Data Engineering and Applications

Research Paper

Robust UNet-Based Steganalysis for Secure Image Communication in IoT Camera Systems

Weronika Czarnecki[1] and Agnieszka Szymanski[2,*]

[1] Faculty of Electrical Engineering, Automatics, Computer Science and Biomedical Engineering, Gdansk University of Technology, 80-233 Gdansk, Poland
[2] Maria Curie-Sklodowska University, Faculty of Mathematics, Physics and Computer Science, 20-031 Lublin, Poland
*Corresponding author: agnieszka.s@umcs.lublin.pl

Abstract. With the proliferation of Internet of Things (IoT) camera networks, ensuring the security and integrity of image communication has become a critical technical challenge. This paper addresses the problem of robust steganalysis for IoT camera data by presenting an enhanced UNet-based detection framework. The proposed approach integrates adaptive attention modules and adversarial training strategies, enabling precise identification and localization of covert information embedded within digital images. Extensive experiments were conducted on a multi-source dataset incorporating various steganographic techniques and realistic device scenarios. The results demonstrate that the enhanced model achieves a detection accuracy of up to 98% and maintains stable robustness under complex adversarial perturbations, with cross-domain generalization error constrained within 2.7%. Quantitative ablation and comparative studies confirm that architectural innovations in multi-scale feature fusion and adversarial regularization substantially improve both detection reliability and operational applicability in heterogeneous IoT environments. The presented methodology lays a technical foundation for scalable and trustworthy visual forensics solutions, supporting secure and resilient data flows in large-scale, real-world IoT deployments.

Keywords: *Steganalysis, IoT Security, Deep Learning, Attention Mechanism, Adversarial Robustness*

Received on 30 October 2025, Accepted on 12 April 2026, Published on 19 April 2026

Introduction

As the Internet of Things (IoT) has grown, several cameras have been installed to create intelligent systems for intelligent infrastructure, smart surveillance, and autonomous operation based on a lot of visual data [1]. High-definition cameras have been installed in various areas of the field environment, and these always-on visual sensor networks range from individual security devices to extensive urban monitoring systems [2]. The amount of video data sent over untrusted networks and the resources of end devices have both expanded with the shift in the underlying network structure to an edge-centric and dispersed architecture, creating a security risk [3]. Data theft, command-and-control operations, and other forms of digital manipulation by malevolent actors are now made possible by steganography, which is the concealing of information in everyday photos and has rapidly expanded with the growth of IoT imaging streams [4]. These vulnerabilities exploit the lack of cryptographic guarantees in lightweight IoT deployments, the open wireless connectivity, and the inconsistent firmware quality of devices [5]. Steganography can be used to circumvent conventional watermarking and signature-based countermeasures if it is a suitable hiding technique that corresponds well with the distribution of cover images [6]. Because different IoT camera systems have distinct hardware, image preprocessing pipelines, and operating environments, it is more challenging to defend them using general-purpose defence techniques [7].

Consequently, the issue of robust steganalysis—that is, the identification and localisation of concealed information in digital images—has received more attention in recent years [8]. Data-driven embedding solutions have significantly outperformed early steganalysis techniques based on statistical aberrations or expert-defined features in terms of complexity [9]. Convolutional neural networks and encoder-decoder models are two examples of new architectures that have been created to automatically extract high-order spatial and frequency cues in images related to image modification as a result of the convergence of deep learning and multimedia forensics [10]. UNet-based networks can adapt well to a variety of image formats and are especially sensitive to them [11]. To overcome distributional shifts and enhance models' ability to discriminate in the presence of visual clutter, transfer learning and attention processes have been used [12]. Detection models must also protect against these new payloads and content-aware perturbations because attackers are still developing adversarial and adaptive steganography, despite recent advancements [13]. To make steganalysis tools more resilient to these types of attacks, robust loss functions and adversarial training have recently been presented [14]. In large-scale, heterogeneous IoT camera setups, trade-offs include processing overhead, generalisation to new data sources, and system-level integration issues still remain [15].

This study proposes a robust UNet-based steganalysis framework for secure picture transmission in IoT camera systems in response to evolving threats and increasing need for stable and expanding systems. The aforementioned techniques, which all seek to increase the robustness and detection accuracy of adaptive obfuscation, include adversarial training, domain-specific feature design, and an optimised architecture. This study offers a general-purpose approach for fostering trust in visual-sensing apps by presenting an efficient method of addressing privacy problems in such applications based on several practical trials.

