CAUNet: Cross-Attention UNet for Image Steganography

Bingxin Wei, Haewoon Nam
Department of Electrical Engineering, Hanyang University Ansan, South Korea
wei0911@hanyang.ac.kr, hnam@hanyang.ac.kr

***Abstract*—Image steganography is an important branch in the field of information hiding, aiming to covertly embed secret information into the cover image. Existing UNet-based steganographic methods demonstrate strong feature extraction capability, but the simple skip connection performed between the encoder and decoder often leads to low feature fusion efficiency and easily introduces redundant information. To solve this problem, this paper proposes a new network structure called CAUNet (Cross-Attention UNet) for image steganography. The core contribution lies in designing a lightweight cross-attention module and embedding it into the skip connection of UNet. This module dynamically learns the interdependence between encoder and decoder features to generate an attention map for weighting and refining the encoder features, thereby achieving more effective and targeted feature fusion. Experimental results show that, compared with the conventional UNet-based steganography model, CAUNet significantly improves the visual quality of the reconstructed stego images, demonstrating that the proposed cross-attention mechanism effectively enhances the performance of steganographic networks.**

***Index Terms*—Image steganography, Cross-Attention, Deep learning**

I. INTRODUCTION

IMAGE steganography aims to covertly embed secret information in the cover image while maintaining the visual quality of the image, to achieve secure communication. The core challenge lies in balancing two key performance indicators: imperceptibility, which minimizes the visual difference between the stego image and the original cover image, and robustness, which means embedding the information while maintaining the ability to resist [1]. In recent years, the rapid development of deep learning technology has provided powerful tools to address these challenges, significantly pushing the performance boundaries of steganography algorithms.

Among the numerous encoder-decoder structures based on deep learning [2], [3] [4], especially the UNet structure, because of its powerful feature extraction and pixel-level reconstruction capabilities, it is widely applied in image steganography tasks. UNet [5] extracts multi-scale features through a contraction path (encoder) and restores spatial resolution through an expansion path (decoder). To preserve high-resolution details, UNet relies on skip connections to directly transfer advanced features from the encoder to the corresponding layers of the decoder for feature concatenation. However, this simple concatenation mechanism has inherent limitations. Passes all features, regardless of their relevance to the steganography task, equally weighted, which easily introduces high-dimensional redundant information and background noise from the encoder, thereby interfering with the decoder's accurate reconstruction of the secret information and ultimately limiting the performance improvement of the steganography network.

To address the limited efficiency of the feature fusion of the skip connections in UNet, this paper proposes a novel image steganography framework termed CAUNet (Cross-Attention UNet). The core contribution of this work is the design and integration of a lightweight cross-attention module that replaces the traditional simple concatenation. Positioned along the feature transmission path between the encoder and the decoder, this module dynamically models the contextual dependencies between their respective feature representations and generates a refined attention map. This attention map is used to weight and calibrate the encoder features, effectively suppressing redundant background features and highlighting the local details that are crucial for embedding secret information. Through this attention-guided feature fusion mechanism, CAUNet can achieve more robust and higher-quality feature representations, thereby significantly improving the quality of stego images and the capacity of information embedding. The main contributions of this paper are as follows.

- This paper proposes a new steganographic network called CAUNet, which is developed based on the UNet architecture. Through structural innovation, the network effectively enhances the embedding performance.
- We design an cross-attention module to enhance the feature fusion between the encoder and decoder of UNet, effectively suppressing redundant information.
- Extensive experimental results show that CAUNet outperforms existing image steganography methods in both visual quality and anti-steganalysis capability.

The rest of the paper is organized as follows. Section II introduces the basic idea and the related techniques. Section III describes the steganography algorithms. Experimental results are given in Section IV. We conclude this paper in Section V.

II. RELATED WORK

Image steganography, as an important research direction in the field of information hiding, has made significant progress driven by deep learning technology in recent years. Traditional steganography methods mostly rely on manually designed embedding rules and statistical models [6], [7], which are difficult to achieve simultaneously concealment and robustness

