Information technology for detecting hidden threats in multimedia based on a hybrid vision transformer architecture*

Dmytro Denysiuk[1,†], Bohdan Savenko[1,†], Oksana Onyshko[2,†], Pavlo Rehida[1,*,†] and Anatoliy Sachenko[3,†]

[1] *Khmelnytskyi National University , 11, Instytuts'ka str., Khmelnytskyi, 29016, Ukraine*
[2] *Khmelnytskyi National University , 11, Instytuts'ka str., Khmelnytskyi, 29016, Ukraine*
[3] *Kazimierz Pulaski University, The Radom, Poland*

Abstract

This paper presents an information technology for identifying concealed cyber threats in multimedia objects, with particular attention to adaptive steganographic insertions generated by modern models and multiformal structures hidden inside multimedia containers. The study shows that detectors relying mainly on local convolutional processing or statistical feature descriptions do not provide sufficient detection capability when the concealed impact is distributed across several representation levels. Their limitation is caused by the inability to capture long-range dependencies and to establish relationships between visual, structural, and behavioral feature domains.

To address these shortcomings, a hybrid detection architecture is proposed. It combines spectral-visual analysis implemented through a Wavelet-ViT module with behavioral sequence processing based on Transformer-XL. The main scientific contribution consists in the use of a controlled cross-modal feature integration mechanism, which enables coordinated analysis of heterogeneous indicators within a unified decision space. Experimental modeling on aggregated datasets demonstrated the capability of the proposed technology to detect hidden threats with an accuracy of 97.4% and a false positive rate of 0.8%.

Keywords

hidden threat detection, multimedia objects, Vision Transformer, wavelet analysis, cross-modal feature fusion, behavioral analysis, multiformal multimedia structures, multimodal cyber threat detection., CEUR-WS

1. Introduction

In modern digital systems, multimedia objects, including images, video materials, and combined container structures, are widely used as regular elements of distributed information environments. Their constant circulation through web resources, social platforms, and cloud services often creates a perception that such objects are neutral and safe data carriers [1]. However, this apparent legitimacy can be used to conceal cyber influence, when a multimedia object preserves its expected visual form while simultaneously carrying hidden commands, embedded fragments of harmful functionality, or control instructions for unauthorized network activity [2]. Thus, multimedia objects should not be regarded only as passive visual data, but also as potential carriers of concealed cyber impact within web-based information systems [3].

* *IntelITSIS'26: The 7th International Workshop on Intelligent Information Technologies & Systems of Information Security, July 3, 2026, Khmelnytskyi, Ukraine*
[1,*] Corresponding author.
[†] These authors contributed equally.
denysiuk@khmnu.edu.ua (D. Denysiuk); savenko_bohdan@ukr.net (B. Savenko); van4o@ukr.net (O. Onyshko); pavlo.rehida@gmail.com (P. Rehida); sachenkoa@yahoo.com (A. Sachenko)
0000-0002-7345-8341 (D. Denysiuk); 0000-0001-5647-9979 (B. Savenko); 0000-0002-2125-4160 (O. Onyshko); 0000-0002-6591-7069 (P. Rehida); 0000-0002-0907-3682 (A. Sachenko)

1.1. Relevance of AI-driven steganographic threats

The relevance of this study is determined by the rapid transformation of methods used to conceal cyber threats inside multimedia objects. The emergence of generative and diffusion-based models has made it possible to create steganographic representations whose statistical properties remain close to those of unchanged visual data. Such objects may preserve natural entropy characteristics and avoid typical noise artifacts, which significantly complicates their detection by conventional statistical analysis methods.

A further challenge is associated with attacks in which hidden functionality is activated during the processing of a multimedia object by web mechanisms, including operations related to the Canvas API. In such cases, the threat is not always visible at the level of the image structure itself, since its manifestation may depend on the interaction between the object and the software environment. Multiformal container structures that simultaneously satisfy the requirements of several formats, for example an image representation combined with executable script logic, create an additional masking layer. Therefore, a multimedia object should be considered not only as a passive carrier of visual information, but also as a potential source of active concealed cyber impact, which requires the development of detection approaches capable of analyzing visual, structural, and behavioral indicators jointly.

1.2. Limitations of CNN/LSTM-based detection models

Previous approaches to the detection of hidden threats were based on two-stage information technologies that combined static analysis of visual data using convolutional neural networks (CNN) with dynamic monitoring of system activity through recurrent architectures (LSTM). These models made it possible to achieve classification accuracy of up to 92.8% [4].

