

Stego Guard: A Residual Cnn-Based Framework for Image Steganalysis and Browser-Integrated Detection

J Omar Farooq¹, Keerthi K. S², Hamdan Shamsuddin Kazi³, Hament Kumar⁴, Bikram Chaurasiya⁵

¹UG-Computer Science and Engineering, Atria Institute of Technology, Hebbal, Bengaluru, Karnataka 560024, India

^{2,3,4,5}Assistant Professor-Computer Science and Engineering, Atria Institute of Technology, Hebbal, Bengaluru, Karnataka 560024, India

Emails: Omarashtab@gmail.com¹, Keerthi_cse@atria.edu², Kazihamdan@gmail.com³, Hamentkumar.8449@gmail.com⁴, Bikramchaurasia733@gmail.com⁵

Abstract

The increasing use of digital platforms has made images one of the most frequently exchanged forms of online content. Alongside legitimate communication, images can also be employed as carriers for concealed information through steganographic techniques. Steganography attempts to hide the existence of a secret message by embedding it within a supposedly ordinary digital medium. This characteristic makes the identification of hidden information an important problem in digital image security. The proposed StegoGuard framework addresses this problem using a Residual Convolutional Neural Network (CNN) for image steganalysis. The system divides input images into 128×128 pixel patches and evaluates these regions individually using the trained Residual CNN. The resulting patch-level predictions are used to obtain an image-level stego score, which is compared against a threshold of 0.66 to classify the image as CLEAN or STEGO. The detection model is integrated with a browser extension through Native Messaging, allowing images encountered during web browsing to be analysed through a local detection application. The experimental evaluation achieved an accuracy of approximately 96% on the dataset used in this study. The results demonstrate the potential of combining Residual CNN-based analysis, patch-level detection, and browser integration for practical image steganalysis. The presence of false-positive classifications also indicates the need for more diverse datasets and broader evaluation.

Keywords: Browser Extension; Deep Learning; Image Classification; Image Steganalysis; Steganography

1. Introduction

Steganography is a technique used to conceal secret information within an apparently ordinary digital medium. In image steganography, information is embedded into image data in a manner intended to minimize noticeable changes to the visual appearance of the image. While this technique can be used for legitimate purposes such as confidential communication and information protection, it can also be misused to conceal information within digital images. The detection of hidden information is addressed through image steganalysis. Traditional steganalysis approaches commonly rely on statistical characteristics and manually designed features to distinguish between clean and steganographic images. Rich-model-based approaches demonstrated the effectiveness of extracting a large set of statistical features for identifying subtle changes introduced by

data embedding (Fridrich and Kodovský, 2012; Kodovský et al., 2012). However, these approaches generally depend on carefully designed feature representations. The development of deep learning has provided an alternative approach in which relevant image representations can be learned directly from training data. Convolutional Neural Networks (CNNs) have been investigated for image steganalysis and have demonstrated the ability to learn features associated with steganographic modifications (Xu et al., 2016; Ye et al., 2017). Subsequent research has explored wider network structures, residual architectures, and other deep-learning approaches to improve the detection of subtle embedding-related characteristics (Zeng et al., 2019; Boroumand et al., 2019). Different deep-learning strategies have also been investigated for

improving image steganalysis under varying conditions. These include the use of visual attention, content-adaptive steganography scenarios, and combinations of conventional steganalysis features with CNN-based classification (Hu et al., 2019; Zhang et al., 2019). Reviews of deep-learning-based steganalysis further indicate the increasing role of CNN architectures in detecting hidden information from digital images (Tabares-Soto et al., 2019). More recent approaches have examined improvements in CNN-based steganalysis through alternative network structures and feature-selection mechanisms. Content-adaptive image steganography and spatial-domain detection have been investigated using CNN-based methods, while selection-channel information has also been incorporated to improve the representation of embedding-related characteristics (Xiang et al., 2020; Ren et al., 2020). Siamese and other deep-learning architectures have further been explored for image steganalysis (You et al., 2021). Studies examining CNN performance have also highlighted the importance of training-data composition, generalization, and evaluation under different image conditions (Tabares-Soto et al., 2021). Despite these developments, practical image steganalysis systems are not always directly integrated into the environment in which users encounter digital images. This creates a gap between the detection capability of a trained model and its practical accessibility during ordinary image browsing. The proposed StegoGuard framework addresses this issue by combining Residual CNN-based image analysis with a browser extension and a local detection application. The framework divides an input image into 128×128 patches, evaluates the individual regions using the trained Residual CNN, and uses the resulting predictions to determine an image-level stego score. The proposed system further connects the detection model to a browser extension through Native Messaging, allowing images selected during browsing to be transferred to the local detection application for analysis. The resulting classification is subsequently returned to the browser interface. This integration provides a unified workflow in which image acquisition, preprocessing, localized analysis, classification, and result presentation are connected within the browsing