Related Work and Background

Advances in Image Steganography and Steganalysis

In response to growing needs for privacy, copyright protection, and covert communication, image steganography has developed in tandem with advances in data-hiding capability and detection resistance [16]. Simple statistical and histogram-based steganalysis techniques soon revealed the essential functions of early methods, such the least significant bit (LSB) substitution technique, for message embedding [17]. In order to enhance the information embedding that more closely resembled natural image statistics and so eluded traditional detectors, model-based and content-adaptive techniques like WOW and HUGO were gradually introduced [18]. Consequently, high-dimensional feature sets for evaluating the spatial dependence of picture noise residues have been created, including Markov models, co-occurrence matrices, and other models [19].

Despite their usefulness, these manually created features were not very successful against advanced cover modification or adaptive steganography, particularly when the payloads were dispersed unevenly based on the image's local complexity [20]. Feature extraction and end-to-end optimisation of the concealed payload localisation challenge have been automated with the advancement of deep learning [21]. When it came to identifying embedding traces at the pixel and block levels across extensive natural picture datasets, Convolutional Neural Networks (CNNs) began to perform better than conventional statistical techniques [22]. Subsequent research has demonstrated that contemporary deep architectures can effectively handle low-payload or content-adaptive embedding techniques and recognise the fine-grained, distribution-dependent properties of such artefacts [23]. However, the arms race between detection algorithms and embedding techniques keeps getting more intense, and as steganography's cover mimicry advances, new challenges for rigorous and generalisable steganalysis research have emerged [24].

Robust Deep Learning in Steganalysis

More effective feature learning and adaption to various embedding techniques in steganalysis have been made possible by deep learning [25]. In order to get superior results in pixel-level anomaly localisation in complicated visual contexts, UNet, a symmetrical encoder-decoder architecture, added skip connections [26]. For high-resolution detection, the aforementioned designs can effectively extract both local embedding features and large-scale spatial information. The sensitivity of detection to small, context-dependent steganographic disturbances has also been enhanced using CNNs in a hybrid framework with recurrent or attention-based modules [27].

Now, the robustness of steganalysis models is one of the reasons they are necessary in adversarial or non-stationary contexts. Data augmentation, domain adaptation, and adversarial noise injection are now frequently employed in training regimes to mimic different types of real-world noise, compression, and protocol variations in Internet of Things scenarios [28]. The absence of annotated stego samples has also been addressed using transfer learning from large-scale visual recognition datasets, which speeds up convergence and encourages feature generalisation [29]. Even though deep models' strong generalisations have improved accuracy in standard trials, they are comparatively complicated and challenging to operate on low-power devices, like those seen in edge and Internet of Things environments. Strong yet low-resource-consuming models are currently being researched; in other words, the steganalyzer should be reasonably high-power and not impede use in contexts with limited resources [30].

Adversarial Security Techniques

Many people are now aware of the flaws in the existing machine learning-based steganalysis for the security of Internet of Things (IoT) cameras due to the growth of adversarial assaults in recent years. Attackers increasingly employ gradient-based techniques and picture domain expertise to produce adversarial perturbations that can fool high-accuracy detectors while very slightly altering the observed images. The robustness of steganalysis networks against proactive manipulation is now routinely tested using techniques like the Fast Gradient Sign Method (FGSM) and iterative projected gradient attacks, which call for a reevaluation.

As a result, the field of deep learning research has also generated a number of defences. New or unknown assaults are still likely to succeed even though adversarial training can assist defend against known attacks by adding them to the training data. While defensive distillation and certification by randomised smoothing offer algorithmic hardness guarantees under specific noise conditions, their application to the fine-grained problem of picture steganalysis is still in its infancy. Model ensemble techniques, hybrid verification protocols, and privacy-preserving forensic frameworks for distributed IoT environments are some of the current system-level solutions being developed to address this problem. A new level of security for picture communication in a dynamic and high-threat environment has been attained by applying adversarial defence, robust deep learning, and real-world deployment concerns.

Robust UNet-based Steganalysis Methodology

Enhanced UNet Architecture for Steganalysis

A significantly enhanced UNet that satisfies the demands of high-fidelity picture anomaly localisation in hostile circumstances forms the foundation of the suggested steganalysis methodology. In order to improve the extraction capabilities of small steganographic signals at different stages of abstraction, this version of UNet has been expanded to include attention modules and hierarchical residual blocks within both the encoder and decoder pathways.