However, the use of convolutional architectures is constrained by the local nature of convolution operations and by the limited receptive field, which complicates the detection of long-range dependencies between spatially distant regions of a multimedia object. Low-capacity generative embedding methods distribute concealed information across the whole object, forming weak dependencies that are not reliably captured by local filters. In addition, shift invariance and pooling operations in CNNs may lead to the loss of microscopic high-frequency changes that are important for steganalysis. LSTM-based recurrent networks also have limitations when processing long sequences of system calls due to gradient degradation. Therefore, unimodal and cascaded CNN/LSTM models are not sufficiently reliable for analyzing AI-driven threats in which hidden impact is distributed across visual, structural, and behavioral domains.

1.3. Purpose, objectives, and scientific contribution of the research

Modern cybersecurity approaches increasingly rely on the Zero Trust Architecture paradigm [4], which requires continuous verification of information objects regardless of their origin, declared format, or presumed trust level. Accordingly, multimedia objects cannot be treated as inherently safe, since images, video materials, and combined containers may include concealed functionality, embedded commands, steganographic insertions, or structural anomalies that become active during processing by a browser, decoder, server-side parser, or other software component.

The purpose of this article is to develop and theoretically justify a multimodal information technology for detecting hidden threats in multimedia objects using a hybrid Vision Transformer-based architecture [6]. The proposed approach combines spectral-visual analysis and behavioral pattern processing, enabling the detection of weakly expressed threats that become identifiable only through joint analysis of heterogeneous indicators. To achieve this purpose, the study formalizes the detection task in a multimodal feature space, develops wavelet-oriented tokenization for high-frequency anomaly analysis, constructs a behavioral encoder for system call sequences,

and applies Gated Cross-Attention Fusion to integrate visual and dynamic features within a unified decision-making space.

The study shifts from a cascaded two-level detection scheme to a hybrid architecture with cross-modal attention. This enables the model to establish relationships between spectral characteristics of multimedia objects and behavioral patterns observed during their processing. Such integration is aimed at detecting hidden threats whose concealed influence is distributed across visual, structural, and behavioral domains. The expected result is increased robustness against AI-generated steganograms, multiformal multimedia objects, and covert command-and-control scenarios, which is consistent with Zero Trust requirements for continuous object verification [7].

2. Related works

To assess the capability of information technologies to detect hidden threats in multimedia objects, this study analyzes three representative groups of solutions used in web environments. These groups include specialized forensic tools, advanced neural network detectors belonging to the State-of-the-Art (SOTA) class [8], and integrated enterprise protection systems of the DLP and WAAP categories. The examination of these approaches is necessary for identifying their methodological constraints under the conditions of evolving cyber threats and for substantiating the need to develop new architectural solutions for multimodal detection.

2.1. CNN-based detection model limitations

The evaluation of steganographic threat detection methods in web environments requires consideration of several categories of existing solutions, including specialized forensic instruments, advanced neural network detectors of the State-of-the-Art (SOTA) class, and integrated DLP- and WAAP-based protection systems [9]. Among intelligent steganalysis approaches, convolutional neural networks (CNN) [10] are widely used because they can automatically form hierarchical feature representations from visual data without manual feature engineering.

A representative architecture in this direction is SRNet (Steganalysis Residual Network) [11], which relies on the sequential accumulation of residual features. Its internal structure includes preprocessing blocks intended to separate high-frequency noise residuals from the dominant low-frequency visual component of the image. This makes it possible to focus the analysis on weak changes that may indicate concealed embedding. However, the functioning of such models is fundamentally determined by the local convolution operation:

$$y_{i,j} = \sum_{m=-r}^{r} \sum_{n=-r}^{r} w_{m,n} \cdot x_{i+m,j+n} \tag{1}$$

where $y_{i,j}$ denotes the convolution result at spatial position i, j, $w_{m,n}$ is the kernel weight, $x_{i+m,j+n}$ is the input pixel value within the local neighborhood, and r defines its radius. Since convolution operates within a limited spatial area, the receptive field remains restricted, which complicates the detection of dependencies between distant regions of a multimedia object. Increasing network depth can partially address this issue, but it may also cause gradient degradation and suppress weak amplitude deviations that contain hidden diagnostic information. This is especially important for diffusion- or GAN-based embedding methods [12], where concealed data are distributed across the entire visual representation and form long-range correlations that local filters cannot reliably capture. An additional limitation is the shift invariance of convolutional architectures, although the position of frequency coefficients is critical in steganalysis. In JPEG images, even a one-pixel shift can change DCT block statistics, while pooling operations may smooth these changes. Moreover, SRNet- and XuNet-type models remain vulnerable to adversarial influence, where structured noise can reduce detection accuracy from about 90% to 10–18%.

