Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

EGDIS: Edge-Guided Document Image Steganography

Shuqin Zhang
Zhongyuan University of Technology
Kaiyu Li
likaiyu@zut.edu.cn
Zhongyuan University of Technology
Xueqing Wang
Zhongyuan University of Technology
Jun Li
Zhongyuan University of Technology
Yuanqing Xia
Zhongyuan University of Technology

Research Article

Keywords: Document Image Steganography, Edge-Guided, Cross-Attention Fusion, Transformer

Posted Date: August 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-7990497/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

EGDIS: Edge-Guided Document Image Steganography

Shuqin Zhang[1], Kaiyu Li[1*], Xueqing Wang[1], Jun Li[1], Yuanqing Xia[1]

[1*]School of Computer Science, Zhongyuan University of Technology, Zhengzhou, 451191, China.

*Corresponding author(s). E-mail(s): likaiyu@zut.edu.cn;
Contributing authors: zhangsq@zut.edu.cn; 1461600041@qq.com; lijun@zut.edu.cn; xyq@zut.edu.cn;

Abstract

Document image steganography is a key information-hiding task aimed at securely and effectively embedding secret information into document images while preserving visual quality. State-of-the-art document image steganography methods typically focus on feature fusion. However, they often overlook edge information when designing fusion strategies. In this work, we observe that the incorporation of edge features leads to improved fusion performance. Based on this observation, we propose a novel method: Edge-Guided Document Image Steganography (EGDIS). The proposed model first applies an adaptive Canny algorithm to extract high-quality edge maps from document images. Subsequently, a Transformer-based Cross-Attention Fusion Network (CAFN) is employed to perform multi-scale feature fusion among the raw image, edge image, and secret message. By leveraging cross-attention mechanisms, the model enables fine-grained integration across modalities and scales. This mechanism captures long-range dependencies in image content, facilitating the deep integration of secret messages with raw and edge images, while preserving edge distribution patterns and global semantic consistency. Extensive experiments demonstrate that EGDIS outperforms existing methods on both DocImgCN and DocImgEN datasets.

Keywords: Document Image Steganography, Edge-Guided, Cross-Attention Fusion, Transformer

1 Introduction

In modern society, the growing demand for secure information storage and transmission has driven the rapid development of various steganographic techniques. To enhance information security, researchers have proposed image-based steganography methods [1–5], which embed secret data into image carriers. These approaches have demonstrated impressive performance on natural image steganography tasks. However, their effectiveness decreases substantially when applied to document images. Unlike natural images, document images generally exhibit uniform backgrounds and limited texture noise, making data embedding significantly more challenging.

To overcome these limitations, Ge et al. [6] proposed an end-to-end document image watermarking scheme based on deep neural networks. This approach incorporates a text-sensitive loss function to effectively constrain modifications within character regions. However, this method still exhibits limited precision in information extraction. Li et al. [7] proposed the DAMS method, which explicitly accounts for the influence of pixel distribution and value deviations during the generation of document images. They introduced a