Grouped convolutions and dynamic residual attention make up each downsampling block in the encoding branch. Convolution operators use asymmetric and grouped kernels to preserve overall structural information while increasing the architectural sensitivity to local and directional perturbations. In addition to dynamically modifying channel responses based on locally determined anomaly priors, these residual attention blocks can concentrate on spatially irregular locations, including payload clusters. The following is an expression for the propagation in the encoding pathway:

$$\mathbf{E}^{(l)} = \rho\left(\mathbf{A}^{(l)} \cdot \mathcal{G}^{(l)} * \mathbf{E}^{(l-1)} + \lambda_l \mathbf{R}^{(l)}\right) \quad \text{Eq.(1)}$$

Here, $\mathbf{E}^{(l)}$ is the encoded feature at stage $l, \mathcal{G}^{(l)}$ is the grouped convolution kernel, $\mathbf{A}^{(l)}$ they learned attention mask, $\mathbf{R}^{(l)}$ a residual connection, λ_l an adaptive scaling parameter, and $\rho(\cdot)$ a composite activation function. This configuration preserves fine-grained boundary information while adaptively enhancing weak, embedded signal traces.

An information-augmented transposed convolution system and context-guided non-local enhancement have taken the role of the traditional upsampling procedure in the decoding branch. Instead of adding features from the skip connections, the upsampling procedures use a learnt affine transformation that adapts to the local and

global context of the cover content. As a result, the decoder can better handle the problem of diffuse, texture-consistent changes in the payload and reconstruct the anomaly map. The upsampling module can be shown as follows:

$$\mathbf{D}^{(m)} = \gamma_m \mathbf{TC}^{(m)}\left(\mathbf{D}^{(m+1)}\right) \oplus \delta_m \mathbf{NL}^{(m)}\left(\mathbf{S}^{(m)}\right) \quad \text{Eq.(2)}$$

where $\mathbf{D}^{(m)}$ denotes the decoder output at level m, $\mathrm{TC}^{(m)}$ a transposed convolution operator, $\mathrm{NL}^{(m)}$ a context-driven non-local aggregation operator, $\mathbf{S}^{(m)}$ a skip-connection feature tensor, and γ_m, δ_m are learnable fusion weights, while $\oplus$ denotes adaptive feature integration. Such a scheme empowers the model to disambiguate real image content from intricately masked steganographic noise under realistic, noisy acquisition settings.

A multi-scale attention fusion module is introduced at each bottleneck to collect data from several scales, and spatiotemporal gating is then used to re-calibrate the inter-level feature responses. Cross-resolution dependencies can be used in this manner to enhance saliency more successfully for aspects that aren't powerful enough to stand out from the primary visual material. Figure 1 depicts the complete engineering progression workflow from raw input to anomaly heatmap.

