $g_y$ with four stages as the analysis transform to produce the latent representation $y \in \mathbb{R}^{\frac{1}{16} \times \frac{1}{16} \times c}$ as depicted in Fig. 5 for medium to large images.

The synthesis transform module $f_y$ has a symmetric architecture with analysis transform $g_y$, including the patch division operation for up-sampling and Transformers. It recovers input images from noisy or quantized latent representations. The hyperprior autoencoder, including $h_a$ and $h_s$, summarizes the distribution of means and standard deviations in $z$. It consists of fully convolutional layers followed by the ReLU activation functions as shown in Fig. 6.

2) Codec Modules $f_c$ and $f_d^2$: The overall architectures of deep JSCC codec modules are illustrated in Fig. 7. $f_c$ learns to transmit the latent representation $y$ with variable transmission rate in accordance to the entropy model $P_y|x$. In particular, the encoder $f_c$ first partitions the latent representation $y$ into patch embedding sequence $(y_1, y_2, \ldots, y_l)$. As we shall note, each $y_i$ is a $c$-dimensional feature vector. The learned entropy model $-\log P_{y_i|x}(y_i|z)$ indicates the summation of entropy along all $c$ dimensions of $y_i$, thus, the information density distribution of $y$ has been well captured. Accordingly, the channel bandwidth cost $\bar{k}_{y_i}$ for transmitting $y_i$ can be determined as

$$\bar{k}_{y_i} = Q^e(k_{y_i}) = Q^e(-\eta \log P_{y_i|x}(y_i|z)) = Q^e(-\eta \log p_{y_i|x}(\hat{y}_i|z)), \quad (25)$$

where the learned entropy model $p_{y_i|x}$ follows [19], $Q^e$ denotes a scalar quantization whose range includes $2^{k_{q}}$ ($k_{q} = 1, 2, \ldots$) integers, and the quantization value set $\mathcal{V} = \{v_1, v_2, \ldots, v_{2^{k_{q}}-1}\}$ is related to the scaling factor $\eta$ and the Lagrange multiplier $\lambda$. Hence, a predetermined $k_{q}$ bits must be transmitted from Alice to Bob as an extra side information to inform the receiver which rate is allocated to each $y_i$.

As a toy visualization demo, Fig. 8 shows the rate adaption results of the left top image, where each patch is assigned a channel bandwidth value within $\mathcal{V}$. It can be observed that complex regions (water, text, and human) are encoded with more channel bandwidth costs, while simple areas (sky and canvas) use fewer. Furthermore, with the decrease of $\lambda$, the NTSCC model tends to learn the latent representation with higher entropy to reduce the distortion. As a result, the learned entropy model guides the $f_c$ to use more channel bandwidth for transmission.

To adaptively map $y_i$ to a $\bar{k}_{y_i}$-dimensional channel-input vector $s_i$, an intuitive solution is dividing the encoder $f_c$ as $l$ parts, each one is responsible to encode a patch embedding $y_i$. To attain this, one needs to train $2^{k_{q}}$ pairs of sub-encoders and sub-decoders. Apparently, this tough method will lead to large training complexity and storage costs. Furthermore, it ignores the contextual dependencies among $y_i$, which will degrade the system performance. Therefore, in this paper, we introduce the dynamic neural network structure [36] into Transformers to realize the deep JSCC encoder $f_c$. Compared to static models of fixed computational graphs and parameters, dynamic networks can adapt their network structures or parameters to different inputs, leading to notable advantages in terms of performance, computational efficiency, etc [36]. Here, we employ $N_f$ shared Transformer blocks for feature extraction and light FC layers to encode $y_i$ into given dimension $\bar{k}_{y_i}$. The process includes two steps:

(a) Indicating the shared Transformer blocks of the channel bandwidth cost $\bar{k}_{y_i}$ of each $y_i$ through self-attention;
(b) Using light FC layers to transform the Transformer output corresponding to $y_i$ into a $\bar{k}_{y_i}$-dimensional vector.

Inspired by positional embedding of Transformer, which enables the network to distinguish the relative positions of patches through position-related bias [30], we develop a rate token vector set $\mathcal{R} = \{r_{v_1}, r_{v_2}, \ldots, r_{v_{2^{k_{q}}}}\}$ to indicate the rate information. As shown in Fig. 7, each $y_i$ will be added its corresponding rate token vector $r_{\bar{k}_{y_i}}$ before fed into Transformer blocks. As a result, the output of shared Transformers can learn to adapt to the entropy of $y_i$, and the following FC layer can efficiently adjust output dimension.