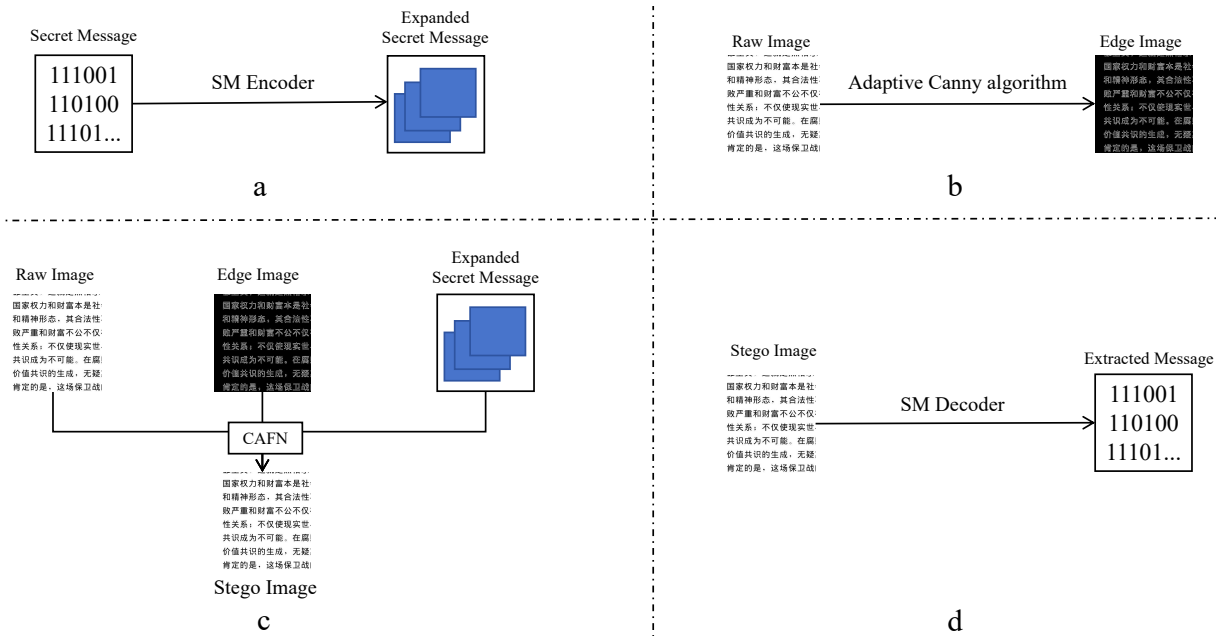


Fig. 1 Edge-Guided Document Image Steganography: a) Message Preprocessing: The secret message is resized to match the dimensions of both the raw and edge images using a SM Encoder. b) Edge Extraction: An enhanced Canny algorithm is employed to generate the edge image from the raw document image. c) Cross-Modality Attention Fusion: The CAFN dynamically assigns weights to effectively integrate the raw image, edge image, and secret message. d) Message Extraction: The SM Decoder extracts and reconstructs the secret message from the stego image.

novel Self-Attention Fusion (SAF) network to enable multi-scale feature extraction and fusion in text regions. The self-attention mechanism significantly enhances the perception of text region features, thereby improving both the effectiveness of secret information embedding and the accuracy of information extraction. However, these methods do not fully exploit textual edge features and struggle to maintain the consistency of semantic edges, indicating room for further improvement in the imperceptibility of steganographic images.

Motivated by the above issues, this paper proposes an edge-guided document image steganography method that enhances feature fusion by incorporating edge information. This approach utilizes edge features to accurately embed secret information into the original image. Specifically, we introduce a Spatial Message Encoder, which transforms binary secret data into a format compatible with the spatial dimensions of the original image. Furthermore, to mitigate local semantic distortions caused by traditional embedding methods, we propose a Transformer-based Cross-Attention Fusion Network (CAFN). This network dynamically integrates edge images, raw images, and secret information through a cross-attention mechanism, enabling adaptive fusion among the three components. As a result, the generated stego image adheres to edge distribution constraints while maintaining global visual fidelity. To ensure accurate message recovery, we also designed a decoder that employs a parameter-sharing mechanism. Through the joint design of the encoder, CAFN, and decoder, experimental results demonstrate that our method achieves promising performance in document image steganography.

The main contributions of this work are summarized as follows:

- We propose an innovative EGDIS that utilizes edge information from document images to enhance imperceptibility.
- We design the CAFN that effectively utilizes edge information for message embedding, enabling accurate message recovery while maintaining high imperceptibility in the resulting stego images.
- Extensive experiments on publicly available benchmark datasets demonstrate that EGDIS achieves state-of-the-art performance.