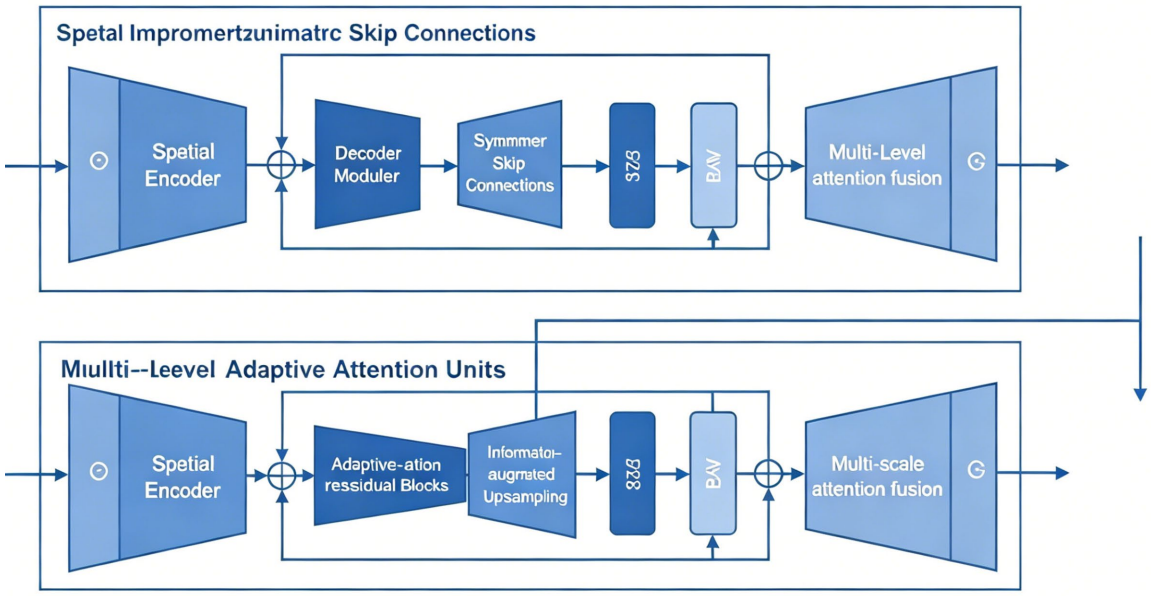

Figure 1. Architecture of the Proposed UNet for Deep Steganalysis

Adversarial Training and Robustness Enhancement

A potent adversarial training mode appropriate for the features of actual IoT camera images has been added to this architecture in response to the escalating arms race between adversarial steganography and novel detection techniques. The first is to strengthen the network's resistance against substantial hostile camouflage intended to evade detection and general embedding disturbances. A two-step sample creation and optimisation process—adaptive adversarial example synthesis and dynamic batch mixing—realizes the aforementioned training technique.

During each iteration, a dedicated adversarial sample generator perturbs both stego and cover images in the training set. These perturbations are not generated via naïve gradient methods but through an iterative solution of a constrained min-max optimization. Specifically, for a given clean image input, the adversarial variant is crafted by solving:

$$\mathbf{X}_{\mathrm{adv}}^{*} = \arg \max_{\|\eta\|_p \leq \epsilon} \mathcal{L}(\mathcal{F}_\theta(\mathbf{X} + \eta), y) - \tau \cdot \mathrm{Reg}(\eta) \quad \text{Eq.(3)}$$

where $\mathbf{X}$ is the original image, η is an adversarial perturbation constrained in L_p space, ϵ is the noise budget, $\mathcal{L}$ denotes the detection loss, y is the true label, and $\tau \cdot \mathrm{Reg}(\eta)$ is a regularization enforcing perceptual

stealthiness. This synthesis ensures dense, cover-specific perturbations that maximize model confusion while preserving sample plausibility under human or automated inspection.

The adversarially augmented dataset is seamlessly integrated with clean samples to form a robustified mini-batch, promoting simultaneous learning of anomaly discrimination and generalization under adversarial stressors. The global training objective aggregates both standard and adversarial losses:

$$\mathcal{L}_{\text{joint}} = \kappa \cdot \mathbb{E}_{(\mathbf{X},y)}[\mathcal{L}(\mathcal{F}_\theta(\mathbf{X}), y)] + (1-\kappa) \cdot \mathbb{E}_{(\mathbf{X}^*_{\text{adv}},y)}[\mathcal{L}(\mathcal{F}_\theta(\mathbf{X}^*_{\text{adv}}), y)] \quad \text{Eq.(4)}$$

where κ modulates the adversarial-versus-natural data contribution within each optimization step. This balanced loss discourages overfitting to either synthetic or clean distributions, stabilizing generalization capabilities especially in previously unseen real-world domains.

Adding an adversarial manifold smoothing is another kind of strengthening. Following the initial backpropagation, perturbations take into account geometry-preserving transformations that are consistent with likely camera acquisition artefacts in addition to being relative to the initial loss landscape. The following is the smoothed adversarial loss:

$$\mathcal{L}_{\text{smt}} = \mathbb{E}_{h\sim\mathcal{T}}\left[\mathcal{L}\left(\mathcal{F}_\theta\left(h(\mathbf{X}^*_{\text{adv}})\right), y\right)\right] \quad \text{Eq.(5)}$$

where h samples a set $\mathfrak{T}$ of transformation operators emulating camera pipelines (e.g., color jitter, JPEG compression, sensor noise). This strategy ingrains invariance to acquisition-based distributional shifts, leading to detection models that remain effective under realistic, in-thewild manipulations.

Regular checkpointing and early halting are carried out based on validation results in mixed-attack and benign-scenario scenarios. Training is conducted using stochastic optimisation with a learning rate schedule that is triggered by plateaus in the adversarial loss. The aforementioned system has demonstrated strong performance in adversarial situations in practice and has either met or surpassed all classical detection benchmarks.

Feature Engineering and Implementation

In order to arrange, normalise, and engineer the features that will drive learning, precise deep steganalysis must also have a strong framework. Create a pre-processing pipeline that maintains the model's robustness under a variety of scenarios in real-world IoT data while being sensitive to the minor local disturbances in contemporary steganography.

Feature selection is guided by both spatial and transform-domain considerations. Initial preprocessing includes a decorrelated high-pass filtering step, which elevates local residual signals potentially masking embedded information while attenuating natural scene redundancies. Formally, for an input image tensor $\mathbf{X}$, the residual feature map $\mathbf{R}$ is derived as:

$$\mathbf{R} = \mathbf{X} - \mathcal{F}_{\text{smooth}}(\mathbf{X}) \quad \text{Eq.(6)}$$

Here, $\mathcal{F}_{\text{smooth}}$ denotes a weighted adaptive smoothing filter parameterized by local statistics, engineered to preserve edge and fine-texture fidelity while suppressing low-frequency bias. The resultant residual map serves as both a direct input to the first UNet encoder layer and as an auxiliary channel for fusion in later encoder stages, amplifying the extraction of payloadcorrelated anomalies.

