Cross-Modal Text Steganography Against Synonym Substitution-Based Text Attack

Wanli Peng, Tao Wang, Zhenxing Qian, *Senior Member, IEEE*, Sheng Li, and Xinpeng Zhang, *Member, IEEE*

***Abstract*—Steganography has received massive attention from the information-hiding community due to its excellent security for covert communication systems. Existing work focuses on improving security on single-modal media while cross-modal media is less explored. However, cross-modal interaction has become a prevalent social manner on current social networks, which arises potential behavioral security issues of single-modal steganography. In this letter, we propose a novel text steganography to explore the practicability of cross-modal steganography. The proposed scheme is composed with image encoder, message encoder, language model, and message extractor networks, where the generated stego texts are semantically consistent with the input reference image. In addition, current generative text steganography schemes are vulnerable to text attack based on synonym substitution since these heuristic algorithms embed information by constructing a mapping between secret messages and candidate tokens. Thus, we design a text attack layer based on synonym substitution to further improve the robustness of generated stego text. Experiments illustrate the superior performance of the proposed cross-modal steganography scheme in terms of security and robustness.**

***Index Terms*—Cross-modal, text steganography, text attack, synonym substitution.**

I. Introduction

OVER the past decade, with the flourishing of online social networks, communication devices, and mobile intelligent terminals, people tend to use not single-modal data but multi-modal data to record everyday life, express different opinions, and make daily communication, etc. Compared with single-modal data, multi-modal data can more accurately deliver immediate emotion and attitude to other people, and more comprehensively show the details of moments shared on online social networks. As a result, exploiting multi-modal data to enhance the performance of classic tasks has emerged as a promising research field in the computer vision and natural language processing communities [1], [2], [3], [4]. Unfortunately, there are few analogous explorations of information-hiding community.

Manuscript received 9 January 2023; revised 1 March 2023; accepted 15 March 2023. Date of publication 17 March 2023; date of current version 31 March 2023. This work was supported by the National Natural Science Foundation of China under Grants U20B2051, U22B2047, U1936214, and 62072114. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Victor Sanchez. *(Corresponding author: Zhenxing Qian.)*

The authors are with the School of Computer Science, Fudan University, Shanghai 200082, China, and also with the Key Laboratory of Culture and Tourism Intelligent Computing of Ministry of Culture and Tourism, Fudan University, Shanghai 200082, China (e-mail: pengwanli@fudan.edu.cn; 22210240296@m.fudan.edu.cn; zxqian@fudan.edu.cn; lisheng@fudan.edu.cn; zhangxinpeng@fudan.edu.cn).

Digital Object Identifier 10.1109/LSP.2023.3258862

Steganography, the crux of covert communication system, generally refers to the technology and science of embedding secret messages into ubiquitous multimedia, while introducing the inevitable distortion to cover media as little as possible [5], [6]. Recently, most research on steganography has focused on single-modal media, including image [7], [8], [9], text [10], [11], [12], [13], video [14], audio [15], [16], etc. These single-modal steganography schemes have achieved elegant security performance, especially resisting machine eavesdroppers, i.e., the steganalysis tools based on deep learning [17], [18], [19], [20], [21]. However, if Alice (sender) and Bob (receiver) always use single-modal data to transmit a secret message on online social networks where most people generally utilize multi-modal data, the covert communication between Alice and Bob should be easily judged as an anomalous manner by Eve (eavesdropper). You may have an intuitive question: Can we solve this potential security issue by using different single-modal steganography methods to transfer secret messages on social networks?

Roughly speaking, it seems to be wise to use various single-modal steganography schemes on covert communication to reduce unwanted attention. While current single-modal steganography ignores semantic relevance between different modal data, which can not ensure semantic consistency between the generated different media since the commonly used generators are unconditional generation models, so as to arise the suspicion of eavesdroppers. In order to tackle this problem, Hu et al. [22], [23] proposed multi-modal steganography based on image-text matching network. These methods are essentially a cover selection steganography, leading to low embedding capacity.