environment. The objective of this work is therefore to develop and evaluate a browser-integrated image steganalysis framework capable of distinguishing clean images from images containing possible steganographic modifications. The proposed approach combines localized 128×128 patch analysis, Residual CNN-based feature learning, image-level score generation, and browser connectivity into a single detection framework. The experimental evaluation investigates the effectiveness of this approach on the dataset considered in this study.

2. Method

2.1. StegoGuard Framework

The proposed StegoGuard framework integrates image preparation, localized image analysis, deep feature extraction, classification, and browser communication into a unified detection process. The system first prepares the image data and divides each input into fixed-size regions. These regions are then examined using a Residual Convolutional Neural Network, whose outputs are used to determine whether the image is likely to contain steganographic content. The trained detection model is further connected to a browser extension using Native Messaging. The major stages of the proposed framework are described below. As shown in Figure 1 Overall workflow of the proposed StegoGuard framework [1 – 5].

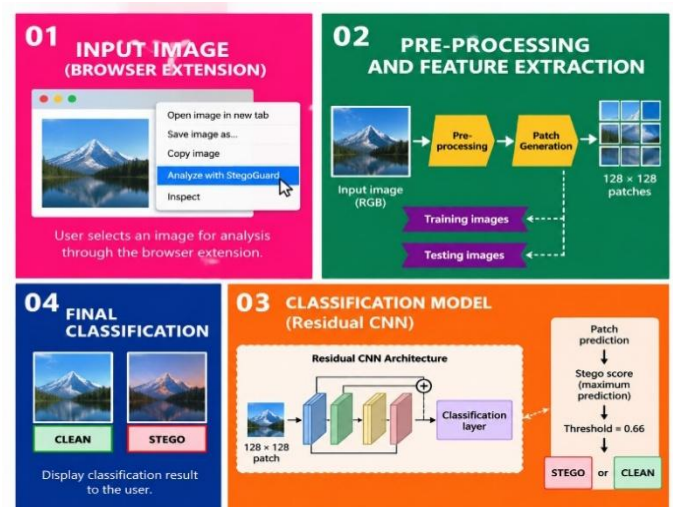


Figure 1 Overall workflow of the proposed StegoGuard framework

2.2. Image Preparation and Patch Generation

Each image is prepared in a consistent format before being passed to the detection model. The input images are represented in RGB format and separated into 128×128 pixel patches. The patch-based strategy enables the system to inspect localized areas rather than relying exclusively on a representation of the complete image. This is especially useful for steganalysis because embedding-related changes may be subtle and may occur within specific portions of an image. Examining smaller portions of the image enables the network to focus on local characteristics that may be difficult to distinguish when the entire image is processed at once. After extraction, every patch is supplied to the trained Residual CNN for further analysis [6 – 10].

2.3. Residual CNN-Based Analysis

The principal learning component of StegoGuard is a Residual Convolutional Neural Network. The network is trained to learn image representations that can help separate clean image patches from patches associated with steganographic content. During processing, the convolutional layers transform the input patch layer by layer, producing higher-level feature representations. These representations can capture image patterns that may not be apparent during ordinary visual examination. The use of residual architecture provides shortcut connections between different stages of the network. Shortcut connections within the residual architecture help preserve and transmit information between deeper network layers. Rather than forcing every layer to learn an entirely new transformation, the residual structure allows the network to learn the required changes relative to information already available from earlier layers. This characteristic is useful for detecting subtle image alterations. Steganographic embedding may produce weak and distributed signals that are difficult to identify visually. The Residual CNN therefore learns representations from the input patches that can assist in recognizing these characteristics. The feature representations generated by the network are then passed to the classification stage. A prediction is produced for each processed image patch, providing the basis for the final image-level decision. As shown in Figure 2 Residual CNN architecture used for image-patch classification.