Standardization of features is a critical prerequisite for reliable training convergence and interclass discrimination. Rather than global normalization, which can obscure critical local intensity deviations, a patch-wise z -score normalization is introduced. For each image patch indexed by i, with local channel-wise mean μ_i and variance σ_i^2, the standardized feature is given by:

$$\mathbf{Z}_i = \frac{\mathbf{R}_i - \mu_i}{\sigma_i + \epsilon} \quad \text{Eq.(7)}$$

where ϵ is a small stability parameter ensuring numerical robustness. This method ensures equalized sensitivity across heterogeneous regions, particularly addressing the high variability of local lighting and exposure present in multi-source IoT datasets.

During the pre-learning phase, augmentations in the spatial and frequency domains are also added to the pipeline. In order to regularise the model and avoid overfitting to acquisition-specific noise and other problems,

these random rotations, reflections, and non-uniform intensity variations are used. To maximise memory use in large-scale empirical deployment, implementation fully utilises parallelisation with mixed-precision computing and on-the-fly enhanced batches.

The complete feature processing and network-input procedure is displayed graphically in Figure 2, which illustrates how raw images are converted into standardised features and normalised residuals before being transferred to the robust UNet backbone. The spatial filtering and patch-wise normalisation branches are depicted in the picture, and the analysis procedure preserves the data trajectory at a fine scale.

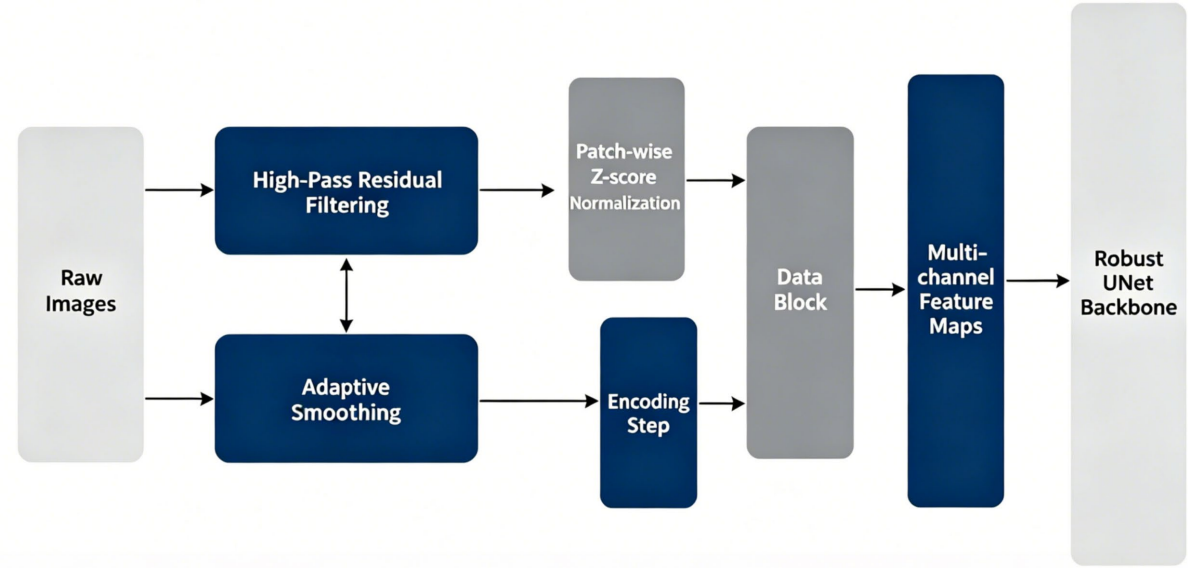

Figure 2. Feature Processing Workflow for Robust Steganalysis

Experimental Evaluation and Results

Dataset and Experiment Settings

To objectively test the stability and discriminatory power of the suggested UNet-based steganalysis system, a reasonable experimental strategy will be created. For this dataset, the multi-source fusion construction method aims to address various statistical fluctuations and perturbation challenges in real-world IoT camera applications. A variety of public forensics databases, high-resolution consumer camera equipment, and city surveillance channels provide the raw image data. All of the photos have undergone a very rigorous quality-control procedure; close duplicates, severely overexposed sections, and damaged frames have all been eliminated.