From aforementioned analysis, in this letter, we proposed a novel steganography scheme tailored for text carrier : cross-modal text steganography (CM-TStega.[1]) which generates stego text semantically consistent with input reference images and is suitable for the cross-modal interaction manner in the social networks. Similar to the proposed CM-TStega, Li et al. [24] generated stego image description using an off-the-shelf language model and a heuristic embedding algorithm based on dynamical synonym substitution, while the method still have low embedding capacity (0.297 bits/sentence). In addition, current generative text steganography methods may suffer from the main drawback that the mapping between secret messages and tokens is constructed heuristically to implement information hiding, which can not resist synonym substitution-based text attack since each token represents a corresponding binary bit. If some stego tokens are replaced by their synonyms, the secret messages would not be accurately extracted from the stego texts. Thus, in order to improve the robustness of the proposed

[1]Code are available at https://github.com/hunanpolly/CMS/CM-TStega

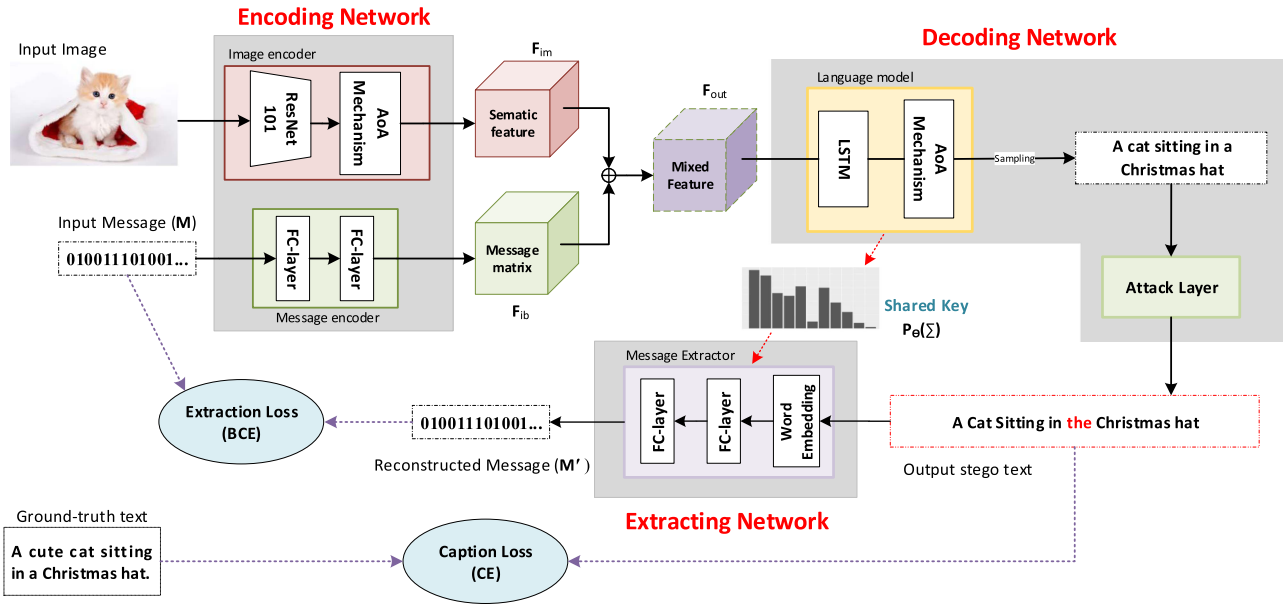

Fig. 1. The overall framework of the proposed cross-modal text steganography scheme. $\oplus$ denotes point-wise add operation. "CE" and "BCE" represent "cross-entropy" and "binary cross-entropy," respectively.

method, an attack layer is designed to simulate the synonym substitution attack based on random sampling in the training phase. Experimentally, we evaluated the proposed CM-TStega in MSCOCO dataset and results illustrate the superior performance of the proposed cross-modal steganography scheme in terms of security and robustness.