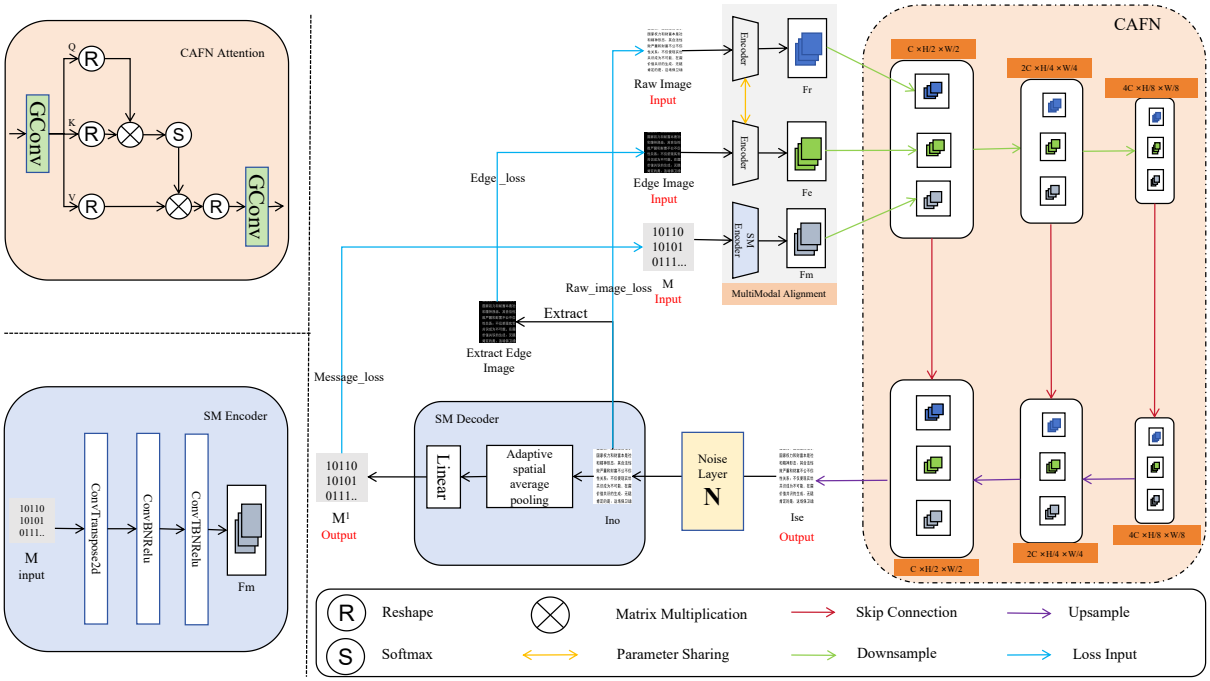


Fig. 2 The EGDIS framework comprises four core components: the SM Encoder, the CAFN, the noise layer, and the SM Decoder. The SM Encoder preprocesses the secret message for embedding; the CAFN dynamically fuses the raw image, edge image, and secret message; the noise layer simulates real-world interference to enhance robustness; and the SM Decoder reconstructs the complete secret message from the stego image.

2 Related Work

2.1 Image Steganography

In recent years, the rapid advancement of deep learning has significantly propelled the development of image steganography techniques, particularly in handling complex image backgrounds, where numerous deep learning-based approaches have achieved remarkable results [1, 4–13]. Researchers have proposed end-to-end image steganography methods aimed at enhancing model robustness. For example, HiDDeN [1], the first neural network-based end-to-end image encoding and decoding architecture, has achieved notable success. However, it still faces challenges such as limited imperceptibility of hidden messages and image quality degradation when applied to document image backgrounds.

To address these issues, Ge et al. [6] introduced a document-specific steganographic method that improves recovery accuracy but still suffers from limitations in imperceptibility. Building upon this, Li et al. [7] proposed the Dual-Attention Multi-Scale (DAMS) Encoder-Decoder Architecture, which balances the embedding and extraction of information through a Channel Attention Network (CAN), while integrating multi-scale textual region features using a Self-Attention Fusion Network (SAF). Although DAMS significantly improves recovery accuracy, its enhancement of imperceptibility metrics remains constrained due to limited consideration of edge semantic information and regional feature consistency.

2.2 Feature Fusion Techniques

Effective multi-scale feature fusion plays a vital role in various vision tasks by integrating complementary features across different levels or representations, thereby enhancing the model's expressive capacity and generalization performance. Early studies have explored this direction from different perspectives. For example, Li et al. [14] introduced a Multi-scale Edge-Guided Learning Network (MEGLN), which utilizes global edge

information to guide the network in more effectively capturing and restoring local details. Similarly, Wei et al. [15] proposed a novel Multi-Modality Cross Attention (MMCA) Network that jointly models intra- and inter-modality relationships between image regions and sentence words through a cross-attention mechanism.

Recent approaches have advanced the integration of multi-scale and attention mechanisms. Wen et al. [16] presented a multi-scale context feature and cross-attention method based on convolutional neural networks, wherein multi-scale context modules are constructed at different network depths, and both cross-attention modules and deep supervision are applied in each branch, followed by output integration. Likewise, Xu et al. [17] introduced a cross-attention fusion network (CAF-Former) that bridges CNN and Transformer architectures to leverage the local feature extraction strengths of CNNs and the long-range dependency modeling capabilities of Transformers for hyperspectral image classification. Inspired by these advancements, our work proposes a dynamic multi-scale feature fusion strategy combined with a cross-attention mechanism, enabling dynamically weighted aggregation to effectively improve the invisibility of stego images.

3 Method

3.1 Overall architecture

As shown in Figure 2, the edge-guided document image steganography framework primarily comprises the CAFN, the SM Encoder, the noise layer, and the SM Decoder. The workflow is divided into four stages: First, the SM Encoder encodes the secret message M into a feature map F_m. Second, the CAFN dynamically fuses the raw feature map F_r, the edge feature map F_e, and the message feature map F_m using a cross-attention mechanism to generate the stego image I_{se}. Third, the noise layer adds simulated noise to the stego image I_{se}, producing a noisy image I_{no} to enhance robustness. Finally, the SM Decoder reconstructs the recovered secret message M' from the noisy image I_{no}, completing the end-to-end steganography and extraction process.

