GENERATIVE MODEL BASED HIGHLY EFFICIENT SEMANTIC COMMUNICATION APPROACH FOR IMAGE TRANSMISSION

Tianxiao Han[1], *Jiancheng Tang*[1], *Qianqian Yang*[1], *Yiping Duan*[2], *Zhaoyang Zhang*[1], *Zhiguo Shi*[1]

[1]College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China
[2]Department of Electronic Engineering, Tsinghua University, Beijing, China

ABSTRACT

Deep learning (DL) based semantic communication methods have been explored to transmit images efficiently in recent years. In this paper, we propose a generative model based semantic communication to further improve the efficiency of image transmission and protect private information. In particular, the transmitter extracts the interpretable latent representation from the original image by a generative model exploiting the GAN inversion method. We also employ a privacy filter and a knowledge base to erase private information and replace it with natural features in the knowledge base. The simulation results indicate that our proposed method achieves comparable quality of received images while significantly reducing communication costs compared to the existing methods.

Index Terms— Semantic communion, Generative model, Image transmission

1. INTRODUCTION

With the explosion of visual content in our daily life, such as pictures, movies, and ever the undergoing metaverse applications, the efficient transmission of images becomes more important to improve the performance and experience of the current wireless communication systems. Semantic-oriented communication methods have brought a tremendous improvement in transmission efficiency because of the joint extraction and compression of the semantic information from the source leveraging deep learning methods. In [1], the authors are the first to apply the autoencoder model for joint source and channel coding and achieve better transmission efficiency of images. The authors [2] proposed a multi-level Semantic communication system to extract both the high-level and low-level semantic information of an image. However, the compression ratio of all these works ranges between $1/60$ and $1/20$, which may be insufficient due to the explosion of demands in the near future. A task-specific semantic communication system was proposed in [3] for image transmission targeting person re-identification, which achieves a compression ratio of $1/128$. But it considers the specific task and is not able to reconstruct the image at the receiver. In this paper, we aim to propose a highly efficient semantic communication method for image transmission with a low compression ratio while preserving the perceptual quality.

The generative adversarial network (GAN) models have the capability to generate high dimensional contents from a low dimension vector. Some existing works on semantic communication systems exploit this property of GAN in the image transceivers to achieve a small compression ratio. The authors in [4] use conditional GAN at the receiver to reconstruct the image, which requires the transmitter to send the semantic segmentation labels and the residual images, resulting in additional transmission overhead. The authors in [5] use GAN along with the Deep JSCC model [1] to achieve better transmission efficiency and perceptual metrics. However, the smallest compression ratio achieved so far is $1/50$.

In this paper, we exploit the recent advance of interpretable generative models [6] [7], in particular, semantic StyleGAN proposed in [8], to extract disentangled semantic information from input images and further improve the transmission efficiency. This is achieved by training the network with segmentation labels in a supervised manner. For example, the different parts in latent codes obtained by semantic StyleGAN trained on the human face dataset present different parts of a human face. We also take into account the privacy concern of the image transmission. Note that the straightforward method is to erase the private part and transmit the masked images directly. However, it results in unnatural perception to the viewer because the removal of private parts may make the generated images fall out the distribution of natural images [9]. We tackle this issue by ensuring the modified latent code to be within the latent space [10], in order to protect private information while reconstructing the natural image simultaneously.

The contributions of this paper can be summarized as follows: (i) A generative model based semantic communication framework is proposed with the inversion method of semantic StyleGAN to extract semantic information and a normalizing flow model with its inversion to help achieve an extremely small compression ratio, i.e., $1/3072$, which is only 1% of existing works. (ii) We demonstrate that manipulating the latent codes of semantic StyleGAN with a privacy filter and a knowledge base helps reconstruct images naturally while

protecting privacy. (iii) As numerical experiments show, we achieve a comparable quality of received images while significantly reducing communication costs compared to the existing methods, especially under harsh communication conditions.

2. SYSTEM MODEL

In this section, we present the system model of the considered semantic communication system for privacy-preserving image transmission. We also introduce the metrics to evaluate the performance of the proposed model.

