Perceptual Learned Source-Channel Coding for High-Fidelity Image Semantic Transmission

Jun Wang*, Sixian Wang*, Jincheng Dai*, Zhongwei Si*, Dekun Zhou† and Kai Niu*‡

*Key Laboratory of Universal Wireless Communications, Ministry of Education
*Beijing University of Posts and Telecommunications, Beijing, China
†Information & Communication Branch of State Grid Hubei Electric Power Company, China
‡Peng Cheng Laboratory, Shenzhen, China

Abstract—As one novel approach to realize end-to-end wireless image semantic transmission, deep learning-based joint source-channel coding (deep JSCC) method is emerging in both deep learning and communication communities. However, current deep JSCC image transmission systems are typically optimized for traditional distortion metrics such as peak signal-to-noise ratio (PSNR) or multi-scale structural similarity (MS-SSIM). But for low transmission rates, due to the imperfect wireless channel, these distortion metrics lose significance as they favor pixel-wise preservation. To account for human visual perception in semantic communications, it is of great importance to develop new deep JSCC systems optimized beyond traditional PSNR and MS-SSIM metrics. In this paper, we introduce adversarial losses to optimize deep JSCC, which tends to preserve global semantic information and local texture. Our new deep JSCC architecture combines encoder, wireless channel, decoder/generator, and discriminator, which are jointly learned under both perceptual and adversarial losses. Our method yields human visually much more pleasing results than state-of-the-art engineered image coded transmission systems and traditional deep JSCC systems. A user study confirms that achieving perceptually similar end-to-end image transmission quality, the proposed method can save about 50% wireless channel bandwidth costs.

I. INTRODUCTION

End-to-end wireless image transmission systems based on deep joint source-channel coding (deep JSCC) are becoming an active area of research [1], [2]. Compared to traditional image coded transmission systems, such as BPG image compression (the state-of-the-art (SOTA) engineered image compression codec [3]) combined with 5G low-density parity-check (LDPC) channel coded modulation [4], deep JSCC shows competitive performance due to its source image semantics-aware and channel-aware abilities. Despite this, current deep JSCC image transmission systems are typically optimized under traditional distortion metrics such as peak signal-to-noise ratio (PSNR) or multi-scale structural similarity (MS-SSIM) [5]. These simple and shallow metrics fail to account for human perception in terms of many image nuances. Especially for low transmission rates, due to the imperfect wireless channel, traditional distortion metrics, such as mean-squared error (MSE) and MS-SSIM, lose significance as they favor pixel-wise preservation [6]. This is apparently not aligned with human perception in semantic communications, which tend to focus on global semantic information and local texture. To tackle this, it is of great importance to account for human perception during the optimization of deep JSCC systems.

Recent deep learning techniques in computer vision suggest that perceptual metrics and adversarial training strategy play...
important roles in image compression tasks that can well improve the visual quality of reconstructed images. As for the image quality assessment, Zhang et al. [7] has proposed the learned perceptual image patch similarity (LPIPS) model, measuring the Euclidean distance between deep representations. It has been shown that feature maps of different neural network architectures have “reasonable” effectiveness in accounting for human perception of image quality, LPIPS has been widely identified as a generalization of the “perceptual loss”. To improve human visual quality, adversarial losses were employed in image compression tasks [6], [8]–[10]. Motivated by these emerging techniques to augment reconstructed image visual quality, it is the very time to design a new deep JS defense system to coincide with human perception, which is consistent with the goal of semantic communications.

In this paper, we propose a new learned deep JS defense method for end-to-end wireless transmission of full-resolution images. Our method integrates both perceptual loss and adversarial loss to capture global semantic information and local texture thus yielding deep JS defense systems that produce human virtually appealing images. In contrast to SOTA engineered image coded transmission systems and recently proposed deep JS defense systems, our new deep JS defense system combines encoder, decoder/generator, and discriminator together, which can be jointly trained under both perceptual and adversarial losses. Our proposed end-to-end wireless image transmission system tends to preserve the full image, also falling back to the synthesized content instead of blocky/blurry regions caused by the limited channel bandwidth and undetermined channel fading/noise. Extensive experimental results and user study verify that our approach is favorable to yield human perception more pleasing results versus SOTA methods. Achieving the perceptually similar end-to-end image transmission quality, the proposed method can save about 50% wireless channel bandwidth costs. Fig. 1 intuitively shows visual comparisons between our proposed method and other image coded transmission schemes. This is the first work integrating perceptual ingredients and generative method to optimize image coded transmission systems, which is enlightening for future semantic communication system design.