In the receiver, noisy channel vector $\hat{s}_i$ with different length will be reshaped to the unified dimension by FC blocks, then added the rate tokens and sent into $N_f$ shared Transformer blocks in parallel. The tentatively recovered version of $y_i$ is $\hat{y}_i$. As stated in [10], the deep JSCC decoder can further refine the reconstructions $\hat{y}$ with the help of the learned prior $\hat{\mu}$ and $\hat{\sigma}$ that is realized by a two-layer MLP network in $f_d^2$.

IV. EXPERIMENTAL RESULTS

A. Experimental Setup

1) Datasets: We quantify the wireless image transmission performance on RGB small-size image dataset CIFAR10 [14] (50,000 training images and another 10,000 test images with $32 \times 32$ pixels), medium-size dataset Kodak [37] (24 images, 768 $\times$ 512 pixels), and large-size dataset CLIC2021 [35] (60 images, up to $2048 \times 1890$ pixels). The content and resolution of these datasets are diversified, thus have been widely used to measure the performance of image-related algorithms. In addition, the dataset for training the proposed NTSCC model for large images consists of 500,000 images sampled from the Open Images Dataset [38]. During model training, images are randomly cropped into $256 \times 256$ patches.
2) **Comparison Schemes:** Our comparison schemes include both the emerging deep JSCC scheme proposed in [9] and the classical separation-based source and channel coding schemes. Specifically, we employ the powerful image codec BPG [39] combined with a practical LDPC code [28] or an ideal capacity-achieving channel code for the latter scheme (marked as “BPG + LDPC” and “BPG + Capacity”, respectively). As we shall note, the ideal “BPG + Capacity” scheme can be viewed as a performance bound on the separation-based source and channel coding schemes. In practical implementation, to be aligned with previous works [9], we convert two consecutive real symbols in \( s \) as one complex channel-input symbol and add complex Gaussian noise.

3) **Evaluation Metrics:** We qualify the end-to-end image transmission performance of the proposed NTSCC models and state-of-the-art models using both the widely used pixel-wise metric (i.e., the peak signal-to-noise ratio (PSNR)) and the perceptual metric (i.e., the multi-scale structural similarity index (MS-SSIM) [40]) and the emerging deep learning based perceptual metric (i.e., the learned perceptual image patch similarity (LPIPS) [16]). PSNR corresponds to the pixel-wise \( \ell_2 \) Euclidian distance, thus when evaluating PSNR performance, we set the distortion function \( d \) as the mean square error (MSE) between \( x \) and \( \hat{x} \). When evaluating the MS-SSIM performance, the training loss \( d \) is set as “1 - MS-SSIM” so as to minimize \( d \). It is usually known that a higher PSNR/MS-SSIM indicates a better performance.

Although PSNR and MS-SSIM are the most widely used metrics, they are simple and shallow functions, and fail to account for many nuances of human perception [16]. To be more closely aligned with the purpose of semantic communications, we further adopt the emerging deep learning based LPIPS metric [16] as a perceptual loss to quantify the image transmission performance, it can imitate the human perceptual assessment process to give the LPIPS loss score. A lower LPIPS value indicates a lower distortion, with a maximum value of 1. In practical model training, one can observe that the weakness of each metric is exploited by learning algorithms, e.g., checkerboard artifacts may appear when targeting neural network derived by the LPIPS metric, relying on MS-SSIM can cause poor text reconstructions, and MSE yields blurry reconstructions [21], [41]. To tackle this, when the NTSCC model targets at human subjective perception metric, we introduce a conditional generative adversarial network (cGAN) [42] and augment the MSE distortion item with GAN loss and LPIPS loss. The total distortion item for model training is

\[
d(x, \hat{x}_{\text{NTSCC}}) = \beta_M \| x - \hat{x}_{\text{NTSCC}} \|^2 + \beta_L d_{\text{LPIPS}}(x, \hat{x}_{\text{NTSCC}}) - \beta_D \log(D(\hat{x}_{\text{NTSCC}}, \hat{y})),
\]

where \( \beta_M, \beta_L, \beta_D \) control the trade-off between three distor-
tion terms, $d_{LPIPS}$ is the LPIPS distortion, and $D$ denotes the discriminator of the cGAN, which maps the latent representation and reconstruction to the probability that it is a sample from the true distribution $p_{x|y}$ rather than from the synthesis transform. Here, the discriminator adopts a fully convolutional network with the same architecture as [21].