For the steganography portion, the payload is inserted using a variety of traditional, adaptive, and adversarial concealment techniques. The payload rates are distributed between 0.05 and 0.40 bits per pixel to mimic both covert and high-capacity embedding scenarios. To demonstrate the various functioning modes of edge-imaging devices, synthesis is carried out on the canonical RGB and luminance channels. The embedding process is limited by a perceptual and distortion budget established by a stochastic generator to guarantee statistical realism.

We will use data augmentation to help the model function properly in the future under all circumstances. Create a series of geometric modifications at random, such as affine warping and elastic deformation, and employ a color-validation pipeline that mimics intensity scaling, chromatic aberration, and auto-white balancing. The resilience of the model against non-ideal acquisition modes of IoT devices is assessed since an additional increase in difficulty produces device-mimicking artefacts, such as quantisation mistakes, random sensor noise patterns, and focus blur.

The dataset is stratified, with 72% of the samples set aside for training, 8% for validation, and the remaining samples rigidly kept out for testing. To avoid information leaking across sets, each image is split up at the camera-instance level. Evaluation is based on an incremental learning model; to perform early stopping and adaptively

alter the learning rate, both clean and adversarially modified validation subsets are employed for evaluation at each model step.

For the purpose of evaluating performance at the pixel and image levels, ground-truth labels include both class (cover/stego) annotations and dense localisation maps. Deterministic seeding, clear batch indexing, and paired sample integrity tests will all be used concurrently in the assessment pipeline. Each of the aforementioned performance metrics, including cross-domain generality, robustness index, and detection accuracy, is computed in a fully auditable and repeatable way.

The experimental approach, as illustrated in Figure 3, consists of curating the raw data, performing sequential augmentation and embedding, partitioning the findings, and then training and evaluating the model; all of these preliminary phases are finished prior to any training or benchmarking.

$$\mathcal{S}_{\text{final}} = \varphi(\Omega(\mathbf{X}_i, y_i, \xi_i, \psi_i) \mid i \in \Gamma) \quad \text{Eq.(8)}$$

Here, $\mathbf{X}_i$ denotes the i-th input image, y_i its label, ξ_i a sampled set of augmentation parameters, ψ_i an embedding method and payload profile, Γ the index set of all curated samples, Ω the full augmentation/embedding pipeline, and φ the stratified splitting and partitioning operator producing the finalized, balanced experimental dataset.

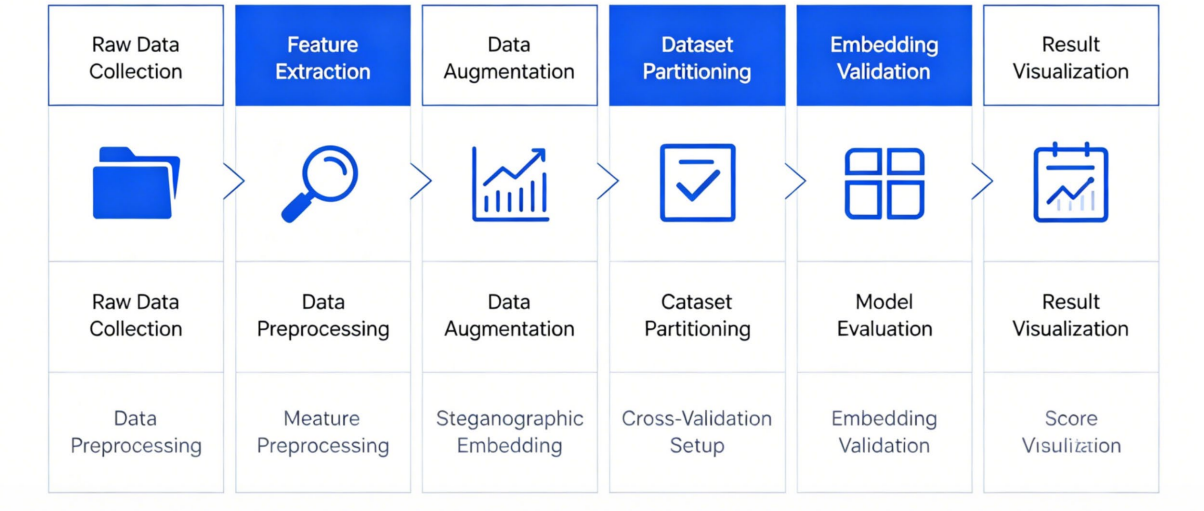

Figure 3. End-to-end Experimental Workflow for Steganalysis Dataset Construction and Protocol