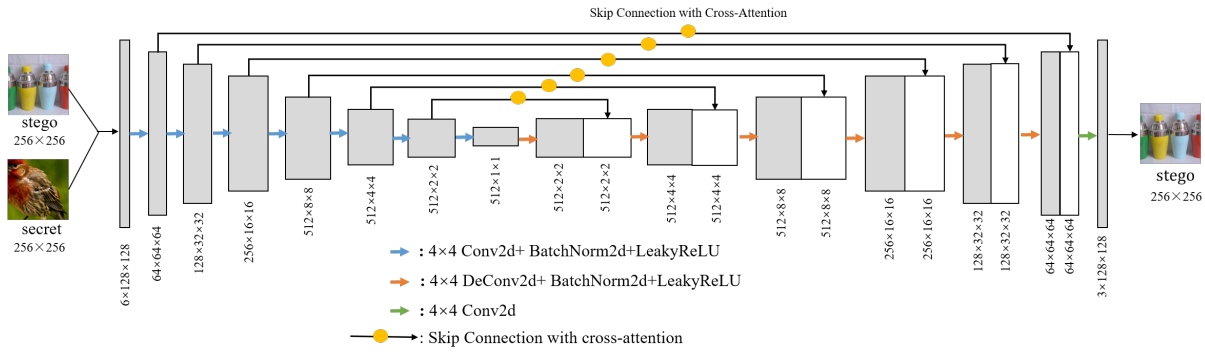

Fig. 1. The CAUNet structure, as the hidden network for image steganography, incorporates a cross-attention module in its skip connections.

in complex scenarios. With the development of convolutional neural networks, deep learning-based steganography methods [2], [8] have gradually become mainstream. Among them, UNet due to its powerful ability to extract multi-scale features is widely used in steganography encoding and decoding tasks. However, skip connections in standard UNet usually adopt a simple feature concatenation method, which directly integrates and may lead to the transmission of redundant information, thereby affecting the embedding effect and the quality of the stego image.

To enhance the efficiency of feature interaction between the encoder and decoder, recent studies have attempted to introduce the attention mechanism to improve the feature fusion process [9]. The attention mechanism can dynamically adjust the weights of features according to task requirements, allowing the network to focus more on key information and reducing unnecessary background interference. Such methods have demonstrated excellent performance in tasks such as image segmentation, super-resolution, and image restoration. Based on the current research status, this paper proposes a new steganographic network CAUNet, and introduces a lightweight cross-attention module to enhance the feature fusion ability between the encoder and decoder. This module can effectively suppress redundant background features and highlight local details conducive to information embedding, thereby improving the quality and embedding performance of the stego image. Further experimental results have verified the superiority of CAUNet on multiple benchmark datasets.

III. PROPOSED METHOD

A. Overall Architecture

The CAUNet adopts a typical encoder-decoder framework. The encoder consists of six downsampling convolutional modules, with the number of channels increasing layer by layer, for extracting multi-scale semantic features. The decoder is composed of six upsampling modules, corresponding one-to-one with the encoder. Unlike the traditional UNet which uses a simple concatenation method, this network introduces a lightweight cross-attention module in each level of skip connection to enhance the quality of feature fusion. The structure of the CAUNet structure is shown in Fig. 1.

The goal of the extraction network is to accurately recover the secret information from the stego image. The extraction network is composed of a deep network with 6 convolutional layers. The number of channels gradually increases ($64 \rightarrow 256$), then decreases ($256 \rightarrow 64$), and finally, through Conv6, the output channel number is mapped to 3. This series of standard convolution operations can learn the subtle changes embedded in stego and map the features back to the secret information.

B. Cross-Attention Module

The objective of this module is to utilize the decoder features as guidance to selectively aggregate the information within the encoder features. Its input consists of the encoder features $F_{enc} \in \mathbb{R}^{C_{enc} \times H \times W}$ and the features from the previous level decoder $F_{dec} \in \mathbb{R}^{C_{dec} \times H \times W}$. Apply a 1×1 convolution to both features with weights W_e and W_d, and unify their channel numbers to C' to facilitate effective fusion. The process is as follows.

$$F'_{enc} = W_e(F_{enc}), \tag{1}$$

$$F'_{dec} = W_d(F_{dec}), \tag{2}$$

where F_{enc} and F_{dec} represents the encoder/decoder feature, $W(\cdot)$ is the convolution operation, F'_{enc} and F'_{dec} represents the transformed encoder/decoder feature.

Subsequently, the transformed features F'_{enc} and F'_{dec} are concatenated to obtain the fused feature F_{fused}. Then, a convolution layer, batch normalization (BN), ReLU activation function, and 1×1 convolution are used to generate the attention map and the generated as follows.

$$F_{fused} = \text{Concat}(F'_{enc}, F'_{dec}), \tag{3}$$

$$A = \sigma(\text{Conv}_{1\times1}(\text{ReLU}(\text{BN}(\text{Conv}_{3\times3}(F_{fused}))))), \tag{4}$$

where σ is the Sigmoid activation function, which is used to normalize the attention weights A to the range of $[0, 1]$.