II. The Proposed Method

In this section, we introduce the details of the proposed CM-TStega. As shown in Fig. 1, the proposed CM-TStega consists of three parts: encoding network, decoding network, and extracting network. We will introduce these parts in detail in the following subsections.

A. Encoding Network

This component is responsible for translating the inputs to expected matrices. The encoding network contains an image encoder and a message encoder.

Image Encoder: For the purpose of extracting good semantic feature representation of input image, we apply a general ResNet-101 encoder [25] followed by attention on attention (AOA) mechanism [26]. We assume the input image $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$, the image encoder can be formulated as follows:

$$\mathbf{F}_{im} = LayerNorm\left(f_{res}\left(\mathbf{X}\right) + f_{aoa}\left(\mathbf{Q}^E, \mathbf{K}^E, \mathbf{V}^E\right)\right) \quad (1)$$

$$\begin{aligned} &f_{aoa}\left(\mathbf{Q}^E, \mathbf{K}^E, \mathbf{V}^E\right) \\ &= \sigma\left(\mathbf{W}_q^g \mathbf{Q}^E + \mathbf{W}_v^g f_{att}\left(\mathbf{Q}^E, \mathbf{K}^E, \mathbf{V}^E\right) + \mathbf{b}^g\right) \\ &\odot \left(\mathbf{W}_q^i \mathbf{Q}^E + \mathbf{W}_v^i f_{att}\left(\mathbf{Q}^E, \mathbf{K}^E, \mathbf{V}^E\right) + \mathbf{b}^i\right) \end{aligned} \quad (2)$$

where $\mathbf{F}_{im} \in \mathbb{R}^{D\times D}$ is semantic feature space of $\mathbf{X}$. f_{att} is multi-head self attention model [27]. $LayerNorm$ is layer normalization [25]. $\mathbf{Q}^E, \mathbf{K}^E, \mathbf{V}^E = \mathbf{W}_q f_{res}(\mathbf{X}), \mathbf{W}_k f_{res}(\mathbf{X}), \mathbf{W}_v f_{res}(\mathbf{X})$ (f_{res} denotes the ResNet-101 extractor). $\mathbf{W}_q^g$, $\mathbf{W}_v^g \in \mathbb{R}^{D\times D}$, $\mathbf{W}_q^i$, $\mathbf{W}_v^i \in \mathbb{R}^{D\times D}$, $\mathbf{W}_q$, $\mathbf{W}_k$, $\mathbf{W}_v \in \mathbb{R}^{D\times D}$, $\mathbf{b}^g$, and $\mathbf{b}^i \in \mathbb{R}^{D\times 1}$ are learnable parameters.

Massage Encoder: A binary secret message $M \in \{0,1\}^L$ is first fed into a multi-layer perceptron (MLP) with two fully connected layers and a ReLU activation function to match the dimension of semantic feature space ($\mathbf{F}_{im}$). The message encoder is defined as follows:

$$\mathbf{F}_{ib} = f_{relu}\left(\mathbf{W}_1\left(\mathbf{W}_0 M + \mathbf{b}_0\right) + \mathbf{b}_1\right) \quad (3)$$

where $\mathbf{F}_{ib} \in \mathbb{R}^{D\times D}$ is projected latent space of the input secret message. f_{relu} denotes the ReLU activation function. $\mathbf{W}_0 \in \mathbb{R}^{L\times D}$, $\mathbf{W}_1 \in \mathbb{R}^{D\times D}$, $\mathbf{b}_0$, and $\mathbf{b}_1 \in \mathbb{R}^{D\times 1}$ are the learnable parameters of the message encoder. Then the semantic feature space ($\mathbf{F}_{im}$) and the latent space ($\mathbf{F}_{ib}$) are concatenated to produce the final semantic feature space, which is formulated as follows:

$$\mathbf{F}_{out} = \mathbf{F}_{im} \oplus \mathbf{F}_{ib} \quad (4)$$

where $\oplus$ is a concatenation operation. $\mathbf{F}_{out} \in \mathbb{R}^{D\times 2D}$ is final semantic feature space with secret messages.