2.2. Statistical rich models and their limitations for detecting AI-generated hidden threats

Before the widespread use of deep learning methods, hidden information detection was mainly based on Spatial Rich Models (SRM) [13]. This approach uses groups of linear and nonlinear high-frequency filters to obtain noise residuals from the image structure. After that, the relationships between neighboring quantized residual values are analyzed. As a result, the feature vector is represented through empirical frequencies of transitions between these values:

$$f_{SRM} = hist\left(D\left(x, y\right)\right) \tag{2}$$

where f_{SRM} represents the feature vector obtained within the Spatial Rich Model, $D\left(x, y\right)$ defines the residual difference function evaluated in a local pixel neighborhood, and $hist\left(D\left(x, y\right)\right)$ describes the empirical distribution of quantized residual values. The principal limitation of SRM-based methods is their predefined feature space, which makes them poorly adaptable to newly developed embedding strategies. Approaches such as HILL and S-UNIWARD [14] are constructed to introduce changes in a way that minimizes distortions in the same residual characteristics analyzed by SRM filters. Moreover, diffusion-based generative steganography can produce statistical properties close to those of natural images. As a result, the detection capability of statistical models is significantly reduced. In the case of JPEG images with low embedding capacity, SRM-based detectors may demonstrate accuracy below 50%.

2.3. Unimodal detection constraints and the need for system context analysis

The use of general-purpose neural network classifiers, such as EfficientNet and ResNet, trained on large-scale datasets like ImageNet, shows limited applicability in steganalysis because of the semantic gap between classification and hidden signal detection. These architectures are mainly oriented toward extracting high-level visual features, while steganographic changes usually have low amplitude and may be close to natural sensor noise. In addition, pooling operations [15] aggregate local data, which can suppress high-frequency anomalies at the early stages of processing:

$$y = \max_{i \in R}\left(x_i\right) \tag{3}$$

where R denotes the local region over which aggregation is performed. A single-modal detection approach remains incomplete because it does not include the behavioral context in which a multimedia object is processed. Analysis limited only to the pixel domain reduces the ability to detect Stegosploit-type threats or multiformal container structures that may trigger unauthorized actions through web APIs or system calls without producing noticeable statistical deviations in the image itself. Previous research [16] demonstrated that the combination of static analysis with dynamic monitoring based on recurrent networks (LSTM) can achieve an accuracy of 92.8%. However, this scheme is still constrained when long event sequences must be processed. The separation between structural analysis of a multimedia object and the observation of its dynamic effects remains a significant obstacle to building reliable protection systems.

2.4. Transformer architectures as a basis for multimodal threat analysis

Current CNN- and SRM-based detectors are not sufficient for hidden threat detection in multimedia objects, since they mainly focus on local or predefined statistical features and do not reliably

capture relationships distributed across the entire object. This becomes a critical limitation in the Zero Trust paradigm, where each multimedia object must be analyzed as a potentially untrusted entity. Vision Transformer (ViT) models provide a way to overcome this constraint by applying self-attention, which enables the analysis of dependencies between all image tokens from the first processing stages [17]. The attention matrix is defined through the interaction of token representations:

$$Attention(Q, K, V) = \text{softmax}\left(\frac{Q K^T}{\sqrt{d_k}}\right)V \tag{4}$$

where Q, K, V denote the query, key, and value vectors, while d_k defines the dimensionality of the latent feature space. This mechanism enables the model to detect inconsistencies distributed across different regions of a multimedia object, which is important for adaptive steganographic embedding. Cross-Attention [18] further supports multimodal detection by comparing visual indicators with behavioral descriptors. The combination of ViT-based visual analysis with the Zero Trust concept [19] helps overcome the locality limitations of convolutional models and increases detection robustness [20]. Therefore, a hybrid architecture integrating visual attention with system event encoders is considered a promising basis for achieving detection accuracy above 95% [21].

3. Concealed threat detection in multimedia objects

The proposed information technology is based on the assumption that hidden threats in multimedia objects can be detected through the relationship between frequency-domain microstructural deviations and the behavior of system processes initiated during object processing. Unlike conventional approaches that treat a media object as a static byte sequence, this technology considers it as an active element whose impact may appear through system call sequences and addressed memory operations [22]. Therefore, the detection task is formalized as finding an optimal mapping Φ from the joint space of visual data X and behavioral sequences S to the probability space of malicious intervention:

$$\Phi : X \times S \tag{5}$$

where the space $X \subset R^{H \times W \times C}$ describes the pixel representation of the multimedia object, while $S = \{e_1, e_2, \ldots, e_T\}$ defines an ordered sequence of system events recorded during its processing. Such events may include API calls associated with process creation, memory allocation, or writing data into an addressed memory area, including operations such as *CreateProcess*, *VirtualAllocEx*, and *WriteProcessMemory*. The presence of a hidden threat is then estimated as a conditional probability that depends on the joint analysis of the visual representation and the behavioral sequence:

$$\Phi(X, S) = P(y = 1 | X, S) \tag{6}$$

where the condition $y = 1$ indicates that the multimedia object is associated with a compromised or malicious state. The proposed approach differs from cascaded Late Fusion schemes [23], since it introduces cross-modal interaction at earlier stages of feature processing. This allows the model to identify dependencies between anomalous frequency coefficients and atypical memory-related system activity already within intermediate neural network layers. The technology is therefore

based on spectral-behavioral correlation, where anomalies are verified simultaneously in two independent representation domains, increasing the reliability of hidden threat detection [24].

3.1. Tokenization of multimedia objects using wavelet decomposition

To reduce the influence of high-level semantic image information, which may mask weak steganographic deviations, a modified tokenization procedure for the Vision Transformer is introduced. Instead of dividing the original RGB representation directly into spatial patches, the input image $I \in R^{H \times W}$ is first processed using a first-level discrete wavelet transform with the Haar basis [24]. This transformation decomposes the image into four subbands $I_{DWT} = \mathrm{DWT}(I) = \{I_{\mathrm{LL}}, I_{LH}, I_{HL}, I_{HH}\}$. The low-frequency component I_{LL} is excluded from further processing because it mainly contains semantic visual information and may reduce sensitivity to hidden low-amplitude changes [26]. In contrast, the high-frequency subbands I_{LH}, I_{HL}, I_{HH} preserve local variations, edge structures, and noise-like residuals that are more informative for steganalysis. These components are divided into non-overlapping patches of size $P \times P$. For each spatial position p, the corresponding high-frequency fragments are concatenated into a single feature vector:

$$x_p = \mathrm{Concat}\left(I_{LH}^p, I_{HL}^p, I_{HH}^p\right), x_p \in R^{3P^2}. \tag{7}$$

To match the dimensionality D of the transformer latent space, each combined vector is transformed by a linear projection with a trainable weight matrix E. As a result, the initial token embeddings are obtained in the form $z_p = x_p E$. Since the self-attention mechanism does not inherently preserve the spatial order of input tokens, positional information is introduced by adding the encoding matrix $E_{pos}$to the generated embeddings. The formed matrix of visual representations$Z_{vis}^{(0)}$ is then passed to the sequence of Vision Transformer blocks. At this stage, the model establishes global relationships between different image regions, which makes it possible to detect hidden changes distributed across the multimedia object, including those produced by diffusion models or generative adversarial networks.

3.2. Transformer-XL encoder for system behavior analysis

Within the proposed technology, behavioral analysis is implemented by replacing earlier LSTM-based modules with the Transformer-XL architecture [27]. This transition makes it possible to process very long sequences of system events while reducing the limitations associated with gradient vanishing. Each event e_t from the API call vocabulary is converted into a numerical representation $x_t = Embed(e_t)$, after which the full event sequence is divided into segments of length D_b [27]. Context preservation between neighboring segments is ensured by the segment-level recurrence mechanism, where the hidden state $h_\tau$of the current segment is calculated as:

$$h_\tau = T_{Block}\left(X_\tau, \mathrm{SG}\left(h_{\tau\text{-}1}\right)\right) \tag{8}$$

The stop-gradient operator $SG()$ is used to maintain computational stability during segment recurrence and to prevent uncontrolled gradient propagation across long event sequences. This allows the model to capture subtle combinations of API calls that may indicate delayed activation, including Heap Spraying [29] or multi-stage loading of malicious code. Due to deeper modeling of temporal dependencies between system operations, this approach improves the identification of suspicious behavioral patterns [30]. The final behavioral representation Z_{beh} is obtained by

aggregating the hidden states of all processed segments, forming a dynamic profile of the multimedia object for further cross-modal verification.

3.3. Adaptive gated cross-attention fusion mechanism

The general structure of the proposed hybrid multimodal architecture is presented in Figure 1.