The remainder of this paper is organized as follows. In Section II, we present the proposed generative deep JS defense framework combining the perceptual losses and adversarial losses. In Section III, we elaborate introduce the network architecture and training strategy. In Section IV, we provide the experimental results to quantify the performance gain of the proposed method. Finally, we conclude this paper in Section V.

**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 $\mathbf{x}$, which may also represent a subvector of $\mathbf{x}$ as described in the context. Bold uppercase letters (e.g., $\mathbf{X}$) denote matrices, and $\mathbf{I}_m$ denotes an $m$-dimensional identity matrix. $p_x$ denotes a probability density function (pdf) with respect to the continuous-valued random variable $x$. In addition, $\mathbb{E} [\cdot]$ denotes the statistical expectation operation, and $\mathbb{R}$ denotes the real number set. Finally, $\mathcal{N}(\mu, \sigma^2)$ denotes a Gaussian distribution with mean $\mu$ and variance $\sigma^2$.

## II. System Architecture

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

The proposed generative deep JS defense framework for end-to-end wireless image transmission is illustrated in Fig. 2. The source image formulated as a vector $\mathbf{x} \in \mathbb{R}^m$ is mapped to a vector of continuous-valued channel-input symbols $\mathbf{s} \in \mathbb{R}^k$ via a deep neural network (DNN) based encoding function $\mathbf{s} = f_e(\mathbf{x})$, where $f_e$ can be parameterized as a convolutional neural network (CNN). We typically have $k < m$, and $R = k/m$ is named channel bandwidth ratio (CBR) [1] denoting the overall coding rate. Then, the analog sequence $\mathbf{s}$ is directly sent over the communication channel. The wireless channel introduces random corruptions to the transmitted symbols, denoted as a function $W(\cdot)$. The received sequence is $\hat{s} = W(s)$, whose transition probability is $p_{\hat{s}|s}(\hat{s}|s)$. In this paper, we exemplarily consider the most widely used additive Gaussian noise (AWGN) channel model such that the transfer function is $\hat{s} = W(s) = s + n$ where each component of the noise vector $n$ is independently sampled from a Gaussian distribution, i.e., $n \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_k)$, where $\sigma^2$ is the average noise power. Other channel models can also be similarly incorporated by changing the channel transition function.

The receiver also comprises a generator/decoder function $\hat{\mathbf{x}} = f_g(\hat{s})$ to recover the corrupted signal $\hat{s}$ as $\mathbf{x}$, where $f_g$ can also be a format of CNN [1]. In contrast to traditional deep JS defense, the target of $f_g$ is not only recovering the transmitted image $\mathbf{x}$ to a certain degree, but also trying to generate $\hat{\mathbf{x}}$ that is consistent with the real conditional distribution $p_{\hat{\mathbf{x}}|s}$. The latter is contributed by the adversarial loss given by the discriminator $f_d$ [12] as shown in Fig. 2, this idea comes from conditional generative adversarial networks (cGANs) [13]. Therefore, the objective for our generative deep JS defense can be expressed as

$$
\min_{f_e, f_g} \max_{f_d} \left( \mathbb{E} \left[ - \log f_d(\mathbf{x}, \mathbf{s}) \right] + \mathbb{E} \left[ - \log (1 - f_d(f_g(\hat{s}), \mathbf{s})) \right] \right) + \beta \mathbb{E} \left[ d(\mathbf{x}, f_g(\hat{s})) \right] + \lambda R,
$$

where $d(\cdot)$ denotes a metric to measure the distortion $d(\cdot)$.