B. Decoding Network

This part is responsible for producing stego text ($\mathbf{S} = \{s_0, s_1, \ldots, s_n\}$) from the semantic feature space, which contains an LSTM language model and an attack layer based on synonym substitution.

LSTM Language Model: The main goal of the language model is to compute the conditional probabilities of words in the vocabulary, where one LSTM layer and AOA mechanism are leveraged:

$$\begin{aligned} &p\left(\Sigma_t | \mathbf{s}_{1:t-1}, \mathbf{F}_{out}\right) = f_{soft}\left(\mathbf{W}_o \mathbf{c}_t\right) \\ &= f_{soft}\left(\mathbf{W}_o f_{aoa}\left(\mathbf{Q}_t^D, \mathbf{K}^D, \mathbf{V}^D\right)\right) \end{aligned} \quad (5)$$

$$\mathbf{Q}_t^D = \mathbf{W}'_q \mathbf{h}_t;\ \mathbf{K}^D = \mathbf{W}'_k \mathbf{F}_{out};\ \mathbf{V}^D = \mathbf{W}'_v \mathbf{F}_{out}; \quad (6)$$

$$\mathbf{h}_t, \mathbf{r}_t = LSTM\left[(\mathbf{W}_e \mathbf{V}_t, \mathbf{F}_{out} + \mathbf{c}_{t-1}), \mathbf{h}_{t-1}, \mathbf{r}_{t-1}\right] \quad (7)$$

where $\mathbf{W}_e \in \mathbb{R}^{D\times|\Sigma|}$ is a word embedding matrix for vocabulary Σ, and $\mathbf{V}_t$ is one hot encoding of the input word s_t at time step t. $\mathbf{h}_t$ and $\mathbf{r}_t$ are the hidden states of the LSTM. $f_{soft}(\cdot)$ is softmax activation function. $\mathbf{W}_o \in \mathbb{R}^{D\times|\Sigma|}$. $\mathbf{s}_{1:t-1}$ denotes the prefix at first $t-1$ time steps. The language model iteratively samples a word until generating the end token "$< EOS >$" or the maximum length. It is noteworthy that the computed probabilities are used to the key which is transmitted to the extracting network on the secure channel, assisting in secret message extraction.

Attack Layer: Due to crafting mapping between binary secret messages and candidate tokens, existing generative text steganography methods are susceptible to the synonym substitution attack. Inspired by the drawback of generative text steganography, an attack layer based on synonym substitution is designed to improve the robustness of generated stego texts. Specifically, we first exploit the pre-trained Glove vector space [28] and cosine similarity to compute the semantic relevance between words in the vocabulary, and then four candidate words with high relevance are selected to build a thesaurus. Finally, we randomly choose a target word from the sentence generated by the LSTM language model and randomly select the corresponding synonym to replace the target word, producing the final stego texts. The cosine similarity between i-th word and target word is formulated as follows:

$$Sim(\boldsymbol{w}_i, \boldsymbol{w}_t) = \frac{\boldsymbol{w}_i \cdot \boldsymbol{w}_t}{||\boldsymbol{w}_i|| \cdot ||\boldsymbol{w}_t||} \quad (8)$$

where $\boldsymbol{w}.$ denotes the corresponding word embedding vector calculated by the pre-trained Glove vector space.

C. Extracting Network

We design an extractor to extract secret messages from the generated stego texts, which contains a word embedding layer, and an MLP with two fully connected layers. Subsequently, the outputs of the MLP are fed into a sigmoid activation function to yield the predicted binary secret messages:

$$\mathbf{M}' = f_{sig}\left[f_{mlp}\left(\mathbf{S}, '\mathbf{P}_\theta\left(\Sigma\right)\right)\right] \quad (9)$$

where f_{sig} is sigmoid activation function. f_{mlp} denotes an MLP network with a word embedding layer. $\mathbf{P}_\theta(\Sigma)$ is the shared sampling probability distribution from the LSTM language model with parameter θ. It is noteworthy that sharing the probability distribution between Alice and Bob plays an important role in the message extraction since the discrete sampling operation is difficult to retain embedded traces of the semantic feature space.