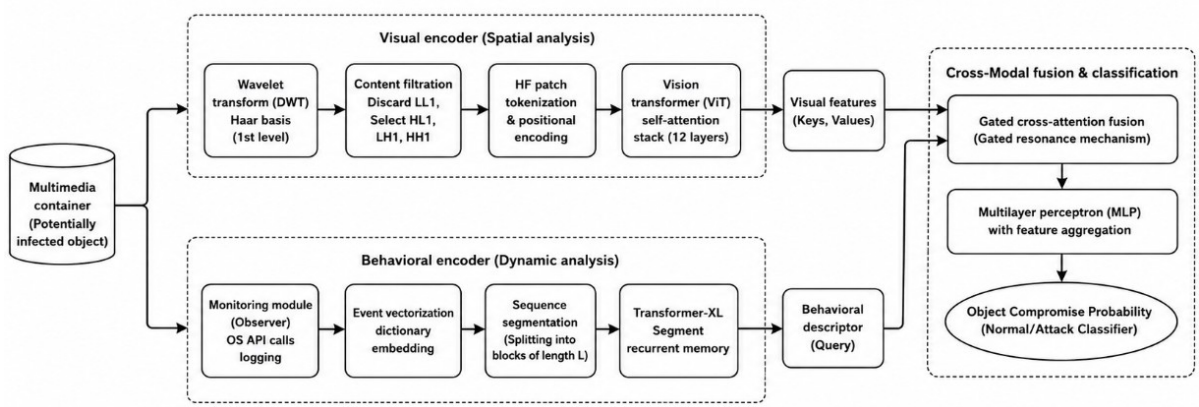

Figure 1: General conceptual scheme of the proposed hybrid multimodal architecture.

A key stage of multimodal feature integration is the use of the Cross-Attention layer, in which the behavioral descriptor Z_{beh} is interpreted as a query representation, while the visual features Z_{vis} serve as keys and values [31]. The cross-attention matrix A is computed as $A = \mathrm{softmax}\left(Q K^T / \sqrt{d_k}\right)$ in order to determine which image patches are most strongly associated with the current system activity. On the basis of this matrix, the contextual vector $C = AV$ is formed and then passed to a controlled fusion mechanism regulated by the scalar parameter λ. The contribution of each modality is adjusted through the activation coefficient $\alpha = \tanh(\lambda)$, which makes it possible to construct the final combined representation vector:

$$Z_{fused} = (1 - \alpha) Z_{vis} + \alpha C \tag{9}$$

This architecture provides adaptability to different categories of hidden cyber threats. When the analyzed multimedia object contains mainly static steganographic embedding without an active code component, the model increases the importance of the visual analysis branch $\alpha \to 0$. Conversely, when the object triggers complex activation behavior, including Stegosploit-type scenarios, the final decision is formed with greater reliance on behavioral resonance $\alpha \to 1$. In this way, the proposed mechanism supports the Zero Trust strategy by enabling continuous cross-validation of each multimedia object, regardless of its origin, declared format, or presumed level of trust.

3.4. Training criterion and multilevel optimization of the detection architecture

The training procedure of the proposed multimodal architecture is aimed at minimizing a modified objective function that combines binary cross-entropy with an L_2 regularization term used to stabilize the modality fusion parameter λ. Classification of the analyzed multimedia object is performed by a multilayer perceptron, which processes the fused representation and produces the predicted probability of threat presence. For a batch containing N samples, the final loss function is defined as follows:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log\left(\hat{y}_i\right) + \left(1 - y_i\right)\log\left(1 - \hat{y}_i\right)\right] + \beta\lambda^2 \tag{9}$$

The coefficient β is introduced to limit the excessive influence of a single modality during training and to encourage the model to identify meaningful relationships between different feature domains. As a result, the proposed information technology supports the detection of adaptive steganographic insertions in multimedia objects under near-real-time processing conditions. This creates an additional protection layer against concealed harmful functionality embedded in multimedia objects of web resources and increases the resilience of information infrastructure to AI-driven cyber threats.

4. Experimental results

To test the hypothesis that the proposed approach improves hidden threat detection accuracy, experimental simulation of the Hybrid ViT-Observer architecture was carried out. The modeling was implemented in a high-level data analysis environment using tensor computation libraries, which made it possible to form a latent space for the interaction between visual and behavioral features. The hardware configuration included NVIDIA H100 tensor-core accelerators, enabling the training of Transformer-XL modules on long sequences of system calls. The experimental sample was constructed from four representative data sources. The visual component was formed using BOSSBase 2.0 and ALASKA II, which contain steganographic objects with low embedding capacity measured in bpp. To reflect the characteristics of recent AI-driven threats, the CIFAKE dataset, containing objects generated by diffusion models and GANs, was also included. The behavioral component was based on system call traces from MalNet, which contains more than 1.2 million API call sequences typical of modern software implants and botnet activity. Each experimental instance was represented as a multimodal tuple X_v, X_b, where X_v denotes the spectrally decomposed image and X_b denotes the corresponding system activity log.