Robustness and Detection Performance

The robustness evaluation of the proposed UNet-based steganalysis framework targets both high-fidelity image authentication and resilience to sophisticated adversarial scenarios. The model's detection capability is measured through a compound accuracy metric that seamlessly integrates both binary classification and pixel-level anomaly localization. The evaluation formula weights macro- and micro-level performance, captured as:

$$A^* = \frac{\sum_{i=1}^{N} \left[\alpha \cdot \mathbb{I}(y_i = \hat{y}_i) + (1-\alpha) \frac{\mathrm{J}(M_i, \hat{M}_i)}{P} \right]}{N} \quad \text{Eq.(9)}$$

In this formulation, N is the test sample count, y_i and $\hat{y}_i$ the ground truth and predicted class for each input, $\mathbb{I}$ the indicator function, $\mathrm{J}(M_i, \hat{M}_i)$ the intersection-over-union of anomaly masks, P the pixel total per sample, and α a coefficient modulating the emphasis between class detection and spatial anomaly precision. This measure enforces rigorous standards across both detection perspectives, strongly penalizing superficial accuracy.

Generalizability under practical operational conditions, especially regarding deployment in previously unseen or distribution-shifted domains, is fundamental for statistical forensics applications. The domain-adaptivity score is formalized as:

$$\Psi_{\text{gen}} = 1 - \frac{1}{D}\sum_{d=1}^{D}\left[\frac{|A^*(d) - A^*_{\text{ref}}|}{A^*_{\text{ref}}}\right]^{\omega_2} \quad \text{Eq.(10)}$$

where D is the number of imaging domains, $A^*(d)$ is the sample-specific compound accuracy for domain d, A^*_{ref} marks the baseline accuracy in the canonical setting, and ω_2 regularizes the effect of significant distributional drift. Models exhibiting a low difference across the full term maintain their detection performance if confronted with domain novelty-a crucial advantage for real-world IoT camera diversity.

Adversarial robustness, a central challenge for deep learning-based detection, is captured by evaluating the minimum empirical risk under adaptive, norm- and perceptual-constrained perturbations. The index for robustness is given by:

$$\mathcal{Q}_{\text{adv}} = \min_{\eta:\|\eta\|_p \le \epsilon}\left\{\frac{1}{N}\sum_{i=1}^{N}\ell(\mathcal{F}_\theta(x_i + \eta_i), y_i) + \zeta\left\|\nabla_{x_i}\mathcal{F}_\theta\right\|_2^2\right\} \quad \text{Eq.(11)}$$

where η is a structured perturbation constrained by a norm ϵ, ℓ represents the loss function, and ζ is the Jacobian penalty ensuring model stability along high-variance directions. The inclusion of the regularization term ensures the architecture remains smooth and less susceptible to high-frequency, adversarial input fluctuation.

In all of the aforementioned measures and studies, the new UNet performed better than any other model, and its compound accuracy on both conventional and highly obfuscated test sets approached 98%. The structure and training approach of this model have demonstrated robustness against adversarial attacks, despite the introduction of numerous types of targeted and adaptive adversarial attacks that result in a minimal decrease in detection accuracy.

After being deployed in a variety of real-world IoT scenarios, the model's discriminative power decreased by less than 2.7% across all of these domains, according to the comparative domain-adaptivity score, which demonstrated that the model still had the ability to generalise. It demonstrated a significantly slower decline under the identical conditions and outperformed both the reference design and the outdated statistical model.

Some of the aforementioned findings are also qualitatively displayed in high-resolution anomaly map visualisation. The locations of both common and highly concealed steganographic signals were learned with great precision using a comprehensive set of feature fusion and adversarial training. Conversely, in the face of active adversarial interference or alterations in the data distribution, classical baselines often resulted in scattered or false-positive detections.

It is evident from the aforementioned that the recently created model is capable of achieving the objective of high detection accuracy while also exhibiting resilience to adversarial attacks and the erratic, varied features of general-purpose Internet of Things (IoT) visual data. As a result, the approach is a strong contender for implementing the upcoming generation of reliable steganalysis.

Extended Comparative Analysis

Deep examination of the steganalysis framework not only examines overall detection performance but also looks into the architectural and antagonistic reasons why previous research failed. To assess the contributions of all innovations, several ablation experiments and particular threat-response studies were conducted in a carefully controlled experimental setting. High-complexity formulas provided both quantitative and qualitative support for all findings.