4) Model Training Details: As aforementioned, we set up the analysis transform $g_a$ and the synthesis transform $g_s$ with 4 stages with $N_1 = 1, N_2 = 1, N_3 = 2, N_4 = 6$ Transformer blocks for Kodak and CLIC2021 datasets while using 2 stages with $N_1 = 2, N_2 = 6$ Transformer blocks for the tiny CIFAR10 dataset. In all experiments, we use $N_c = N_d = 4$ transformer blocks in JSCC codec $f_e$ and $f_d^{JSCC}$, the number of heads in MHSA is 8, and the channel dimension $c$ is set to 256. The quantization value set $V$ consists of 16 values, which are evenly distributed from 16 to 256. Therefore, an extra side information of $k_y = 4$ bits will be transmitted to inform the receiver which channel bandwidth cost is allocated for each patch. Besides, the complexity of the self-attention calculation is quadratic to the length of the token sequence, resulting in unaffordable computation/memory cost for the transmission of high-resolution images (e.g., $2048 \times 1890$ pixels of CLIC2021 dataset). To tackle this, we employ the shifted-windows-based self-attention mechanism (Swin Transformer) in [31] for large-scale images, which significantly reduces the computational and memory overhead by computing MHSA within non-overlapped local windows size of $16 \times 16$.

During the training process, we use the Adam optimizer [42] with a learning rate of $10^{-4}$. The mini-batch size is set as 10. All implementations were done on Pytorch [44], and it takes about 4 days from pretrained NTC models to train the whole NTSCC model using single RTX 3090 GPU.

B. Results Analysis

1) PSNR Performance: Fig. 9 demonstrates the RD results among various transmission methods at the AWGN channel with channel SNR = 10dB, where the distortion is measured in terms of PSNR. To achieve different RD tradeoffs, we use NTSCC models trained with $\lambda \in \{1024, 256, 64, 16, 4\}$ and set $\eta = 0.2$. For separation schemes, we employ a 2/3 rate (4096, 6144) LDPC code with 16-ary quadrature amplitude modulation (16QAM) for the “BPG + LDPC” scheme. The ideal “BPG + Capacity” scheme adopts the Gaussian channel capacity formula [10] as the transmission rate per channel bandwidth. From these results, we can find that the proposed NTSCC method can generally outperform deep JSCC for all CBRs, and their performance gap increases with the resolution of the image dataset and the channel bandwidth ratio. Compared with separation-based methods, the proposed NTSCC shows competitive performance to the ideal “BPG + Capacity” scheme and surpasses the practical “BPG + LDPC” for all three datasets. Furthermore, NTSCC can outperform “NTC + LDPC” which means the gains are not purely from better compression brought by NTC, and there is an element of good match between the learned deep JSCC and the nonlinear transform. In addition, the proposed NTSCC can closely approach the performance of “NTC + Capacity” which is an upperbound of NTC combined with channel coding. In fact, “NTC + Capacity” is only an approximated upperbound of NTSCC. The reason is that NTC suffers from some extra performance loss caused by the quantization of $y$, while our NTSCC can directly exploit $y$ for transmission, that may even bring some performance gain.

As we shall note, with the help of the adaptive rate transmission mechanism, the proposed NTSCC method provides comparable coding gain as that of the BPG series in most cases, while conventional deep JSCC degrades rapidly with the increase of bandwidth ratio. We also note that the NTSCC may not maintain this coding gain all the time, especially at very high bandwidth ratio regions. The phenomenon stems from the ANN structure of $g_a$, where we set the number of channels at the output layer as $c = 256$, thus the following deep JSCC encoder $f_e$ can only support $k_{y_i} = 256$ dimensions for each patch embedding $y_i$ at most. Otherwise, it may need an extra FC layer to expand the output dimension, this operation will degrade the coding gain. As a result, many patches at the high bandwidth ratio region have been saturated to 256 transmission dimensions. Besides, we also note that the NTSCC on the tiny CIFAR10 dataset performs even slightly worse than the deep JSCC scheme. The reason is that the coding gain brought by adaptive rate allocation and codec refinement may be taken away by the cost of transmitting side information, especially
for low bandwidth ratio regions. However, the relative overhead for transmitting side information will become negligible for large-scale images.