![Fig. 2. Architecture of our generative deep JS defense, where the data links denoted by dashed lines are only used for model training.](image-url)
where $\lambda$ balances the trade-off between the CBR against the distortion and GAN loss terms, and $\beta$ balances between the two types of losses. Herein, in deep JSCC models, the CBR term $R$ is predetermined by the input/output dimensions of the encoder network. Even though it is more reasonable to estimate the entropy of codeword sequence as [14] for further saving the bandwidth cost, it also requires explicit entropy-modeled rate as a loss term. We leave the exploration of this issue for future work. In this paper, we also follow the standard model of deep JSCC to manually set $R$, in this simplified setting, the CBR term in (1) is indeed a fixed term.

Given the CBR $R$, our generative deep JSCC model training is divided into two phases in each iteration. The goal is to create a pair of encoder $f_e$ and generator $f_g$ to "fool" the discriminator $f_d$ into believing the reconstructed image samples $\hat{x}$ are real while also considering the influence of an imperfect wireless channel. The loss functions are written as follows.

- Training the encoder $f_e$ and generator $f_g$ with the loss function $L_{E,G}$:

$$
\min_{f_e,f_g} L_{E,G} = \frac{1}{2} \mathbb{E}_{x,s} \left[ \log \mathbb{P}( s \ | \ x ) \right] + \beta_d \mathbb{E}_{x,\hat{s}} \left[ -\log f_d( f_g(\hat{s}), s ) \right] + \\
\beta_p \mathbb{E}_{x,\hat{s}} [d_{\text{Percept}}(x, f_g(\hat{s}))] + \beta_m \mathbb{E}_{x,\hat{s}} [d_{\text{MSE}}(x, f_g(\hat{s}))]
$$

(2)

The discriminator function $f_d$ in (2) includes two inputs which obey the format of cGANs [13]. Each data sample $\hat{s} = f_g(\hat{s})$ is associated with the channel corrupted codeword $\hat{s}$ as the condition. The generator $f_g$ is trained to map the received samples $\hat{s}$ from a fixed distribution $p_{\hat{s}|x}$ to the real distribution $p_{x|s}$, where $p_{\hat{s}|x}$ is determined by the encoder function $f_e$ and channel transition function $W$. $d_{\text{Percept}}$ denotes the perceptual LPIPS loss function [7]. As the learning-based LPIPS metric has many different configurations, we exemplarily adopt the Alexnet-based [15] LPIPS loss to help model training in a perception-aware way. $d_{\text{MSE}}$ denotes the regular MSE distortion function. $\beta_d, \beta_p, \beta_m$ control the trade-off between three distortion terms.

- Training the discriminator $f_d$ with the loss function $L_D$:

$$
\max_{f_d} L_D = \min_{f_e,f_g} \left( \mathbb{E}_{x,\hat{s}} \left[ -\log (1 - f_d( f_g(\hat{s}), s ) ) \right] + \mathbb{E}_{x,\hat{s}} [ d_{\text{Percept}}(x, f_g(\hat{s})) ] \right)
$$

(3)

The channel corrupted sequence $\hat{s}$ is associated with the source image sample $x$ in (3), i.e., $\hat{s} = W(x)$. The discriminator $f_d$ is mapping the input tuple $(x,s)$ to the probability that it is a sample from the real distribution $p_{x|s}$ rather than the fake sample generated by $f_g$.

### III. NETWORK ARCHITECTURE

Our network architecture is shown in Fig. 3, including the encoder $f_e$, generator $f_g$, and discriminator $f_d$. The encoder $f_e$ uses convolution neural networks (CNNs) to extract features of the source image $x$ and reshape them to power normalized channel-input symbols $s$. The decoder $f_g$ performs the inverse operations of the encoder.

Our discriminator network $f_d$ is basically the same as that in PatchGAN [16], which penalizes structure at the scale of image patches, i.e., it tries to identify if each patch of the reconstruction is real or fake. The patch size is set to $16 \times 16$ pixels in our implementation. Besides, we also make use of the structure of cGAN [13] by concatenating the reconstructed image with an upscaled deep JSCC codeword to provide the contextual information. In particular, four cascaded modules of the form “convolution-SpectralNorm-LeakyReLU” [17] are used to downsample the concatenated feature map progressively, and the number of kernels in convolutions $C_i$ is 64, 128, 256, and 512, respectively. After the last layer, a convolution with the Sigmoid activation indicates the discriminator's true or false judgments for each patch.
During the model training, we firstly pretrain the encoder \( f_e \) and generator \( f_g \) only using the perceptual loss \( d_{\text{LPIPS}} \) and the regular MSE loss \( d_{\text{MSE}} \). The function of the pretraining step is to accelerate the experiments and makes them more stable. In the second step, we freeze the encoder \( f_e \) and generator \( f_g \) and only optimize the discriminator \( f_d \), which enables it to penalize structure at image patches produced by pretrained weights. Finally, we adopt an iterative approach to alternately training the generative network (\( f_e \) and \( f_g \)) and the discriminator network (\( f_d \)).