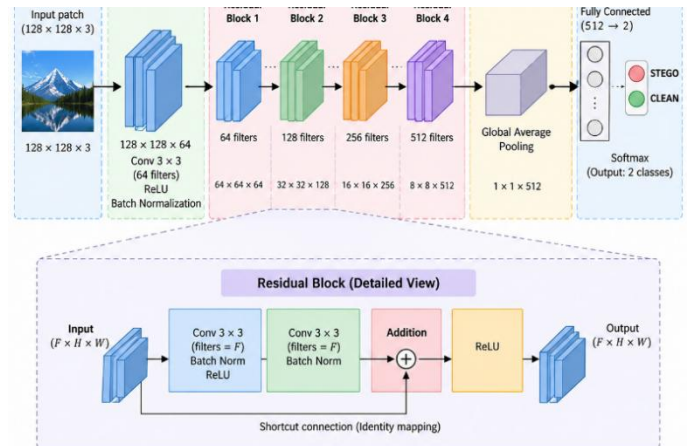


Figure 2 Residual CNN architecture used for image-patch classification

2.4. Image-Level Classification

After feature extraction, the network generates an individual prediction for each 128×128 image patch. The patch-level outputs are used to determine an overall score for the original image. In the proposed implementation, the highest prediction among the processed patches is selected as the image-level stego score. The image-level score is then assessed using 0.66 as the decision threshold. A score equal to or greater than 0.66 results in a STEGO classification, whereas a score below 0.66 results in a CLEAN classification. This approach allows a localized area that provides a strong indication of steganographic modification to affect the classification of the image as a whole [11].

2.5. Browser Integration Using Native Messaging

StegoGuard incorporates browser-based image analysis by linking the trained detection model with a browser extension and a local processing application. The extension serves as the interface through which the user initiates image inspection, while the local application performs the required analysis. Once an image is selected, the extension forwards the image to the local application for processing. The browser extension communicates with the local application through Native Messaging. The received image undergoes preprocessing within the local application. It is converted to the required

representation and partitioned into 128×128 patches. The resulting patches are individually evaluated by the Residual CNN, which produces a prediction for each region. These individual predictions are thereafter considered to obtain a single stego score for the complete image. The score is then evaluated using 0.66 as the classification threshold. Images with a score of 0.66 or above are classified as STEGO, while those with a score below 0.66 are classified as CLEAN. Once the classification is completed, the result is returned to the browser extension through Native Messaging. The workflow of the proposed system can be expressed as: Image acquisition $\rightarrow$ image preparation $\rightarrow$ 128×128 patch division $\rightarrow$ Residual CNN analysis $\rightarrow$ patch predictions $\rightarrow$ stego score calculation $\rightarrow$ CLEAN/STEGO decision $\rightarrow$ browser display. This integration allows the detection procedure to be carried out within the web-browsing environment, eliminating the need for the user to separately operate the image detection application[12].

3. Results And Discussion

3.1.Results

The proposed StegoGuard model was evaluated using the testing subset of the dataset to assess its ability to distinguish between clean images and images containing steganographic modifications. The model achieved an accuracy of approximately 96% on the evaluated dataset, demonstrating that the Residual CNN was able to distinguish between the two image categories within the experimental setting. The patch-based detection process was applied to the test images by dividing each input image into 128×128 pixel regions. Each region was independently processed by the Residual CNN, producing an individual prediction for the corresponding image patch. These patch-level predictions were then used to determine the final stego score for the complete image. For each test image, the highest prediction obtained from the processed patches was selected as the image-level stego score. The resulting score was compared with the predefined classification threshold of 0.66. Images receiving a score equal to or greater than 0.66 were classified as STEGO, whereas images with a score below 0.66 were classified as CLEAN. The evaluation also demonstrated the operation of the browser-integrated detection process. Images

selected through the browser extension were transferred to the local detection application using Native Messaging. The local application performed the required preprocessing and Residual CNN inference, after which the resulting classification was returned to the browser interface. The experimental results therefore demonstrate that the proposed StegoGuard framework can perform image classification based on localized patch analysis while integrating the detection process with a browser environment[13].