D. Loss Function and Training Details

The aforementioned networks are jointly trained by the overall loss function formulated as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{cap} + \lambda * \mathcal{L}_{ext} \quad (10)$$

$$\mathcal{L}_{cap} = -\sum_{t=1}^{T} \log\left(p_\theta\left(\mathbf{y}_t^* | \mathbf{y}_{1:t-1}^*\right)\right) \quad (11)$$

$$\mathcal{L}_{ext} = -\sum_{i=1}^{L} m'_i \log(p_\phi(m'_i)) + (1 - m'_i)\log(1 - p_\phi(m'_i)) \quad (12)$$

where $\mathcal{L}_{cap}$ is cross-entropy loss function, and $\mathbf{y}_{1:T}^*$ denotes the target ground truth sequence. $\mathcal{L}_{ext}$ denotes the binary cross-entropy loss function, and $p_\phi(m'_i)$ presents the predicted probability of the i-th extracted binary secret message, which is computed by the message extractor with parameter ϕ.

The network training is divided into two stages: in 0–4 epochs, we just train the encoding and decoding networks, and the λ is set to 1×10^{-8}. From the 5-th epoch, all networks are jointly trained on the total loss function guidance and the λ is set to 10. The maximum length of generated sentence is set to 20 and the batch size is set to 16. Our experiments are conducted on PyTorch and an NVIDIA RTX 3090.

III. Experiments and Discussions

A. Experimental Setup

We evaluate the effectiveness of the proposed CM-TStega scheme on MSCOCO [33], which is a widely used benchmark dataset. The MSCOCO consists of images of complex scenes with people, animals, and common everyday objects in their contexts. It contains 113,287 images for training, 5,000 images for validation, and 5,000 images for testing. Each image in the training set and validation set has been annotated by humans with 5 relatively unbiased sentences. We compare the proposed CM-TStega with three state-of-the-art generative text steganography methods, HC-based [11], AC-based [13], and ADG-based [19] methods. For a fair comparison, the language model of the proposed method is used to implement these compared generative text steganography methods.

B. Comparison With Baselines

We compared the proposed method with SOTA generative text steganography methods in terms of statistical distribution consistency and security performances. The Kullback-Leibler distance (KLD) [19] is used to evaluate statistical distribution consistency and three steganalysis tools (LS-CNN [29], R-BIC [30], SeSy [31] and BERT-FT [32]) are used to measure security performance. In this experiment, the generated texts without secret messages are used as cover texts.

From Table I, the experimental results in the last line demonstrate that the proposed CM-TStega outperforms HC-based and AC-based methods regarding statistical distribution consistency. In addition, with the increase of embedding capacity, the statistical distribution consistency of the proposed CM-TStega gradually decreases since a large embedding capacity arises a big disturbance of the semantic space of the input image. The trend is different from those of AC-based and HC-based methods since the two methods leverage entropy coding to embed secret messages, which is called "Psic-effect" in [34]. For security performance, from Table I, compared with HC-based and AC-based methods, the proposed CM-TStega can significantly achieve superior security performance against several advanced text steganalysis tools based on deep learning. The steganalysis accuracy of the proposed CM-TStega method is an average of 33.38% 29.45% lower than those of HC-based and AC-based methods, respectively. Meanwhile, the proposed

TABLE I
THE EXPERIMENTAL RESULTS OF STATISTIC DISTRIBUTION CONSISTENCY AND STEGANALYSIS ACCURACY COMPARED WITH THREE SOTA TEXT STEGANOGRAPHY METHODS