2.1. Transmitter

As shown in Fig. 1, the input image $\boldsymbol{F}$ is first transformed into a latent representation $\boldsymbol{L}$ by Semantic StyleGAN with the inversion method. Then a privacy filter is applied to filter out the privacy information and obtain a privacy-preserving latent representation $\boldsymbol{L'}$. Then we select features in the semantic knowledge base to replace the erased information and get $\boldsymbol{P}$, which helps the natural reconstruction of the images with the private information replaced. Normalizing flow transformation [11] is applied on $\boldsymbol{P}$ to get a latent representation $\boldsymbol{N}$ with a distribution that are optimized for transmission and reconstruction. Finally, the channel encoder maps the $\boldsymbol{N}$ into symbol sequences $\boldsymbol{X}$ to be transmitted that can be robust to imperfect channel.

2.2. Receiver

The received signal at the receiver is given by

$$\boldsymbol{Y} = \boldsymbol{h} * \boldsymbol{X} + \boldsymbol{n}, \tag{1}$$

where $\boldsymbol{h}$ represents the channel coefficients, and $\boldsymbol{n}$ denotes the Gaussian noise of channel. The received signal, $\boldsymbol{Y}$, is then mapped back into the latent semantic representation $\widehat{\boldsymbol{L}}$ by a channel decoder and the inverse normalizing flow, which is then converted into the reconstructed images $\widehat{\boldsymbol{F}}$ by the generator of Semantic StyleGAN.

2.3. Metrics

We evaluate our system performance with both the objective and subjective metrics. For the objective evaluation, we compare the input images and reconstructed images at the pixel level with Peak Signal to Noise Ratio(PSNR). PSNR is calculated by the Mean Square Error(MSE) between the two images, denoted by

$$PSNR = 10 \times log_{10} \frac{Max(F, \widehat{F})^2}{MSE(F, \widehat{F})}, \tag{2}$$

where $Max(F, \widehat{F})$ is the maximum possible pixel value of the image $\boldsymbol{F}$ and $\widehat{\boldsymbol{F}}$. We also use the deep learning based perceptual metric, LPIPS[12], to evaluate our system performance. LPIPS evaluate the reconstructed image from human perceptual perspective by computing the distance between the original and the reconstructed images in the feature space derived by the pretrained VGG network. We have

$$LPIPS(F, \widehat{F}) = \sum_l \frac{1}{H_l W_l} \sum_{i,j} \| c_l \odot (f_{ij}^l - \hat{f}_{ij}^l) \|_2^2, \tag{3}$$

where f^l and $\hat{f}^l$ denote the normalized latent feature map output by layer l of VGG network with F and $\widehat{F}$ as input, respectively. H_l, W_l and subscript ij denote the height, width and (i, j)th element of the feature map, respectively. Notation $\odot$ denotes the scale operation and c_l represents the pretrained weights for the features in layer l which is used to scale the feature map.

3. PROPOSED METHOD

In this section, we present the details of the proposed model depicted in Fig. 1. We first introduce the proposed generative models based transmission scheme and then the privacy-preserving mechanism to protect the privacy of the transmitted images by utilizing a privacy filter and a knowledge base.

3.1. Generative Model Based Transmission

In this paper, we adopt semantic StyleGAN[8], which generates a latent representation of the input images with different parts representing different semantic information. This is achieved by training this generative model with semantic segmentation labels as supervised signals, where a dual-branch discriminator models the joint distribution of RGB images and semantic segmentation labels. Due to this property, we exploit the inversion method of a pretrained semantic StyleGAN [13] to extract the disentangled semantic information in latent spaces from the input image. The inversion method works as follows: it starts with an initial vector in the latent space, which is then input into the Semantic StyleGAN to generate an image. Then we calculate the MSE loss between the generated image and input image $\boldsymbol{F}$, by which we derive gradient descents of the vector in the latent space. After several iterations, we the optimal latent code of the input image in the latent space and output this code as the latent representation $\boldsymbol{L}$.