3.2 Spatial Message Encoder

Before initiating our research on message processing, we conducted a systematic review of the relevant literature. Among the reviewed methods, the DAMS approach proposed by Li et al. [7] served as a key inspiration for this work. Building on its message processing mechanism, we designed a processing pipeline comprising the following core components: the encoder takes as input a binary message $M \in \mathbb{R}^{\text{batch}\times L}$ of length L. The message M is then transformed into a message feature map $F_m \in \mathbb{R}^{\text{batch}\times L\times H\times W}$ through dimensional expansion and upsampling, ensuring spatial alignment between the two-dimensional message representation and the raw image. This process is formally defined by the following equations:

$$F_m = Upsample(Reshape(\left\lfloor \sqrt{L} \right\rfloor, M)) \quad (1)$$

3.3 Cross-Attention Fusion Network

This study aims to enhance the concealment of M within textural edges and improve the efficiency of feature fusion, while accurately preserving critical multimodal information. To this end, we propose the CAFN, whose core workflow is as follows: First, feature alignment is achieved through dedicated encoders that extract multimodal features from the raw image, edge image, and secret message. These features are then projected into query ($\mathbf{Q}$), key ($\mathbf{K}$), and value ($\mathbf{V}$) triplets using a feature mapping strategy. Attention weights are computed via dot products and subsequently normalized using the Softmax function. The normalized weights guide the dynamic weighted aggregation of the value vectors $\mathbf{V}$. Next, the feature scales are restored through concatenation, followed by the application of convolutional layers with ReLU activation to enable deep cross-modal feature fusion. Finally, an output convolutional layer maps the fused features to the target space, resulting in a stego image $I_{se} \in \mathbb{R}^{\text{batch}\times 3\times H\times W}$ that effectively balances visual quality and message embedding performance. This design significantly enhances the imperceptibility and robustness of the stego image by utilizing a locally sensitive yet globally coordinated attention mechanism for multimodal feature fusion, which

is formally expressed as:

$$I_{\text{se}} = \sum_{i \in \{\text{Fr,Fe,Fm}\}} \text{Softmax}\left(\frac{\mathbf{Q} \cdot \mathbf{K}_i}{\sqrt{d_k}}\right) \odot \mathbf{V}_i \quad (2)$$

3.4 Spatial Message Decoder

The SM Decoder module is designed to accurately extract embedded secret messages from stego or noisy images. It follows a symmetrical encoder-decoder architecture where the decoder shares parameters with the encoder and maintains hierarchical structural alignment. This symmetry ensures consistent feature attention and homologous feature extraction throughout the pipeline, reducing feature distortion during transmission and improving reconstruction accuracy and robustness. The architecture leverages prior feature knowledge from the encoder to guide decoding, employs parameter sharing to minimize error accumulation, and establishes end-to-end feature mapping to preserve message integrity. The secret message extraction process is formally defined by the following mathematical equations:

$$M' = \begin{cases} Linear(SM(I_{se})) & noise = identity \\ Linear(SM(I_{no})) & noise \neq identity \end{cases} \quad (3)$$

where *Linear* is a fully connected message extraction network.

3.5 Objectives

In this paper, we propose EGDIS and design a comprehensive loss function ($Loss$) to optimize image quality, message hiding capacity, and cross-modal feature fusion effectiveness. The loss function comprises raw image distortion loss (L_{ssid}), message hiding loss (L_{mhd}), and edge image distortion loss (L_{edge}), collectively ensuring the model's robustness and the quality of the hidden message.

The image distortion loss L_{ssid} is computed using the Mean Squared Error (MSE) between the raw and stego images, ensuring visual quality. The corresponding loss formulation is as follows:

$$L_{\text{ssid}} = \frac{1}{C \times W \times H} \sum_{i,j} \left(I_{\text{r}}(i,j) - I_{\text{se}}(i,j)\right)^2 \quad (4)$$