3.2.Discussion

The results indicate that the proposed StegoGuard framework can effectively distinguish between clean images and images containing steganographic modifications within the evaluated dataset. The approximately 96% accuracy obtained during testing suggests that the Residual CNN learned image characteristics that were useful for identifying the differences associated with steganographic modification. The use of 128×128 patches allows the model to examine localized regions of an image during detection. This is particularly relevant to steganalysis because embedding-related characteristics may be subtle and may occur within specific portions of an image. By processing individual regions, the model can consider localized information rather than relying solely on characteristics derived from the complete image. For the final image-level decision, the highest prediction among the processed patches is selected as the stego score. This approach allows a localized region producing a strong indication of steganographic modification to influence the classification of the complete image. The predefined threshold of 0.66 provides the decision boundary for assigning the image to either the STEGO or CLEAN category. The browser-integrated architecture provides a practical means of using the detection system during normal browsing activity. The browser extension acts as the user-facing component, while the local application performs image preprocessing and Residual CNN-based analysis. Communication between these components through Native Messaging enables the classification result to be returned to the browser after processing. The evaluation also revealed the presence of false-positive classifications, where some clean

images were identified as steganographic. Such classifications indicate that certain characteristics naturally present in clean images may resemble patterns associated with steganographic content learned by the model[14]. This limitation emphasizes the importance of dataset diversity when developing an image-steganalysis system. A training dataset with limited variation may not adequately represent the range of image characteristics encountered during practical use. Increasing the size and diversity of the dataset, together with testing on a wider range of previously unseen images, could therefore provide a stronger assessment of the model's generalization capability. Overall, the findings suggest that the combination of Residual CNN-based feature learning, localized patch analysis, and browser integration provides a practical direction for image steganalysis. Further evaluation with larger and more diverse datasets would be required to determine whether the observed performance can be maintained under broader real-world image conditions[15].

Conclusion

The proposed StegoGuard framework presents an integrated approach for detecting possible steganographic content in digital images using a Residual Convolutional Neural Network and browser-based image analysis. The framework divides input images into 128×128 pixel patches and evaluates these regions individually to obtain patch-level predictions. The highest prediction is used to determine the image-level stego score, which is classified using a threshold of 0.66. The experimental evaluation achieved an accuracy of approximately 96% on the dataset used in this study, demonstrating the ability of the trained model to distinguish between clean and steganographic images within the evaluated experimental setting. The browser extension and Native Messaging integration further enable the detection process to be incorporated into the web-browsing environment. Although the results are encouraging, the presence of false-positive classifications indicates that the framework may incorrectly identify certain clean images as steganographic. Therefore, further evaluation using larger and more diverse datasets is necessary to assess the generalization of the proposed approach under a wider range of image conditions. Future

improvements may include expanding the training dataset, evaluating additional steganographic techniques, and optimizing the detection and browser communication processes. Overall, StegoGuard demonstrates the potential of combining deep-learning-based image steganalysis with browser integration to provide a practical mechanism for detecting possible hidden information in digital images.

Acknowledgements

The authors would like to express their sincere gratitude to Prof. Keerthi K. S. for her valuable guidance, encouragement, and support throughout the development and evaluation of the proposed StegoGuard framework. The authors also extend their gratitude to the Department of Computer Science and Engineering, Atria Institute of Technology, Bengaluru, Karnataka, India, for providing the academic environment and resources necessary to carry out this work.