However, it is hard to combine the semantic StyleGAN with the channel encoder and channel decoder in an end-to-end method during training, because the training process of semantic StleGAN, which consists of 28 feature generators with 10 FC layers each, along with a stack of convolution layers, is quite slow to train and often fails to converge. Therefore, we need diffeomorphic mapping [14] from the semantic latent code to a new and easily trainable representation. We adopt a normalizing flow model [11] to learn an invertible

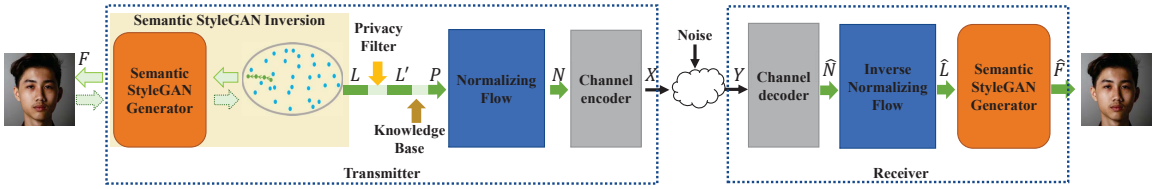
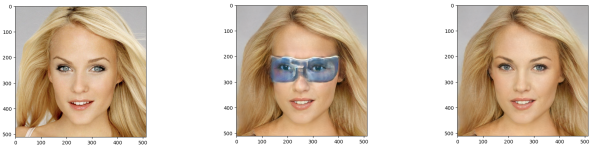


Fig. 1. The overall architecture of the proposed generative model based semantic communication system for image transmission.

mapping between latent spaces of semantic StyleGAN and a group of Gaussian distributions. At the receiver side, we then use the inverse normalizing flow model to map the received X back to the semantic latent representation $\widehat{\boldsymbol{L}}$, by taking advantage of the invertible property of normalizing flow in computational efficiency. More specially, we use the RealNVP model proposed in [15], which consists of 8 coupling layers, and each coupling layer consists of 6 FC layers. The RealNVP is quite simple but efficient to train compared to semantic StyleGAN. As revealed by experiments, the usage of normalizing flow helps improve the quality of reconstructed images.

We then employ relatively simple channel encoder and decoder, each consisting of two cascaded FC layers, are used to map latent codes into symbol sequences $\boldsymbol{X}$ at the transmitter, and map received symbol sequences $\boldsymbol{Y}$ back to latent representation $\widehat{\boldsymbol{N}}$ at the receiver.



(a) Input image from Celeb-HQ dataset [16]

(b) Reconstructed image w/o KB

(c) Reconstructed image with KB

Fig. 2. Generated images with and without the knowledge base(KB).

3.2. Privacy preserving mechanism

We include a privacy preserving mechanism by exploiting a privacy filter and a knowledge base. Since private parts and non-privacy parts in latent representation L for are disentangled, we design a privacy filter that adds a bias to the private parts and obtain a privacy-preserving representation $\boldsymbol{L'}$. The bias in the privacy filter is set to ensure the Euclidean distance between the reconstructed image and the input image to be larger than a given threshold, i.e.,

$$d(\boldsymbol{L'}, \boldsymbol{L}) > \lambda_1 * \boldsymbol{L}, \tag{4}$$

where λ_1 is a task-specific parameter.

However, the randomly added bias may cause the latent representation $\boldsymbol{L'}$ to fall into an out-of-distribution area and generate an unreasonable feature, as the Fig. 2 shows. Therefore, we further exploit a semantic knowledge base (KB) to replace the private parts with certain latent codes that generate a natural image suitable for human perception. The semantic KB uses all the real images in the dataset to obtain the average latent codes $\boldsymbol{L}_m$, which is a guideline for a reasonable generation. Then we add another bias to $\boldsymbol{L'}$ and obtain a constrained representation $\boldsymbol{P}$. The added bias is chosen such that the distance between the reconstructed image and the natural images while forming an upper bound for the distance between the reconstructed image and the privacy-protected images, i.e.,

$$d(\boldsymbol{P}, \boldsymbol{L'}) < \lambda_2 * d(\boldsymbol{P}, \boldsymbol{L}_m), \tag{5}$$

where λ_2 is less than 1 and can control the diversity of reconstructed image. It is noted that the bias added start with a small value and gradually increase until (5) is satisfied.