The message hiding loss L_{mhd} ensures the integrity of the embedded message by calculating the MSE between the original message M and the extracted message M'. The corresponding loss formulation is as follows:

$$L_{\text{mhd}} = \frac{1}{L} \sum \left(M(p) - M'(p)\right)^2 \quad (5)$$

Additionally, we introduce the edge loss L_{edge} to regulate the edge feature fusion process within the CAFN. This loss computes the mean squared error between the extracted stego edge features F_{ex} and the edge feature map F_{e}, ensuring effective integration of edges, the raw image, and the steganographic message. The corresponding loss formulation is as follows:

$$L_{\text{edge}} = \frac{1}{N} \sum \left(F_{\text{ex}} - F_{\text{e}}\right)^2 \quad (6)$$

The total loss function is defined as:

$$Loss = \lambda_{\text{ssid}} L_{\text{ssid}} + \lambda_{\text{mhd}} L_{\text{mhd}} + \lambda_{\text{edge}} L_{\text{edge}} \quad (7)$$

By jointly optimizing these loss functions, we ensure a balance among image quality, message embedding integrity, and cross-modal fusion effectiveness, thereby enhancing both the robustness and imperceptibility of the embedded message.

4 Experimental

We evaluate our method on the DocImgCN and DocImgEN datasets so that we can train and evaluate the proposed tasks simultaneously.

4.1 Datasets

We conducted experiments on two large-scale document image datasets: the DocImgCN dataset, consisting of 230,000 training, 10,000 validation, and 10,000 test images of Chinese document crops (400×400) from CNKI; and the DocImgEN dataset with identical splits, containing English document images uniformly resized to 400×400 from IEEE Xplore.

Table 1 Comparative Analysis of Different Models on DocImgCN and DocImgEN with a 64-bit Message Length.

Model	PSNR	SSIM	Acc(%)						Dataset
			Crop p=30%	Cropout p=30%	Resize p=50%	Dropout p=30%	JPEG Q=50	CN	
HiDDeN [1]	36.43	0.9	76.25	73.01	78.58	77.66	92	70.9	DocImgCN
MBRS [2]	33.67	0.89	73.27	74.9	-	74.97	67.81	69.51	
IGA [3]	11.41	0.59	56.77	58.75	65.09	55.79	66.54	54.94	
Ge et al. [6]	38.84	0.92	76.03	77.28	79.83	81.64	75.31	73.15	
DAMS [7]	38.89	0.93	94.7	80.71	99.75	85.69	99.56	79.01	
Ours	**40.12**	**0.95**	**95.1**	**81.64**	**99.80**	**86.32**	**99.64**	**80.12**	
HiDDeN [1]	34.38	0.80	75.69	72.51	76.77	75.74	81.09	68.33	DocImgEN
MBRS [2]	32.09	0.90	72.19	75.98	-	76.56	69.31	89.14	
IGA [3]	13.57	0.62	58.53	52.23	60.08	53.36	62.06	54.72	
Ge et al. [6]	38.17	0.90	75.11	76.31	77.63	79.35	73.52	71.31	
DAMS [7]	38.87	0.91	87.03	89.92	89.97	89.18	88.07	90.02	
Ours	**39.83**	**0.93**	**88.30**	**89.94**	**90.12**	**89.42**	**89.03**	**90.42**	

4.2 Metrics

We selected three key metrics for quantitative analysis: decoding accuracy (Acc), Peak Signal-to-Noise Ratio (PSNR) [18], and Structural Similarity Index Measure (SSIM) [19]. Decoding accuracy measures the model's ability to extract hidden information, reflecting its robustness. PSNR and SSIM assess the visual quality and structural similarity of stego images in comparison to the original images.

4.3 Implementation Details

We implement our study on PyTorch and use a single NVIDIA RTX 3080Ti GPU for our experiments. The model encoder was pre-trained separately on the DocImgCN and DocImgEN datasets. During training, each batch consisted of 8 images and 8 random bit strings, and the number of training epochs was set to 300. The initial learning rate was set to 1×10^{-3}, and the weights of the loss function were configured as $\lambda_{\text{ssid}} = 0.5$, $\lambda_{\text{mhd}} = 1$, and $\lambda_{\text{edge}} = 1 \times 10^{-3}$. Additionally, a standardized noise set was employed for model training to enhance robustness and ensure fair quantitative evaluation.

4.4 Comparisons

During the experimental phase, we selected two standard document image datasets, DocImgCN and DocImgEN, for systematic evaluation. To ensure rigorous comparisons, all methods used uniformly sized raw images with a resolution of 128×128 and a fixed embedded message length of 64 bits. As shown in Table 1, across various common interference scenarios (Crop, Cropout, Dropout, Resize, and JPEG [2, 3, 6, 20, 21]), our method demonstrated significant advantages in three key metrics: Acc, PSNR, and SSIM.