IV. EXPERIMENTAL RESULTS

A. Experimental setup

1) Training Details: The training set consists of 720,000 images sampled from the Open Images dataset [18]. During training, we form a batch of 12 patches by randomly cropping a \( 256 \times 256 \) patch from each image. We first pretrain the network except for the discriminator \( f_d \) by about 200,000 iterations jointly using the LPIPS perceptual loss and the regular MSE loss with \( \beta_p = 1.0 \) and \( \beta_m = 0.01 \) under a uniform distribution of \( \text{SNR}_{\text{train}} \in \{1, 4, 7, 10, 13\} \) dB. Then, we freeze the encoder and generator and optimize the discriminator for about 10,000 iterations. Finally, we alternately train the generative network (\( f_e \) and \( f_g \)) by setting \( \beta_g = 0.5 \). We use the Adam optimizer with a learning rate of \( 10^{-4} \).

2) Evaluation Protocols: We evaluate our method on the widely-used Kodak dataset [19] (24 images, \( 768 \times 512 \) pixels) and the test set of the CLIC2021 dataset [20] (60 images, up to \( 2048 \times 1890 \) pixels). The comparison schemes include the deep JSCC optimized for MSE (marked as “deep JSCC (MSE)”) and the classical coded transmission schemes (BPG + LDPC). In particular, for the deep JSCC scheme, we adopt the Deep JSCC (no feedback) scheme in [2], which achieves better performance than the original paper [1]. For traditional coded transmission schemes, we employ the BPG [3] image codec for source coding and LDPC followed by quadrature amplitude modulation (QAM) for channel coding and modulation. We further adopt BPG combined with an ideal capacity-achieving channel code, which can be viewed as the performance upper bound of the traditional coded transmission scheme. For brevity, “BPG + LDPC” and “BPG + Capacity” refers to the two comparison schemes, respectively.

3) Metrics: We evaluate our models and the comparison schemes in classical metric PSNR and MS-SSIM [5] as well as the emerging deep learning based perceptual metric LPIPS [7]. DIISTS [21], FID [22]. PSNR corresponds to the regular MSE distortion in (2). MS-SSIM is the most widely used perceptual metric despite failing to account for many nuances of human perception [7]. LPIPS and DIISTS both can predict well with human perceptual judgment by computing the distance in the feature space of the image classification network. FID is
widely used to assess the quality and diversity of generated images, especially from GAN, which measures the similarity between the distribution of reference images and generated images in the feature space of an Inception network. As the image resolutions vary in a large range, we use the non-overlapping 2155 patches size of 256 pixels for the CLIC2021 dataset and 144 same size patches for the Kodak dataset.

Inspired by [6], we also carry out a user study to more holistically evaluate the visual quality of our results. Considering the tiny distortion between the original image and the reconstructed image at the high CBR, we design the user study for two models with CBR = 1/48 and CBR = 1/24 and BPG at CBR ranging from 0.029 to 0.058 at SNR = 10dB. We crop the reconstructed images on the Kodak dataset to 256 × 256 patches which are respectively produced from our models and BPG + LDPC. Then we compose a questionnaire comprising 46 reconstructed patch pairs. Each patch pair include two patches produced by different methods but from the same patch. 20 random participants are asked to choose which patch corresponds to human visual perception. We report the mean percentage of preference for each choice.