3.3. Training Methods

We use a two-stage training method, which achieves better performance. In the first stage, we train the normalizing flow and ignore the channel encoder and decoder by directly inputting the output of the normalizing flow at the transmitter to the inverse normalizing flow at the receiver. We use MSE loss functions between the reconstructed image and the input image, together with the normalizing flow loss. In the second stage, We train the whole network together with an additional MSE loss function between $\boldsymbol{N}$ and the received $\widehat{\boldsymbol{N}}$ and use a training method during which the noise gradually increases to prevent the model from crashing or collapse at the beginning.

4. SIMULATION RESULTS AND ANALYSIS

In this section, we evaluate the performance of the proposed semantic communication system for image transmission in terms of the transmission efficiency and privacy preserving. For simplicity, We assume the AWGN channel. We use the Celeb-HQ dataset [16] for training and testing, which is a human face dataset with semantic labels. We use the existing semantic communication approach by [1], referred to as Deep JSCC, as the benchmark approach to compare to. We

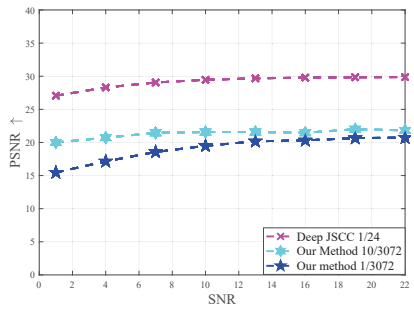


Fig. 3. PSNR versus SNR for different approaches.Our methods' compression ratio are $1/3072$ and $10/3072$, while the compression ratio of Deep JSCC is $1/24$.

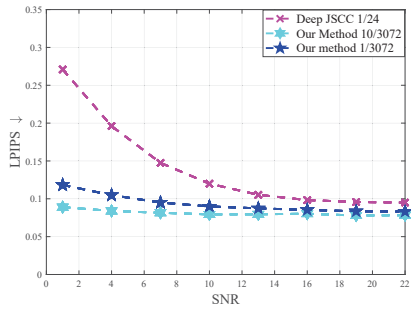


Fig. 4. LPIPS versus SNR for different approaches. Our methods' compression ratio are $1/3072$ and $10/3072$, while the compression ratio of Deep JSCC is $1/24$

note that all approaches are trained and tested with the same datasets. During training, we set channel conditions with SNR varying randomly from 5dB to 10dB. The compression ratio is also defined in [1], which is the ratio of the dimension of vectors (k) to be transmitted and the size of the input images (n).

The performance comparison of different approaches in terms of PSNR and LPIPS is presented in Fig. 3 and Fig. 4. Higher PSNR presents better pixel level restoration, while smaller LPIPS scores represent better objective perception restoration. We can observe that with only 1% the compression ratio of the existing method, our method achieves better LPIPS scores, and slightly worse PSNR. This demonstrates that semantic information we extract and transmit is highly compact and preserve the perceptual contents. We also observe that PSNR and LPIPS by our approach stay almost the same under different SNRs, which illustate the robustness of the proposed semantic communication system to noisy channel. We also compare the numerical results achieved by the proposed method with and without the normalizing flow in Fig. 5, which proves the benefits of normalizing flow in improving the quality of image transmission.