4.1. Parameter settings and training procedure of the proposed architecture

During the simulation, the main hyperparameters were selected separately for each functional module of the proposed information technology. The Wavelet-ViT block was configured to process image patches of 16×16 pixels, which corresponded to the input vector dimensionality $x_p \in R^{768}$ after aggregation of the wavelet components $\{LH, HL, HH\}$. The Self-Attention stack consisted of 12 layers with 8 attention heads in each layer, providing a global receptive field for detecting nonlinear relationships between texture anomalies. The Behavioral Transformer-XL module was configured with a segment memory length of 1024 calls, which allowed the model to preserve the context of previous operations without significant gradient loss. The training procedure was organized according to a Self-Supervised Learning strategy and used the AdamW optimizer with $learning_{rate} = 10^{-4}$ and $weight_{decay} = 0.01$. To reduce the risk of overfitting, Label Smoothing and Stochastic Depth techniques were applied. Special attention was given to the optimization of the cross-modal fusion parameter λ, whose initial value was set to 0, while its training dynamics were regulated by the coefficient $\beta = 0.05$.

4.2. Comparative analysis of experimental results and hypothesis validation

The obtained simulation results confirmed that the proposed multimodal technology outperforms existing unimodal CNN-based detectors. The combination of visual attention mechanisms with behavioral resonance enabled anomaly classification accuracy of 97.4%. A particularly important aspect is the model's robustness under conditions of extremely low embedding capacity. At an

embedding rate of 0.01 bpp, where classical statistical methods and SRNet-class architectures usually approach random-level performance 52-58%, the proposed technology preserved detection capability at 84.2%.

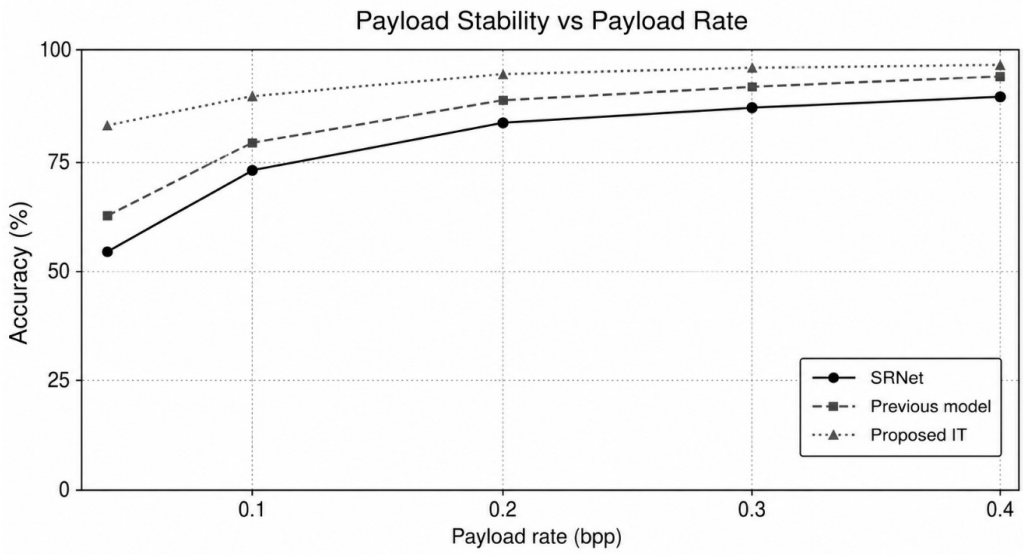

Figure 2: Robustness to Payload Graph (Accuracy vs bpp).

The obtained improvement is mainly associated with the Cross-Attention mechanism, which increases the importance of weak spectral-visual indicators when they are confirmed by suspicious behavioral events. In particular, low-amplitude deviations in the visual representation become more significant when they coincide with atypical operations related to process memory access or other abnormal system activity. This demonstrates that the joint analysis of visual and behavioral modalities provides more reliable detection than the isolated processing of each feature domain.

Table 1
Comparative effectiveness of detection models (simulation results)

Model	Accuracy (%)	Recall (%)	Precision (%)	FPR (%)
SRNet (SOTA CNN)	89,4	86,2	88,7	4,2
XuNet (SOTA CNN)	87,1	84,5	86,3	5,8
Previous Author Model (LSTM+CNN)	92,8	90,4	91,5	2,6
Proposed Hybrid ViT-Observer	97,4	96,1	97,2	0,8

To provide additional interpretation of the obtained decisions, attention maps were generated during the modeling process. These maps made it possible to identify the regions of a multimedia object that had the greatest influence on the final classification result. In Stegosploit-type scenarios, the model focused not only on semantic inconsistencies in the visual representation, but also on anomalies in the service structures of the object, while simultaneously detecting API calls associated with dynamic library loading. The results also confirmed the importance of Haar wavelet decomposition at the tokenization stage, since this procedure reduces the influence of the main visual objects and shifts the model's attention to high-frequency residual components, where traces of anomalous embedding may be present. As a result, the proposed architecture achieved a false positive rate of 0.8%, which is important for practical deployment in protection systems for high-load web resources and API gateways operating within the Zero Trust strategy.