Model →	ADG-based [19]	HC-based [11]			AC-based [13]			CM-TStega (Ours)		
bpw →	4.77	1.00	1.79	2.94	1.33	2.14	3.05	1.03	2.11	3.08
LS-CNN [29]	0.5680	0.9475	0.9170	0.8550	0.9240	0.8675	0.8100	0.5320	0.5745	0.6115
R-BI-C [30]	0.5710	0.9425	0.9215	0.8595	0.9210	0.8625	0.8140	0.5355	0.5770	0.6140
SeSy [31]	0.5530	0.9335	0.9010	0.8445	0.9125	0.8555	0.7820	0.5245	0.5670	0.6010
BERT-FT [32]	0.5840	0.9580	0.9395	0.8690	0.9370	0.8815	0.8255	0.5425	0.5870	0.6215
KLD	0.98	7.45	5.34	4.11	6.54	5.32	3.65	0.71	0.96	1.05

TABLE II
THE QUANTITATIVE ANALYSIS OF THE CM-TSTEGA

bpw→	1.03	2.11	3.08	4.15
BLEU-4	29.30	27.40	27.60	23.70
METERO	26.80	24.20	22.50	21.30
ROUGE-L	54.70	52.10	48.10	47.10
Ext-ACC	1.00	1.00	0.9323	0.8705

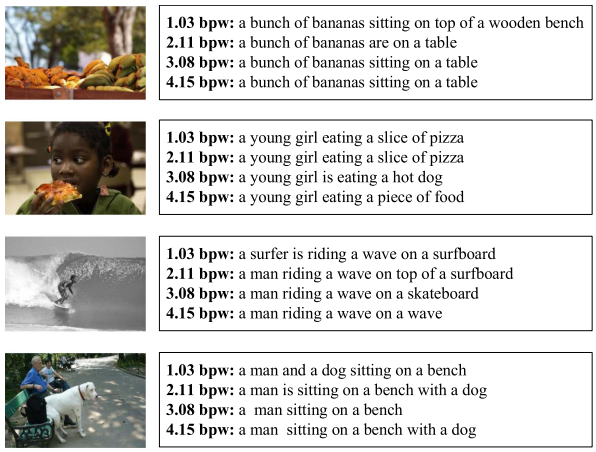

Fig. 2. The examples of generated stego texts.

CM-TStega method can achieve comparable results to those of the ADG-based method.

C. Performance of CM-TStega

Quantitative Analysis: We use BLEU [35], METEOR [36], ROUGE-L [37] and extracting accuracy(Ext-ACC) to quantitatively evaluate the performance of the proposed scheme. The experimental results are shown in Table II. From Table II, the embedding capacity, i.e., bit per word (bpw), is an average number of binary secret messages embedded into each word of the generated stego texts. Due to the different lengths of stego sentences, the embedding capacity is not an integer number but a float number. With the increase of embedding capacity, the text quality of stego texts gradually decreased and the extraction error rate increased. It is noteworthy that when the embedding capacity is less than 3 bpw, the extraction accuracy can be over 90%.

Qualitative Analysis: Fig. 2 compares the generated stego texts of the proposed scheme at different embedding capacities, indicating that the secret messages are successfully concealed in the stego texts. It is noteworthy that the proposed method maybe generates the same stego texts under different embedding capacities because the embedding disturbance probably just changes the sampling probabilities of the marginal words for the semantic feature of the input reference image.

TABLE III
THE ROBUSTNESS ANALYSIS OF THE CM-TSTEGA

bpw →	1.03	2.11	3.08	4.15
No attack	1.00	1.00	0.9323	0.8705
1 word	1.00	0.9977	0.9284	0.8665
2 words	1.00	0.9885	0.9221	0.8600
3 words	0.9861	0.9745	0.9115	0.8447
4 words	0.9644	0.9417	0.8874	0.8322

"*n*-words" denotes the number of substituted words.