We also evaluate the privacy filter's effect in Fig. 6, where Fig. 6(a) is the original image. By assuming the eyes are

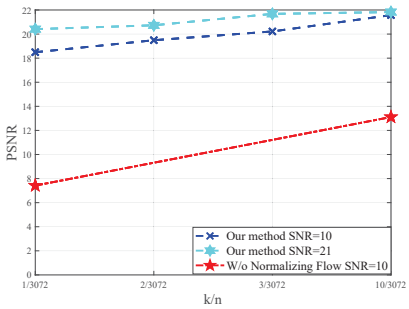
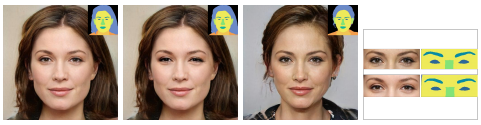


Fig. 5. PSNR versus k/n and the PSNR without normalizing flow.



(a) (b) (c) (d)

Fig. 6. Example of the privacy protection.

the private information to protect, and we can get Fig. 6(b), where all the other parts stay the same while the eyes are different. For another case, assuming all the features are private information except eyes, and we can get Fig. 6(c). By comparing the eyes of the input image and reconstructed image, as shown in Fig. 6(d), we can see that similar eyes are kept while all the other information is changed. This shows the disentanglement of different semantic parts helps the protection of privacy.

5. CONCLUSIONS

In this paper, we proposed a novel generative model based image transmission approach with the ability to significantly reduce the transmission overhead while preserving the private information. We utilized the inversion method of Semantic StyleGAN to acquire the disentangled latent codes of input images, and proposed a privacy filter that modifies the latent codes with the Euclidean distance rule with the help of the knowledge base. Simulation results demonstrated that our proposed method achieves significantly better transmission efficiency with less than 1% of the compression ratio by the existing method while achieving the comparable reconstruction quality and able to preserve private information.

6. 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] Zhenguo Zhang, Qianqian Yang, Shibo He, Mingyang Sun, and Jiming Chen, "Wireless transmission of images with the assistance of multi-level semantic information," *arXiv preprint arXiv:2202.04754*, 2022.

[3] Mikolaj Jankowski, Deniz Gündüz, and Krystian Mikolajczyk, "Wireless image retrieval at the edge," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 1, pp. 89–100, 2020.

[4] Danlan Huang, Xiaoming Tao, Feifei Gao, and Jianhua Lu, "Deep learning-based image semantic coding for semantic communications," in *2021 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2021, pp. 1–6.

[5] Jun Wang, Sixian Wang, Jincheng Dai, Zhongwei Si, Dekun Zhou, and Kai Niu, "Perceptual learned source-channel coding for high-fidelity image semantic transmission," *arXiv preprint arXiv:2205.13120*, 2022.

[6] Tero Karras, Samuli Laine, and Timo Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.

[7] Junling Liu, Yuexian Zou, and Dongming Yang, "Semanticgan: Generative adversarial networks for semantic image to photo-realistic image translation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 2528–2532.

[8] Yichun Shi, Xiao Yang, Yangyue Wan, and Xiaohui Shen, "Semanticstylegan: Learning compositional generative priors for controllable image synthesis and editing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11254–11264.

[9] Blaž Meden, Peter Rot, Philipp Terhörst, Naser Damer, Arjan Kuijper, Walter J Scheirer, Arun Ross, Peter Peer, and Vitomir Štruc, "Privacy–enhancing face biometrics: A comprehensive survey," *IEEE Transactions on Information Forensics and Security*, 2021.

[10] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath, "Generative adversarial networks: An overview," *IEEE signal processing magazine*, vol. 35, no. 1, pp. 53–65, 2018.

[11] Ivan Kobyzev, Simon JD Prince, and Marcus A Brubaker, "Normalizing flows: An introduction and review of current methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 11, pp. 3964–3979, 2020.

[12] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[13] Rameen Abdal, Yipeng Qin, and Peter Wonka, "Image2stylegan: How to embed images into the stylegan latent space?," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4432–4441.

[14] John Milnor and David W Weaver, *Topology from the differentiable viewpoint*, vol. 21, Princeton university press, 1997.

[15] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio, "Density estimation using real nvp," *arXiv preprint arXiv:1605.08803*, 2016.

[16] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo, "Maskgan: Towards diverse and interactive facial image manipulation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.