4.4.1 Quantitative Comparison

Due to the limited number of steganography methods specifically designed for document images, this study selected state-of-the-art natural and document image steganography algorithms for comparison, including HiDDeN, MBRS, IGA [1–3], Ge et al.'s method [6], and DAMS [7]. Experimental results demonstrate that our method achieves superior performance on DocImgCN (PSNR = 40.12 dB, SSIM = 0.95) and DocImgEN (PSNR = 39.83 dB, SSIM = 0.93), outperforming all benchmarks. Both objective metrics and subjective evaluations confirm its advantages in structural preservation and concealment quality.

4.4.2 Qualitative Comparison

The experimental results presented in Table 1 demonstrate that steganography methods originally designed for natural images perform poorly on document images. To further highlight the effectiveness of our proposed approach, we compare it with multiple state-of-the-art methods, as shown in Figure 3. Additionally, we conduct a qualitative analysis with DAMS, a representative document image steganography method. Our edge-guided approach embeds secret information into text edge regions through a combination of cross-attention mechanisms and dynamic weight

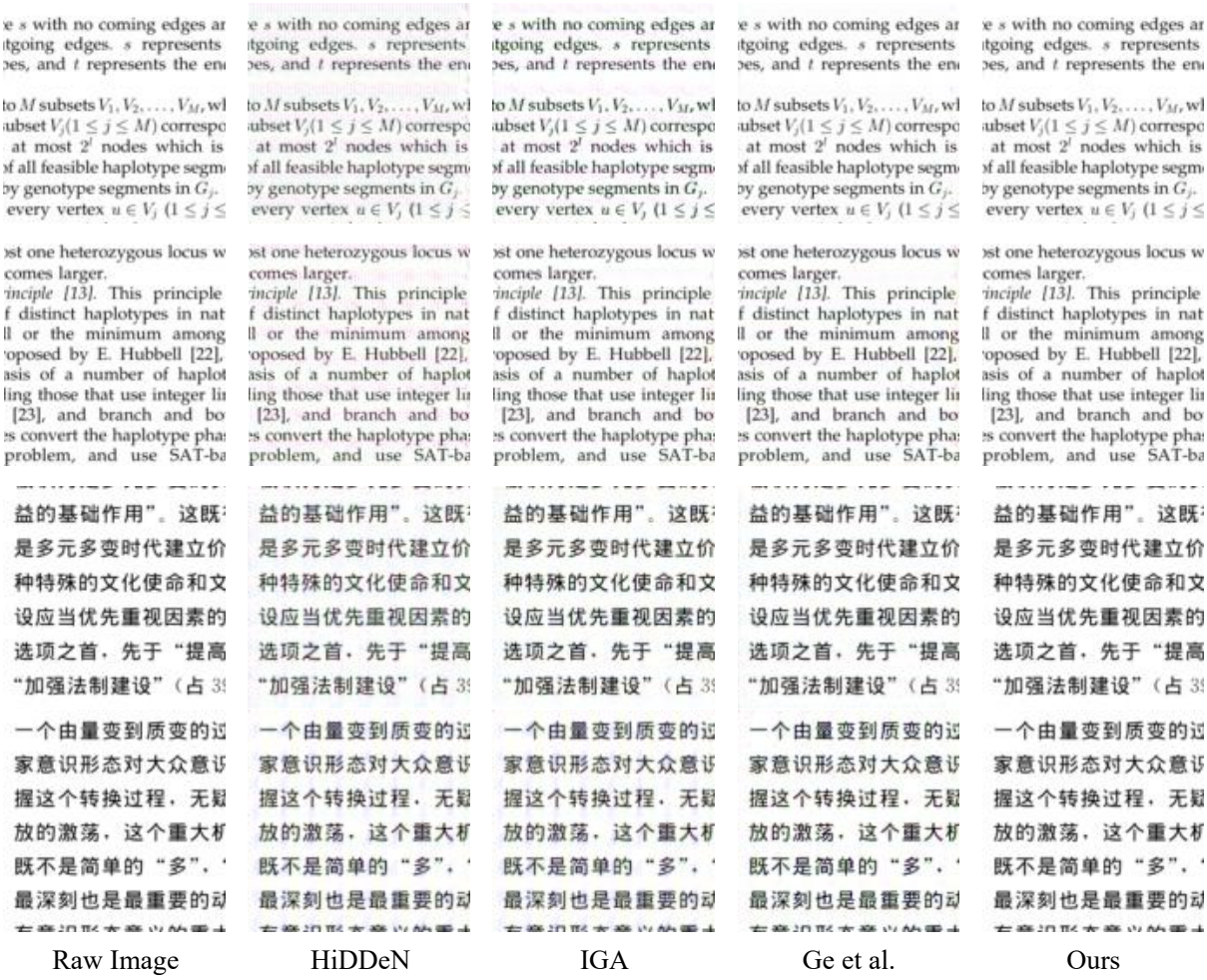


Fig. 3 Visual comparison of EGDIS and other methods with a 30-bit message length on the DocImgCN and DocImgEN datasets.