Fig. 10(a) and Fig. 10(b) show the PSNR results of high-resolution images versus the change of channel SNR, where the average bandwidth ratio is $R = k/m = 1/16$. For the “BPG + LDPC” scheme, similar to [10], we evaluate the performance of different combinations of LDPC coding rate and modulations and present the envelope of the best performing configurations at each SNR. Since the NTSCC method learns an adaptive coding rate allocation mechanism, we traverse the scaling factor $\eta \in [0.05, 0.2]$ and fine-tune the NTSCC model with $\lambda = 16$ at each SNR to ensure the maximum CBR lower than 1/16, thus achieving a fair comparison. Here, for all models, the training SNR equals the testing SNR. We find that the proposed NTSCC method brings considerable performance gain, outperforming the standard deep JSCC by a margin of at least 1dB. Furthermore, the coding gain of NTSCC holds with the increase of SNR. Like that in Fig. 2, NTSCC can also outperform “NTC + LDPC” and closely approach the performance bound of “NTC + Capacity”.

Besides, as observed in Fig. 10(c), the proposed NTSCC also achieves graceful degradation as deep JSCC does when the testing SNR decreases from the training SNR, while the performance of separation-based “BPG + LDPC” transmission scheme reduces drastically (known as the cliff effect).

Table I provides BD-CBR and BD-PSNR results where the relative metric “BD-X” has been widely used to evaluate the performance of different image/video compression systems [45]. Here, the BD-CBR represents the average percentage of CBR savings compared with the baseline scheme at the same PSNR. A negative number of BD-CBR stands for bandwidth saving, and a positive number means bandwidth cost increasing. The BD-PSNR represents the PSNR gains at the same average CBR. A positive number of BD-PSNR means image transmission quality improvement and vice versa. We view the “BPG + LDPC” as the baseline scheme. Results indicate that the proposed NTSCC method with two-stage $g_k$ can achieve 28.91% bandwidth savings or PSNR gains of 2.64dB on the CIFAR10 dataset, and the NTSCC with four-stage $g_k$ can achieve 17.96% and 28.09% bandwidth savings or PSNR gains of 0.81dB and 1.31dB on the Kodak dataset and the CLIC2021 dataset, respectively. In fact, the most appropriate number of Transformer stages in the nonlinear transform module can be different for various source image resolutions, it will be explored in future. As analyzed before, deep JSCC performs worse on high-resolution datasets, thus there are 45.48% and 54.31% bandwidth costs increase.

2) MS-SSIM Performance: Fig. 11(a) and Fig. 11(b) show the RD performance in terms of MS-SSIM at the AWGN channel with channel SNR = 10dB. Since MS-SSIM yields values between 0 (worst) and 1 (best), and most values are higher than 0.9, we converted the MS-SSIM values in dB to improve the legibility. We focus on the high-resolution images. The NTSCC models are trained with $\lambda \in \{1, 1/4, 1/16, 1/32, 1/128\}$ and $\eta = 0.4$, a lower value of $\lambda$ leads to a larger bandwidth ratio $R$. The difference in the value of hyperparameters $\lambda$ and $\eta$ stems from the difference in the loss function value. Results indicate that the proposed NTSCC method outperforms “BPG + Capacity”
and deep JSCC schemes by a large margin, and it achieves a greater improvement on high-resolution images and high CBR regions. Compared to the PSNR results, we can find that the BPG series are inferior to the learning-based schemes because BPG compression is designed to be optimized for squared error with hand-selected constraints [21].

We also provide BD-CBR and BD-MS-SSIM results in Table II. It can be observed that in the high-resolution CLIC2021 dataset, the proposed NTSCC can save up to 64.27% CBR, while deep JSCC only saves 31.77%. Similar improvements can be found in Fig. 11(c) and Fig. 11(d), which demonstrate the MS-SSIM values with a range of channel SNRs.

3) LPIPS Performance: Apart from the above PSNR and MS-SSIM distortion metrics, regarding the goal of semantic communications, we further use the human-perception orientated LPIPS loss function [16] to train the NTSCC model. This learning-based metric is aligned with the perceptual quality. In addition, as aforementioned, when the NTSCC model targets at human perceptual metric, the NTSCC distortion term in the loss function for model training refers to (26). Fig. 12 shows the LPIPS results versus the average CBR, a lower LPIPS indicates a lower distortion. For learning-based schemes (deep JSCC and NTSCC), we mark PSNR and MS-SSIM in parentheses to indicate the model training target. The “NTSCC (Perceptual)” curves in Fig. 12 indicate the performance of NTSCC model optimized for learned perceptual distortion in (26). Clearly, the perceptually optimized NTSCC overpasses other schemes by a large margin.