B. Results Analysis

1) Quantitative Comparison: Fig. 4 shows SNR-distortion and SNR-perception results at the CBR = 1/48 on the Kodak dataset and the CLIC2021 testset. Solid red line and solid blue line plot the performance of our model and deep JSCC, where the training SNR value is consistent with the testing SNR value. For the BPG + LDPC scheme, we present the envelope of the best performing configurations at each SNR among different combinations of LDPC coding rate and modulations. Moreover, as shown in Fig. 5, we further present the CBPerception and CBPer-Perception curve over AWGN channel at SNR = 10dB on CLIC2021. For the BPG + LDPC scheme, following [2], we use a 2/3 rate (4096, 6144) LDPC code with 16-ary QAM to ensure reliable transmission and adjust the BPG to the largest compression rate within the CBR constraint.

Results indicate that our model dominates in terms of all three perceptual quality indices (LPIPS, DISTS, and FID). In particular, our scheme can achieve comparable or even better perceptual quality values than other approaches while saving 1.5x to 2.5x bandwidth cost in the same SNR or otherwise achieving 3dB to 5dB SNR gain using the same CBR. It validates the effectiveness of the proposed compound loss function in (1). Even though perceptual metrics match better with human judgments [23], more convincing results can be made through subjective testing in the following user study. Besides, we also note that our scheme presents some negligible performance degradation in the classic metrics metric compared to the MSE optimized deep JSCC scheme. These verify that our scheme prefers to generate clear textures resulting in lower pixel-wise metrics than comparison schemes. Comparing our model to the baseline (no GAN only using same distortion loss), the performance of classic metrics become slightly worse and the perceptual performance (FID and DISTS) improves which indicates the GAN loss can improve the perceptual quality.

2) User Study: Fig 6 shows the mean preference percentage comparing our proposed models to BPG + LDPC on the Kodak dataset at channel SNR = 10dB. The upper figure illustrates the results at the CBR = 1/48. It can be seen that our method is the preferred scheme even if the CBR cost is 47% smaller than that of BPG + LDPC. It is well known that the BPG suffers blocky artifacts, which are unfriendly to human visual perception, but our method can well preserve global semantic information and synthesizes realistic texture. The lower figure shows the results at the CBR = 1/24, the overall image quality improvement leads to a relatively smaller gain of our model, but it is still preferred to the BPG + LDPC scheme when saving 18% channel bandwidth cost. These results verify that our perceptual learned deep JSCC can reconstruct higher visual quality images by using less channel bandwidth cost.

C. Model Variants

As one type of ablation study, we explore the effects of introducing perceptual loss $d_{\text{LPIPS}}$ and adversarial loss. Firstly we set the $\beta_g = 0$ and compare the reconstructed images produced respectively from our models with $\beta_m = 0.1, 0.01$ to the deep JSCC model in [2]. Fig. 7 indicates that when $\beta_m$ decreases, the blurring artifacts are removed gradually. We further introduce the adversarial loss respectively setting $\beta_g = 0.05, 0.5, 5.0$ in our model. We can see that the more realistic details appear in the image with the help of the discriminator. Our GAN can synthesize some clear local textures to achieve better visual quality. However, at $\beta_g = 5.0$ we find significant griding artifacts appear on generative images. We think the main reason is probably the GAN synthesizes too fake details. To explore the effect of the perceptual loss, we only add the discriminator to the deep JSCC model in [2] and use the same training strategy depicted in section II. The
results show that only adding GAN also improves the visual quality though still missing some textures compared to our proposed model. These results verify that our generative deep JSCC can well reconstruct closer to human visual perception images by combing the perceptual loss and adversarial loss.

V. CONCLUSION

This paper has proposed a new deep JSCC system optimized toward human visual perception for semantic communications. It has introduced perceptual and adversarial losses, enabling the proposed model to capture global semantic information and local texture. Visualization results have demonstrated that our method can produce visually pleasing images instead of blurring/blocky artifacts for low transmission rates. Extensive numerical results and user study have verified that the proposed scheme can generally save communication channel bandwidth costs compared to SOTA engineered image coded transmission systems as well as the standard deep JSCC.