Robustness Analysis: In this part, we evaluate the robustness performance of the proposed method against text attack based on synonym substitution. To the best of our knowledge, we are first to consider the attack for generated stego texts. From Table III, the experimental results show the extraction accuracy gradually decreases as the number of substituted words increases. Note that the state-of-the-art generative text steganography schemes construct a mapping between secret messages and candidate tokens, which can not resist the text attack based on synonym substitution.

IV. CONCLUSION

In this letter, we proposed a novel text steganography scheme called CM-TStega tailored for cross-modal interaction on popular social networks. The CM-TStega is jointly trained with an encoding network, decoding network, and extracting network. Meanwhile, we first design a text attack layer based on synonym substitution to improve the robustness of generated stego texts. Experimental results show that the proposed CM-TStega can achieve superior security and robustness performance than current text steganography. Moreover, the proposed CM-TStega generates fluent stego texts with good semantic consistency for an input reference image.

ACKNOWLEDGMENT

The authors would like to sincerely thank the editors and reviewers for their valuable comments and suggestions.

REFERENCES

[1] Y. Chen et al., "Cross-modal ambiguity learning for multimodal fake news detection," in *Proc. ACM Web Conf.*, 2022, pp. 2897–2905.

[2] Z.-Y. Dou et al., "An empirical study of training end-to-end vision-and-language transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 18166–18176.

[3] Y.-L. Sung, J. Cho, and M. Bansal, "VL-Adapter: Parameter-efficient transfer learning for vision-and-language tasks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5227–5237.

[4] C. Ju, T. Han, K. Zheng, Y. Zhang, and W. Xie, "Prompting visual-language models for efficient video understanding," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 105–124.

[5] J. Fridrich, *Steganography in Digital Media: Principles, Algorithms, and Applications*. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[6] V. Holub and J. Fridrich, "Digital image steganography using universal distortion," in *Proc. first ACM workshop Inf. Hiding Multimedia Secur.*, 2013, pp. 59–68.

[7] T. Filler, J. Judas, and J. Fridrich, "Minimizing additive distortion in steganography using syndrome-trellis codes," *IEEE Trans. Inf. Forensics Secur.*, vol. 6, no. 3, pp. 920–935, Sep. 2011.

[8] B. Li, J. He, J. Huang, and Y. Q. Shi, "A survey on image steganography and steganalysis," *J. Inf. Hiding Multim. Signal Process.*, vol. 2, no. 2, pp. 142–172, 2011.

[9] Z. Zhou et al., "Secret-to-image reversible transformation for generative steganography," *IEEE Trans. Dependable Secure Comput.*, early access, Oct. 27, 2022, doi: 10.1109/TDSC.2022.3217661.

[10] T. Fang, M. Jaggi, and K. Argyraki, "Generating steganographic text with lstms," in *Proc. ACL, Student Res. Workshop*, 2017, pp. 100–106.

[11] Z.-L. Yang, X.-Q. Guo, Z.-M. Chen, Y.-F. Huang, and Y.-J. Zhang, "Rnn-stega: Linguistic steganography based on recurrent neural networks," *IEEE Trans. Inf. Forensics Secur.*, vol. 14, no. 5, pp. 1280–1295, May 2019.

[12] F. Z. Dai and Z. Cai, "Towards near-imperceptible steganographic text," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 4303–4308 .

[13] Z. M. Ziegler, Y. Deng, and A. M. Rush, "Neural linguistic steganography," in *Proc. 2019 Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process.*, 2019, pp. 1210–1215.

[14] Y. Tew and K. Wong, "An overview of information hiding in h. 264/avc compressed video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 2, pp. 305–319, Feb. 2014.

[15] P. Jayaram, H. Ranganatha, and H. Anupama, "Information hiding using audio steganography–a survey," *Int. J. Multimedia Its Appl.*, vol. 3, pp. 86–96, 2011.

[16] J. Wu, B. Chen, W. Luo, and Y. Fang, "Audio steganography based on iterative adversarial attacks against convolutional neural networks," *IEEE Trans. Inform. Forensics Secur.*, vol. 15, pp. 2282–2294, 2020.