5. Conclusion

The article presents an information technology designed to improve the detection of concealed cyber threats in multimedia containers, including adaptive generative steganographic insertions and multiformal object structures. To address the limitations of local convolutional neural networks (CNN) and recurrent architectures (LSTM), a hybrid neural architecture is proposed and substantiated. It combines a Wavelet-ViT spectral-visual analyzer with a Transformer-XL behavioral encoder. The main contribution of the proposed approach is the use of the Gated Cross-Attention Fusion mechanism, which enables the system to establish correlations between microstructural anomalies in high-frequency Haar wavelet coefficients and dynamic patterns of API calls observed in the operating environment. Experimental modeling using aggregated datasets, including BOSSBase 2.0, ALASKA II, CIFAKE, and MalNet, confirmed the detection capability of the proposed technology. The hybrid model achieved a classification accuracy of 97.4% with a false positive rate of 0.8%. In addition, the technology demonstrated robustness under conditions of extremely low embedding capacity, maintaining an accuracy of 84.2% at a payload rate of 0.01 bpp.

Declaration on Generative AI

During the preparation of this work, the author(s) used Grammarly in order to check grammar, spelling, and improve the linguistic flow and style of the text. Additionally, the author(s) used Gemini to assist with the correct formatting and syntax of mathematical notations. After using these tools, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

References

[1] O. Savenko, S. Lysenko, A. Kryschuk, Multi-agent based approach of botnet detection in computer systems. In: A. Kwiecień, P. Gaj, P. Stera (eds.), Communications in Computer and Information Science, vol. 291, Springer, Heidelberg, 2012, pp. 171–180. https://doi.org/10.1007/978-3-642-31217-5_19.

[2] C. Xiao, S. Peng, L. Zhang, J. Wang, D. Ding, J. Zhang, A transformer-based adversarial network framework for steganography, Expert Systems with Applications 269 (2025) 126391. doi:10.1016/j.eswa.2025.126391.

[3] A. A. Alquwayzani, A. A. Albuali, A systematic literature review of zero trust architecture for military UAV security systems, IEEE Access 12 (2024) 176033–176056. doi:10.1109/ACCESS.2024.3503587.

[4] D. Denysiuk, O. Savenko, S. Lysenko, B. Savenko, A. Kashtalian, Method for detecting steganographic changes in images using machine learning, in: Proceedings of the 2023 13th International Conference on Dependable Systems, Services and Technologies (DESSERT), IEEE, Athens, Greece, 2023, pp. 1–6. doi:10.1109/DESSERT61349.2023.10416453.

[5] M. L. Gambo, A. Almulhem, Zero trust architecture: a systematic literature review, Journal of Network and Systems Management 34 (2026) 25. doi:10.1007/s10922-025-09998-x.

[6] G. Luo, P. Wei, S. Zhu, X. Zhang, Z. Qian, S. Li, Image steganalysis with convolutional vision transformer, in: Proceedings of ICASSP 2022, IEEE, 2022, pp. 3089–3093. doi:10.1109/ICASSP43922.2022.9747091.

[7] S. Rose, O. Borchert, S. Mitchell, S. Connelly, Zero Trust Architecture, NIST Special Publication 800-207, National Institute of Standards and Technology, 2020. doi:10.6028/NIST.SP.800-207.

[8] N. J. De La Croix, T. Ahmad, H. Fengling, Comprehensive survey on image steganalysis using deep learning, Array 22 (2024) 100353. doi:10.1016/j.array.2024.100353.