References

- [1]. Fridrich, J., & Kodovský, J. (2012). Rich models for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, 7(3), 868–882. doi: 10.1109/TIFS.2012.2190402.
- [2]. Kodovský, J., Fridrich, J., & Holub, V. (2012). Ensemble classifiers for steganalysis of digital media. *IEEE Transactions on Information Forensics and Security*, 7(2), 432–444. doi: 10.1109/TIFS.2011.2175919.
- [3]. Holub, V., Fridrich, J., & Denemark, T. (2014). Universal distortion function for steganography in an arbitrary domain. *EURASIP Journal on Information Security*, 2014, 1. doi: 10.1186/1687-417X-2014-1.
- [4]. Xu, G., Wu, H. Z., & Shi, Y. Q. (2016). Structural design of convolutional neural networks for steganalysis. *IEEE Signal Processing Letters*, 23(5), 708–712. doi: 10.1109/LSP.2016.2548421.
- [5]. Ye, J., Ni, J., & Yi, Y. (2017). Deep learning hierarchical representations for image steganalysis. *IEEE Transactions on Information Forensics and Security*, 12(11), 2545–2557. doi: 10.1109/TIFS.2017.2710946.

- [6]. Zeng, J., Tan, S., Liu, G., Li, B., & Huang, J. (2019). WISERNet: Wider separate-then-reunion network for steganalysis of color images. *IEEE Transactions on Information Forensics and Security*, 14(10), 2735–2748. doi: 10.1109/TIFS.2019.2904413.
- [7]. Boroumand, M., Chen, M., & Fridrich, J. (2019). Deep residual network for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, 14(5), 1181–1193. doi: 10.1109/TIFS.2018.2871749.
- [8]. Hu, D., Zhou, S., Shen, Q., Zheng, S., Zhao, Z., & Fan, Y. (2019). Digital image steganalysis based on visual attention and deep reinforcement learning. *IEEE Access*, 7, 25924–25935. doi: 10.1109/ACCESS.2019.2900076.
- [9]. Zhang, T., Zhang, H., Wang, R., & Wu, Y. (2019). A new JPEG image steganalysis technique combining rich model features and convolutional neural networks. *Mathematical Biosciences and Engineering*, 16(5), 4069–4081. doi: 10.3934/mbe.2019201.
- [10]. Tabares-Soto, R., Ramos-Pollán, R., & Isaza, G. (2019). Deep learning applied to steganalysis of digital images: A systematic review. *IEEE Access*, 7, 68970–68990. doi: 10.1109/ACCESS.2019.2918086.
- [11]. Xiang, Z., Sang, J., Zhang, Q., Cai, B., Xia, X., & Wu, W. (2020). A new convolutional neural network-based steganalysis method for content-adaptive image steganography in the spatial domain. *IEEE Access*, 8, 47013–47020. doi: 10.1109/ACCESS.2020.2978110.
- [12]. Ren, W., Zhai, L., Jia, J., Wang, L., & Zhang, L. (2020). Learning selection channels for image steganalysis in spatial domain. *Neurocomputing*, 401, 78–90. doi: 10.1016/j.neucom.2020.02.105.
- [13]. You, W., Zhang, H., & Zhao, X. (2021). A Siamese CNN for image steganalysis. *IEEE Transactions on Information Forensics and Security*, 16, 291–306. doi: 10.1109/TIFS.2020.3013204.
- [14]. Tabares-Soto, R., Arteaga-Arteaga, H. B., Mora-Rubio, A., Bravo-Ortíz, M. A., Arias-Garzón, D., Alzate-Grisales, J. A., Burbano-Jacome, A., Orozco-Arias, S., Isaza, G., & Ramos-Pollán, R. (2021). Strategy to improve the accuracy of convolutional neural network architectures applied to digital image steganalysis in the spatial domain. *PeerJ Computer Science*, 7, e451. doi: 10.7717/peerj-cs.451.
- [15]. Tabares-Soto, R., Arteaga-Arteaga, H. B., Mora-Rubio, A., Bravo-Ortíz, M. A., Arias-Garzón, D., Alzate-Grisales, J. A., Orozco-Arias, S., Isaza, G., & Ramos-Pollán, R. (2021). Sensitivity of deep learning applied to spatial image steganalysis. *PeerJ Computer Science*, 7, e616. doi: 10.7717/peerj-cs.616.