allocation strategies. This design enables precise embedding in edge areas, thereby improving both the hiding accuracy and the visual quality of stego images. Under noiseless conditions, our method achieves over 99.95% accuracy on both the DocImgCN and DocImgEN datasets, as reported in Table 2, and exhibits superior visual performance, as illustrated in Figure 4. The Embedding Location Distribution (ELD) further confirms the effective utilization of edge regions.

4.5 Ablation Study

In this section, we present ablation studies to validate the effectiveness of edge guidance and the cross-attention fusion strategy in the EGDIS model.

4.5.1 The impact of edge information

We evaluated the impact of edge features by comparing two configurations. The Baseline directly fuses the raw image and secret message, while Baseline + CAFN integrates the raw image, edge image, and message for fusion. As shown in Table 2, adding edge information significantly improves PSNR under a fixed 30-bit message length.

4.5.2 The effectiveness of different fusion strategies

We evaluated three feature fusion strategies to validate the effectiveness of the proposed CAFN. CNN-Cat fuses features by concatenating images and secret messages, while the General Attention Fusion Strategy (GAFS) uses a generic attention



Fig. 4 Comparative analysis of EGDIS and DAMS with a 30-bit message length on the DocImgCN and DocImgEN datasets: Accuracy, PSNR, and ELD.

Table 2 The role of edge information in 30-bit message length embedding: A controlled ablation study with and without edge guidance on the DocImgCN and DocImgEN datasets.

Methods	DocImgCN			DocImgEN		
	Acc(%)	PSNR(dB)	SSIM	Acc(%)	PSNR(dB)	SSIM
Baseline	99.76	39.26	0.96	99.42	36.71	0.94
Baseline + CAFN	99.99	41.81	0.99	99.96	38.79	0.97

mechanism. In contrast, CAFN employs cross-attention for precise and adaptive fusion, achieving better preservation of fine-grained textual structures. This results in improved imperceptibility in steganography, as shown in Table 3.

5 Conclusion

This paper proposes EGDIS, a method based on an encoder-decoder architecture. The model employs an adaptive Canny algorithm to extract edge images from raw document images, which serve as guidance for the precise embedding of secret messages. This edge-guided strategy significantly enhances both the imperceptibility of the secret messages and the fidelity of the stego images. To further improve edge-based guidance, we introduce a dedicated CAFN, which jointly optimizes feature representations from the raw image, edge image, and encoded secret message. Experimental results demonstrate that our method achieves state-of-the-art performance in

Table 3 The role of edge-guided fusion strategies in 30-bit message length embedding: An ablation study on the DocImgCN and DocImgEN datasets.

Dataset	Strategy	Metrics		
		Acc(%)	PSNR	SSIM
DocImgCN	CAFN(Ours)	99.99	41.81	0.99
	GAFS	$99.80_{\downarrow 0.19}$	$40.22_{\downarrow 1.59}$	$0.96_{\downarrow 0.03}$
	CNN-Cat	$99.74_{\downarrow 0.25}$	$40.01_{\downarrow 1.80}$	$0.95_{\downarrow 0.04}$
DocImgEN	CAFN(Ours)	99.96	38.79	0.97
	GAFS	$99.46_{\downarrow 0.50}$	$37.16_{\downarrow 1.63}$	$0.94_{\downarrow 0.03}$
	CNN-Cat	$99.35_{\downarrow 0.61}$	$37.05_{\downarrow 1.74}$	$0.92_{\downarrow 0.05}$

terms of imperceptibility and robustness for document image steganography.

Declarations

Conflict of interest The authors declare no competing interests.

Competing interests The authors declare no competing interests.

Funding This work was supported by the Key Research and Development Program of Henan Province under Grant [251111212000].

References

[1] J Zhu. Hidden: hiding data with deep networks. *arXiv preprint arXiv:1807.09937*, 2018.

[2] Zhaoyang Jia, Han Fang, and Weiming Zhang. Mbrs: Enhancing robustness of dnn-based watermarking by mini-batch of real and simulated jpeg compression. In *Proceedings of the 29th ACM international conference on multimedia*, pages 41–49, 2021.