Finally, the attention map A is element-wise multiplied ($\otimes$) with the encoder features F_{enc} to obtain the refined features F_{attn}. The refined feature F_{attn} will replace the original F_{enc} and be involved in the subsequent operations of the decoder (i.e., be concatenated with the upscaled decoder features). The overall cross-attention module is shown in Fig 2.

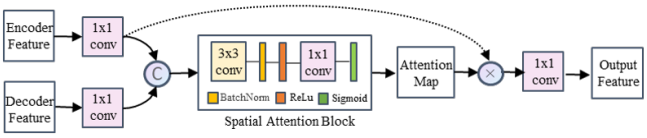

Fig. 2. A lightweight cross-attention structure applicable to image steganography.

C. Loss Function

In order to simultaneously ensure the visual quality of the steganographic image and the reliable recovery of the secret information, this paper divides the training objective into two parts: the hiding loss and the extraction loss.

1) Hiding loss: Measure the similarity between the stego-image I_{stego} and the original cover image I_{cover} to ensure imperceptibility. The loss function of the hiding network is designed as

$$L_{Hid} = \frac{1}{N}\sum_{i=1}^{N}\left(I_{cover} - I_{stego}\right)^2, \tag{5}$$

where N is the number of samples used for training, I_{cover} is the real image and I_{stego} is the image reconstructed by the network.

2) Extraction loss: Measure the difference between the recovered secret image I_{revsec} and the original secret image I_{secret} to ensure the accurate extraction of the information.

$$L_{Ext} = \frac{1}{N}\sum_{i=1}^{N}\left(I_{secret} - I_{resec}\right)^2, \tag{6}$$

where N is the number of samples used for training, I_{secret} is the real image and I_{resec} is the image reconstructed by the network.

3) Total loss: The total loss function consists of hiding loss and extraction loss.

$$L_{Total} = L_{Hid} + \lambda_e L_{Ext}, \tag{7}$$

where λ_e are the hyperparameters to balance the weight between the imperceptibility and the accuracy of secret information recovery.

IV. Experiment

A. Experimental Setup

In this paper, the training data was selected from the small ImageNet dataset. Training data is drawn from the ImageNet dataset, where 40,000 images are randomly selected and uniformly resized to a resolution of 256 × 256. All experiments are carried out on an NVIDIA GeForce RTX 3090 GPU, using PyTorch 1.10.0 and Python 3.6.13 as the implementation environment. During training, the batch size is set to 8 and the initial learning rate is configured to 0.001. The network is trained for a total of 200 epochs to ensure stable convergence.

TABLE 1 Benchmark comparisons on different methods.

Method	PSNR		SSIM	
	Cover-stego	Secret-RevSec	Cover-stego	Secret-RevSec
HiDDeN [2]	35.52	36.10	0.9691	0.9696
Baluja [8]	37.25	36.34	0.9687	0.9516
UNet (Baseline)	39.07	39.12	0.9775	0.9658
CAUNet (Proposed)	**41.80**	**40.35**	**0.9920**	**0.9817**

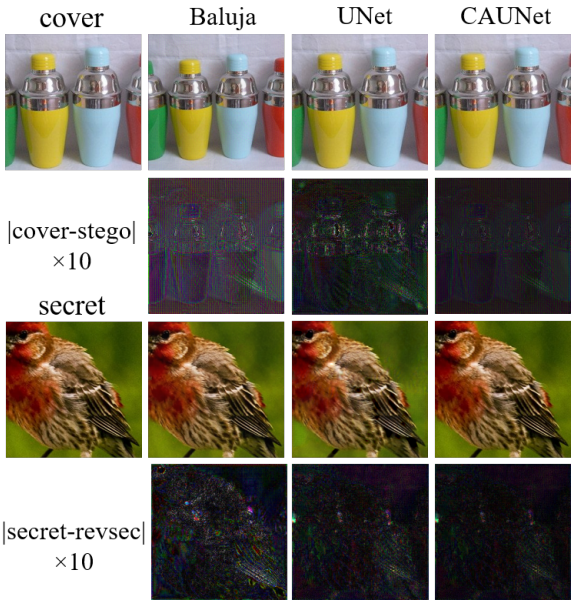

Fig. 3. Qualitative results of the different methods.