To intuitively demonstrate the effect of perceptual optimization, we further pick visible results on the testset as shown in Fig. 13 and Fig. 14. From these examples, we can observe that the proposed NTSCC model learned under perceptual loss
### Fig. 13. Examples of visual comparison. The first column shows the original image. The second column shows the cropped patch in original frame. The third to the sixth column show the reconstructed images by using different transmission schemes over the AWGN channel at SNR = 10dB, where the metrics in parentheses indicate the model training target loss function (PSNR and Perceptual). Red number and blue number indicate the percentage of bandwidth cost increase and saving compared to the baseline “BPG + LDPC” scheme.

| Original | BPG + LDPC | Deep JSCE (PSNR) | NTSCC (PSNR) | NTSCC (Perceptual) |
|----------|------------|------------------|--------------|-------------------|
| 512 × 768 | R / PSNR (dB) | 0.043 (0%) / 34.22 | 0.042 (–2%) / 31.10 | 0.038 (–12%) / 34.22 | 0.039 (–9%) / 33.62 |
| 512 × 768 | R / PSNR (dB) | 0.023 (0%) / 31.97 | 0.042 (+3%) / 31.71 | 0.022 (–4%) / 32.27 | 0.023 (0%) / 31.46 |
| 512 × 768 | R / PSNR (dB) | 0.029 (0%) / 23.12 | 0.042 (+4%) / 23.67 | 0.026 (–10%) / 23.44 | 0.027 (–7%) / 22.99 |
| 1512 × 2016 | R / PSNR (dB) | 0.020 (0%) / 23.71 | 0.021 (+3%) / 25.09 | 0.020 (0%) / 26.67 | 0.018 (–10%) / 25.50 |

### Fig. 14. Examples of visual comparison, where the configurations are the same as that in Fig. 13 except for alternating PSNR as MS-SSIM.

| Original | BPG + LDPC | Deep JSCE (MS-SSIM) | NTSCC (MS-SSIM) | NTSCC (Perceptual) |
|----------|------------|-------------------|----------------|-------------------|
| 1499 × 1002 | R / MS-SSIM (dB) | 0.022 (0%) / 12.84 | 0.021 (–5%) / 11.67 | 0.019 (–14%) / 14.78 | 0.022 (0%) / 15.45 |
| 2048 × 1365 | R / MS-SSIM (dB) | 0.042 (0%) / 14.93 | 0.042 (0%) / 15.85 | 0.025 (–40%) / 14.89 | 0.033 (–21%) / 16.68 |
and deleting the additional rate tokens. Since the network needs to transmit as the side information, thus, all comparisons are fair. For the intact scheme of “NTSCC w/o side info. $\bar{z}$", Table III shows the amount of side information as a function of total CBR. The CBR of hyperprior side information $\bar{z}$ grows with the total CBR. The CBR of side information $\bar{k}$ indicating the allocated channel bandwidth of each patch embedding $y_i$ stays identical, and it is quite lower than the total CBR. With respect to the results of Fig. 15 and Table III, we conclude that exploiting the hyperprior side information $\bar{z}$ at the receiver can indeed obtain performance gain while the bandwidth cost for transmitting $\bar{z}$ is relatively high. We can also adopt the NTSCC architecture without transmitting the hyperprior side information $\bar{z}$, due to the lack of decoder refinement on $f_d$, some performance degradation can be observed, but its performance still greatly surpasses traditional deep JSCC schemes. Therefore, the use of hyperprior side information transmission link in Fig. 4 is optional for practical implementations of NTSCC.

V. CONCLUSION

This paper has proposed a new class of high-efficiency joint source-channel coding methods that can learn to closely adapt to the source distribution under the nonlinear transform. It was collected under the name “NTSCC”. Unlike traditional deep JSCC methods, NTSCC first learns a nonlinear analysis transform to map the source data into the latent space, then transmits the latent representation to the receiver via a group of learned variable-length neural JSCC encoders and a noisy channel. In this way, although the source distribution cannot be known by the transceiver in advance, the proposed NTSCC model can learn the source latent representation, as well as an entropy model on the latent representation to implicitly approximate the true source distribution. Accordingly, novel adaptive rate transmission and hyperprior-aided codec refinement mechanisms have been developed to improve end-to-end transmission performance effectively. For the NTSCC model optimization, both the widely-used PSNR/MS-SSIM distortion metrics and the emerging human perceptual metric LPIPS have been adopted, matching the goal of end-to-end semantic communications. In summary, this paper has proposed a promising method to enable the elaborate design of learning-based joint source-channel coding.

REFERENCES

[1] P. Zhang, W. Xu, H. Gao, K. Niu, and et al., “Toward wisdom-evolutionary and primitive-concise 6G: A new paradigm of semantic communication networks,” Engineering, 2021.