[3] Honglei Zhang, Hu Wang, Yuanzhouhan Cao, Chunhua Shen, and Yidong Li. Robust data hiding using inverse gradient attention. *arXiv preprint arXiv:2011.10850*, 2020.

[4] Atique ur Rehman, Rafia Rahim, Shahroz Nadeem, and Sibt ul Hussain. End-to-end trained cnn encoder-decoder networks for image steganography. In *Computer Vision–ECCV 2018 Workshops: Munich, Germany, September 8-14, 2018, Proceedings, Part IV 15*, pages 723–729. Springer, 2019.

[5] Pin Wu, Yang Yang, and Xiaoqiang Li. Image-into-image steganography using deep convolutional network. In *Advances in Multimedia Information Processing–PCM 2018: 19th Pacific-Rim Conference on Multimedia, Hefei, China, September 21-22, 2018, Proceedings, Part II 19*, pages 792–802. Springer, 2018.

[6] Sulong Ge, Zhihua Xia, Jianwei Fei, Yao Tong, Jian Weng, and Ming Li. A robust document image watermarking scheme using deep neural network. *Multimedia Tools and Applications*, 82(25):38589–38612, 2023.

[7] Kaijiang Li, Yi Qin, Peisen Wang, Chunyi Guo, Junqi Wang, Ruiyang Jia, and Wenfeng Jiang. Dams: Document image steganography with dual attention multi-scale encoder-decoder architecture. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 118–131. Springer, 2024.

[8] Chuanlong Li, Yumeng Jiang, and Marta Cheslyar. Embedding image through generated intermediate medium using deep convolutional generative adversarial network. *Computers, Materials & Continua*, 56(2), 2018.

[9] Horia Porav, Valentina Musat, and Paul Newman. Reducing steganography in cycle-consistency gans. In *CVPR Workshops*, pages 78–82, 2019.

[10] Xintao Duan, Kai Jia, Baoxia Li, Daidou Guo, En Zhang, and Chuan Qin. Reversible image steganography scheme based on a u-net structure. *Ieee Access*, 7:9314–9323, 2019.

[11] Haichao Shi, Jing Dong, Wei Wang, Yinlong Qian, and Xiaoyu Zhang. Ssgan: Secure steganography based on generative adversarial networks. In *Advances in Multimedia Information Processing–PCM 2017:*

18th Pacific-Rim Conference on Multimedia, Harbin, China, September 28-29, 2017, Revised Selected Papers, Part I 18, pages 534–544. Springer, 2018.

[12] Denis Volkhonskiy, Ivan Nazarov, and Evgeny Burnaev. Steganographic generative adversarial networks. In *Twelfth international conference on machine vision (ICMV 2019)*, volume 11433, pages 991–1005. SPIE, 2020.

[13] Malik Qasaimeh, Alaa Abu Qtaish, and Shadi Aljawarneh. Robust steganographic approach using generative adversarial network and compressive autoencoder. *Multimedia Tools and Applications*, pages 1–38, 2024.

[14] Lei Li, Zhiyuan Zhou, Suping Wu, and Yongrong Cao. Multi-scale edge-guided learning for 3d reconstruction. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(3):1–24, 2023.

[15] Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, and Feng Wu. Multi-modality cross attention network for image and sentence matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10941–10950, 2020.

[16] Xin Wen, Shuo Li, Hao Yu, and Yu He. Multiscale context feature and cross-attention network-enabled system and software-based for pavement crack detection. *Engineering Applications of Artificial Intelligence*, 127:107328, 2024.

[17] Fulin Xu, Shaohui Mei, Ge Zhang, Nan Wang, and Qian Du. Bridging cnn and transformer with cross attention fusion network for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[18] Adel Almohammad and Gheorghita Ghinea. Stego image quality and the reliability of psnr. In *2010 2nd International Conference on Image Processing Theory, Tools and Applications*, pages 215–220. IEEE, 2010.

[19] De Rosal Igantius Moses Setiadi. Psnr vs ssim: imperceptibility quality assessment for image steganography. *Multimedia Tools and Applications*, 80(6):8423–8444, 2021.

[20] Rui Ma, Mengxi Guo, Yi Hou, Fan Yang, Yuan Li, Huizhu Jia, and Xiaodong Xie. Towards blind watermarking: Combining invertible and non-invertible mechanisms. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1532–1542, 2022.

[21] Lin Sun, Jiucheng Xu, Shangwang Liu, Shiguang Zhang, Yuan Li, and Chang'an Shen. A robust image watermarking scheme using arnold transform and bp neural network. *Neural Computing and Applications*, 30:2425–2440, 2018.