B. Results and Analysis

1) Quantitative Results: We conducted a quantitative comparison of the proposed CAUNet with three mainstream steganography methods on the standard test set. The results are summarized in Table 1. The evaluation metrics include peak signal-to-noise ratio and structural similarity. In terms of imperceptibility, CAUNet demonstrated the best performance, with PSNR (I_{stego}/ I_{cover}) reaching 41.80 dB, significantly higher than Baseline UNet (39.07 dB) by 2.73 dB. SSIM (I_{stego}/ I_{cover}) reached 0.9920, indicating that the stego images generated by it have almost no structural difference from the original cover image, and the visual quality is at an extremely high level. These results fully prove that the cross-attention mechanism we introduced effectively refined the encoder features, avoided the introduction of redundant features and noise, and enabled the hiding network to generate the highest fidelity stego images. In terms of the extraction of the core secret information, it also showed advantages, with PSNR/SSIM reaching 40.35 dB/0.9817. Overall, the robust feature fusion achieved through the cross-attention mechanism can better utilize the high-resolution information of the encoder to guide the decoder to accurately perform information embedding, significantly improving the visual quality of the stego images.

2) Qualitative Results: To visually demonstrate the superiority of the proposed CAUNet, we conducted a qualitative analysis of the visual quality of the steganographic images and the clarity of recovering the secret information. Fig 3

shows the randomly selected cover images, the secret images generated the corresponding results. By visually comparing the original cover image I_{cover} with the stego image I_{stego} generated by CAUNet, it can be observed that there is an almost imperceptible visual difference between I_{stego} and I_{cover}. Compared with the Baseline UNet model, the stego images generated by CAUNet can retain the original structure and color of the image to the greatest extent, which is consistent with the high PSNR and SSIM performance shown in the quantitative results.

V. Conclusion

This paper proposes an innovative steganographic network structure, CAUNet (Cross-Attention UNet), by embedding a cross-attention module in the skip connection between the encoder and decoder of the UNet. This solves the problem of redundant information interference caused by the simple concatenation of traditional features. Experimental results have strongly demonstrated the superiority of CAUNet. In quantitative analysis, the model not only ensures the quality of the reconstructed image (cover-stego PSNR reaches 41.80 dB) but also improves the extraction quality of the secret image (secret-revsec PSNR reaches 40.35 dB). Future work will focus on improving the adaptability capacity and anti-channel attack ability of the network.

VI. ACKNOWLEDGEMENT

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (2022R1A2C1011862) and China Scholarship Council(NO.202308260032).

References

[1] R. J. Anderson and F. A. Petitcolas, "On the limits of steganography," vol. 16, no. 4, pp. 474–481, 1998.

[2] J. Zhu, R. Kaplan, J. Johnson, and L. Fei-Fei, "Hidden: Hiding data with deep networks," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 657–672.

[3] X. Zhu, Z. Lai, Y. Liang, J. Xiong, and J. Wu, "Generative high-capacity image hiding based on residual cnn in wavelet domain," *Applied Soft Computing*, vol. 115, p. 108170, 2022.

[4] K. A. Zhang, A. Cuesta-Infante, L. Xu, and K. Veeramachaneni, "Steganogan: High capacity image steganography with gans," *arXiv preprint arXiv:1901.03892*, 2019.

[5] X. Weng, Y. Li, L. Chi, and Y. Mu, "High-capacity convolutional video steganography with temporal residual modeling," in *Proc. Int. Conf. Multimedia Retrieval*, 2019, pp. 87–95.

[6] J. Fridrich and T. Filler, "Practical methods for minimizing embedding impact in steganography," in *Security, Steganography, and Watermarking of Multimedia Contents IX*, vol. 6505. SPIE, 2007, pp. 13–27.

[7] J. Fridrich, M. Goljan, P. Lisonek, and D. Soukal, "Writing on wet paper," vol. 53, no. 10, pp. 3923–3935, 2005.

[8] S. Baluja, "Hiding images in plain sight: Deep steganography," *Advances in neural information processing systems*, vol. 30, 2017.

[9] F. Li, Y. Sheng, X. Zhang, and C. Qin, "iscmis: Spatial-channel attention based deep invertible network for multi-image steganography," *IEEE Transactions on Multimedia*, vol. 26, pp. 3137–3152, 2023.