[17] P. Wei, S. Li, X. Zhang, G. Luo, Z. Qian, and Q. Zhou, "Generative steganography network," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 1621–1629.

[18] Z. You, Q. Ying, S. Li, Z. Qian, and X. Zhang, "Image generation network for covert transmission in online social network," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 2834–2842.

[19] S. Zhang, Z. Yang, J. Yang, and Y. Huang, "Provably secure generative linguistic steganography," in *Proc. Findings Assoc. Comput. Linguistics: ACL-IJCNLP*, 2021, pp. 3046–3055.

[20] X. Zhou, W. Peng, B. Yang, J. Wen, Y. Xue, and P. Zhong, "Linguistic steganography based on adaptive probability distribution," *IEEE Trans. Dependable Secure. Comput.*, vol. 19, no. 5, pp. 2982–2997, Sept./Oct. 2022.

[21] Y. Xue, J. Zhou, H. Zeng, P. Zhong, and J. Wen, "An adaptive steganographic scheme for h. 264/AVC video with distortion optimization," *Signal Processing: Image Commun.*, vol. 76, pp. 22–30, 2019.

[22] Y. Hu, Z. Yang, H. Cao, and Y. Huang, "Multi-modal steganography based on semantic relevancy," in *Proc. Digit. Forensics Watermarking: 19th Int. Workshop*, 2021, pp. 3–14.

[23] Y. Hu, H. Cao, Z. Yang, and Y. Huang, "Improving text-image matching with adversarial learning and circle loss for multi-modal steganography," in *Proc. Digit. Forensics Watermarking: 19th Int. Workshop*, 2021, pp. 41–52.

[24] M. Li, K. Mu, P. Zhong, J. Wen, and Y. Xue, "Generating steganographic image description by dynamic synonym substitution," *Signal Process.*, vol. 164, pp. 193–201, 2019.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[26] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, "Attention on attention for image captioning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 4634–4643.

[27] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inform. Process. Syst.*, 2017, pp. 6000–6010.

[28] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. 2014 Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1532–1543.

[29] J. Wen, X. Zhou, P. Zhong, and Y. Xue, "Convolutional neural network based text steganalysis," *IEEE Signal Process. Lett.*, vol. 26, no. 3, pp. 460–464, Mar. 2019.

[30] Y. Niu, J. Wen, P. Zhong, and Y. Xue, "A hybrid R-BILSTM-C neural network based text steganalysis," *IEEE Signal Process. Lett.*, vol. 26, no. 12, pp. 1907–1911, Dec. 2019.

[31] J. Yang, Z. Yang, S. Zhang, H. Tu, and Y. Huang, "SeSy: Linguistic steganalysis framework integrating semantic and syntactic features," *IEEE Signal Process. Lett.*, vol. 29, pp. 31–35, 2021.

[32] W. Peng, J. Zhang, Y. Xue, and Z. Yang, "Real-time text steganalysis based on multi-stage transfer learning," *IEEE Signal Process. Lett.*, vol. 28, pp. 1510–1514, 2021.

[33] T.-Y. Lin et al., "Microsoft coco: Common objects in context," in *Proc. Euro. Conf. Comput. Vis.*, 2014, pp. 740–755.

[34] Z.-L. Yang, S.-Y. Zhang, Y.-T. Hu, Z.-W. Hu, and Y.-F. Huang, "Vae-stega: Linguistic steganography based on variational auto-encoder," *IEEE Trans. Inf. Forensics Secur.*, vol. 16, pp. 880–895, 2021.

[35] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, 2002, pp. 311–318.

[36] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proc. acl Workshop Intrinsic Extrinsic Eval. Measures Mach. Transl. and/or Summarization*, 2005, pp. 65–72.

[37] L. C. Rouge, "A package for automatic evaluation of summaries," in *Proc. Workshop Text Summarization ACL*, 2004, pp. 23–31.