REFERENCES

[1] Eirina Bourtsoulatze, David Burth Kurka, and Deniz Gündüz, “Deep joint source-channel coding for wireless image transmission,” IEEE Transactions on Cognitive Communications and Networking, vol. 5, no. 3, pp. 567–579, 2019.
[2] David Burth Kurka and Deniz Gündüz, “Deepjssc-f: Deep joint source-channel coding of images with feedback,” IEEE Journal on Selected Areas in Information Theory, vol. 1, no. 1, pp. 178–193, 2020.
[3] Fabrice Bellard, “BPG image format,” URL: https://bellard.org/bpg/.
[4] Tom Richardson and Shrinivas Kudekar, “Design of low-density parity check codes for 5G new radio,” IEEE Communications Magazine, vol. 56, no. 3, pp. 28–34, 2018.
[5] Z. Wang, E. P Simoncelli, and A. C Bovik, “Multiscale structural similarity for image quality assessment,” in Proceedings of the Thirty-Seventh Asilomar Conference on Signals, Systems & Computers. IEEE, 2003, vol. 2, pp. 1398–1402.
[6] Eirikur Agustsson, Michael Tschannen, Fabian Mentzer, Radu Timofte, and Luc Van Gool, “Generative adversarial networks for extreme learned image compression,” in 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 221–231.
[7] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 586–595.
[8] F. Mentzer, G. D Toderici, M. Tschannen, and E. Agustsson, “High-fidelity generative image compression,” Advances in Neural Information Processing Systems, vol. 33, 2020.
[9] O. Rippel and L. Bourdev, “Real-time adaptive image compression,” in International Conference on Machine Learning. PMLR, 2017, pp. 2922–2930.
[10] S. Santurkar, D. Budden, and N. Shavit, “Generative compression,” in 2018 Picture Coding Symposium (PCS). IEEE, 2018, pp. 258–262.
[11] B. Xu, N. Wang, T. Chen, and M. Li, “Empirical evaluation of rectified activations in convolutional network,” arXiv preprint arXiv:1505.00853, 2015.
[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” Advances in neural information processing systems, vol. 27, 2014.
[13] M. Mirza and S. Osindero, “Conditional generative adversarial nets,” arXiv preprint arXiv:1411.1784, 2014.
[14] Jincheng Dai, Sixian Wang, Kailin Tan, Zhongwei Si, Xiaoqi Qin, Kai Niu, and Ping Zhang, “Nonlinear transform source-channel coding for semantic communications,” IEEE Journal on Selected Areas in Communications, vol. 40, no. 8, pp. 2300–2316, 2022.
[15] Eirina Bourtsoulatze, David Burth Kurka, and Deniz Gündüz, “Deepjssc-f: Deep joint source-channel coding of images with feedback,” IEEE Journal on Selected Areas in Information Theory, vol. 1, no. 1, pp. 178–193, 2020.
[16] C. Li and M. Wand, “Precomputed real-time texture synthesis with markovian generative adversarial networks,” in European conference on computer vision. Springer, 2016, pp. 702–716.
[17] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, “Spectral normalization for generative adversarial networks,” arXiv preprint arXiv:1802.05957, 2018.
[18] A. Krizhevsky, I. Sutskever, and G. E Hinton, “Imagenet classification with deep convolutional neural networks,” Advances in neural information processing systems, vol. 25, 2012.
[19] M. Mirza and S. Osindero, “Conditional generative adversarial nets,” arXiv preprint arXiv:1411.1784, 2014.
[20] CLIC 2021: Challenge on learned image compression,” in Proceedings of the Thirty-Seventh Asilomar Conference on Signals, Systems & Computers. IEEE, 2003, vol. 2, pp. 1398–1402.
[21] M. Mirza and S. Osindero, “Conditional generative adversarial nets,” arXiv preprint arXiv:1411.1784, 2014.
[22] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 586–595.
[23] F. Mentzer, G. D Toderici, M. Tschannen, and E. Agustsson, “High-fidelity generative image compression,” Advances in Neural Information Processing Systems, vol. 33, 2020.
[24] O. Rippel and L. Bourdev, “Real-time adaptive image compression,” in International Conference on Machine Learning. PMLR, 2017, pp. 2922–2930.
[25] S. Santurkar, D. Budden, and N. Shavit, “Generative compression,” in 2018 Picture Coding Symposium (PCS). IEEE, 2018, pp. 258–262.
[26] B. Xu, N. Wang, T. Chen, and M. Li, “Empirical evaluation of rectified activations in convolutional network,” arXiv preprint arXiv:1505.00853, 2015.