[9] M. Saleem, M. R. Warsi, S. Islam, Secure information processing for multimedia forensics using zero-trust security model for large-scale data analytics in SaaS cloud computing environment, Journal of Information Security and Applications 72 (2023) 103389. doi:10.1016/j.jisa.2022.103389.
[10] A. Brown, M. Gupta, M. Abdelsalam, Automated machine learning for deep learning-based malware detection, Computers & Security 137 (2024) 103582. doi:10.1016/j.cose.2023.103582.
[11] M. Boroumand, M. Chen, J. Fridrich, Deep residual network for steganalysis of digital images, IEEE Transactions on Information Forensics and Security 14 (5) (2019) 1181–1193. doi:10.1109/TIFS.2018.2871749.
[12] K. Gao, C. C. Chang, J. H. Horng, I. Echizen, Steganographic secret sharing via AI-generated photorealistic images, EURASIP Journal on Wireless Communications and Networking 2022 (2022) 119.
[13] M. A. Bravo-Ortiz, E. Mercado-Ruiz, J. P. Villa-Pulgarin, et al., CvTStego-Net: a convolutional vision transformer architecture for spatial image steganalysis, Journal of Information Security and Applications 81 (2024) 103695. doi:10.1016/j.jisa.2023.103695.
[14] S. Huang, M. Zhang, Y. Ke, X. Bi, Y. Kong, Image steganalysis based on attention augmented convolution, Multimedia Tools and Applications 81 (2022) 19471–19490.
[15] R. Tabares-Soto, H. B. Arteaga-Arteaga, et al., Sensitivity of deep learning applied to spatial image steganalysis, PeerJ Computer Science 7 (2021) e616. doi:10.7717/peerj-cs.616.
[16] D. Denysiuk, O. Savenko, M. Kvassay, Method for detecting malicious commands transmitted via images using steganography, CEUR Workshop Proceedings 3963 (2025) 340–350.
[17] Z. Zhou, K. Chen, D. Hu, et al., Global texture sensitive convolutional transformer for medical image steganalysis, Multimedia Systems 30 (2024) 155. doi:10.1007/s00530-024-01344-6.
[18] T. Yao, Y. Pan, Y. Li, C. W. Ngo, T. Mei, Wave-ViT: unifying wavelet and transformers for visual representation learning, in: Computer Vision – ECCV 2022, LNCS 13685, Springer, 2022, pp. 328–345.
[19] S. Lysenko, O. Savenko, K. Bobrovnikova, A. Kryshchuk, Self-adaptive system for the corporate area network resilience in the presence of botnet cyberattacks, Communications in Computer and Information Science 860 (2018) 385–401.
[20] K. K. Wei, W. Q. Luo, S. Q. Tan, J. W. Huang, CTNet: a convolutional transformer network for color image steganalysis, Journal of Computer Science and Technology 40 (2025) 413–427. doi:10.1007/s11390-023-3006-3.
[21] G. O. Ganfure, C. F. Wu, Y. H. Chang, W. K. Shih, Deepware: imaging performance counters with deep learning to detect ransomware, IEEE Transactions on Computers 72 (2023) 600–613. doi:10.1109/TC.2022.3173149.
[22] R. Cogranne, É. Giboulot, P. Bas, ALASKA#2: challenging academic research on steganalysis with realistic images, in: 2020 IEEE International Workshop on Information Forensics and Security (WIFS), IEEE, 2020, pp. 1–6.
[23] Y. Ding, X. Zhang, J. Hu, W. Xu, Android malware detection method based on bytecode image, Journal of Ambient Intelligence and Humanized Computing 14 (2023) 6401–6410. doi:10.1007/s12652-020-02196-4.
[24] O. Pomorova, O. Savenko, S. Lysenko, A. Kryshchuk, K. Bobrovnikova, Anti-evasion technique for botnet detection based on passive DNS monitoring and active DNS probing, Communications in Computer and Information Science 608 (2016) 83–95.
[25] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, R. Salakhutdinov, Transformer-XL: attentive language models beyond a fixed-length context, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, ACL, 2019, pp. 2978–2988. doi:10.18653/v1/P19-1285.
[26] X. Zhang, J. Liu, T. Long, et al., A code completion approach combining pointer network and Transformer-XL network, Applied Intelligence 55 (2025) 451. doi:10.1007/s10489-025-06315-6.

[27] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: transformers for image recognition at scale, in: Proceedings of ICLR 2021, 2021.
[28] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, L. Zhang, CvT: introducing convolutions to vision transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 22–31.
[29] H. Tan, M. Bansal, LXMERT: learning cross-modality encoder representations from transformers, in: Proceedings of EMNLP-IJCNLP 2019, 2019, pp. 5100–5111. doi:10.18653/v1/D19-1514.
[30] H. Kheddar, M. Hemis, Y. Himeur, D. Megías, A. Amira, Deep learning for steganalysis of diverse data types: a review of methods, taxonomy, challenges and future directions, Neurocomputing 600 (2024) 127528. doi:10.1016/j.neucom.2024.127528.
[31] M. Shelest, Y. Pidlisnyi, M. Kapustian, Methods of hiding data in computer networks: from classics to IoT and AI, Computer Systems and Information Technologies 4 (2025) 27–34. doi:10.31891/csit-2025-4-3.