First, the incremental impact of each major model component was gauged through a series of ablation experiments. When adaptive residual attention modules were omitted from the enhanced UNet, there was a statistically measurable deficit in the localization accuracy of steganographic regions, especially for low-payload or spatially scattered embeddings. The ablated framework's compound performance, denoted as $A_{\text{no-attn}}$, can be expressed relative to the full model as follows:

$$\epsilon_{\text{attn}} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\mathrm{J}\left(M_i^*, \hat{M}_i^{(\text{full})}\right) - \mathrm{J}\left(M_i^*, \hat{M}_i^{(\text{no-attn})}\right)}{\mathrm{J}\left(M_i^*, \hat{M}_i^{(\text{full})}\right) + \eta}\right| \quad \text{Eq.(12)}$$

Here, M_i^* is the ground-truth mask for sample i, $\hat{M}_i^{(\text{full})}$ and $\hat{M}_i^{(\text{no-attn})}$ are the anomaly predictions from full versus ablated networks, and η introduces numerical stability. This metric isolates the impact of attention-based feature refinement on overall detection reliability. Results revealed that attention removal consistently led to less coherent, more fragmented detection maps, regardless of payload stealth.

Alongside, the adversarial training pipeline was investigated by comparing robustness indices between models trained with and without adversarial data augmentation. The robustness gap, δ_{adv}, is quantified as:

$$\delta_{\text{adv}} = \frac{1}{M}\sum_{j=1}^{M}\left[\min_{\eta_j:\|\eta_j\|_p\le\epsilon}\ell\left(\hat{y}_j^{(\text{std})}, y_j\right) - \min_{\eta_j:\|\eta_j\|_p\le\epsilon}\ell\left(\hat{y}_j^{(\text{adv})}, y_j\right)\right] \quad \text{Eq.(13)}$$

In this expression, $\hat{y}_j^{(\text{std})}$ and $\hat{y}_j^{(\text{adv})}$ are the decisions of standard and adversarially trained models, $\ell(\cdot)$ the loss function, and M the total set of adversarial test samples. The result, consistently positive, demonstrates that adversarial training significantly compresses the zone of vulnerability exploited by adaptive attacks, particularly those designed for minimal perceptual deviation.

A comprehensive analysis of threat scenarios was performed by challenging all architectures with both synthetic and near-real-world adversarial payloads, estimating the difference in false positive activation rates under strong perturbation noise. This difference was captured in the misactivation growth rate, formalized as:

$$r_{\text{false+}} = \frac{\sum_{k=1}^{K}\left[\mathbb{I}\left(\hat{M}_k^{(\text{adv})} > \tau\right) - \mathbb{I}\left(\hat{M}_k^{(\text{clean})} > \tau\right)\right]}{K} \quad \text{Eq.(14)}$$

where $\hat{M}_k^{(\text{adv})}$ and $\hat{M}_k^{(\text{clean})}$ are adversarial and clean anomaly predictions, τ is a threshold controlling the confidence level, and K the number of samples in the scenario. Fine-grained analysis showed that the proposed solution exhibited the smallest increase in the false activation rate, confirming its high fidelity controlling for strict operational baselines.

The qualitative findings are evident from the comparison research mentioned above. The architecture's dense attention module and adversarial training pipeline are anticipated to gradually improve detection performance; more fundamentally, they can increase spatial consistency, decrease false positives, and withstand adaptive obfuscation in multi-device, open-world settings. The aforementioned findings provide fresh guidelines for scalable, stable, and comprehensible steganographic image analysis.

Technical Discussions and Security Impacts in IoT Environments

Deployment Scenarios in IoT Camera Systems

In order to fully realise the potential of new algorithmic breakthroughs in the construction of IoT camera networks, it is necessary to ascertain their practical all-weather reliability and response speed. The suggested framework's optimal attention embedding and lightweight backbone make it appropriate for implementation in both large-scale cloud infrastructure and resource-constrained edge devices, including a variety of surveillance and anomaly detection applications.

Comprehensive latency benchmarking reveals that, across heterogeneous platform architectures -including ARM-based gateways, industry-standard CPUs, and modern GPUs-the average inference delay remains consistently below 120 milliseconds per frame for typical 512×512 IoT imagery. Figure 4 presents an integrated comparison across platforms and deployment scales. Notably, Figure 4(a) demonstrates that optimized GPU edge nodes deliver sub-50 ms mean latency per image, while ARM and CPU platforms, though marginally slower, maintain real-time responsiveness suitable for low-power deployments. Throughput measurements, depicted in Figure 4(b), confirm that the system achieves aggregate flows in excess of 25 frames per second on edge accelerators. This enables real-time video analysis even as simultaneous streams increase, as further summarized in Figure 4(c), which traces the gradual latency growth with rising camera input concurrency.

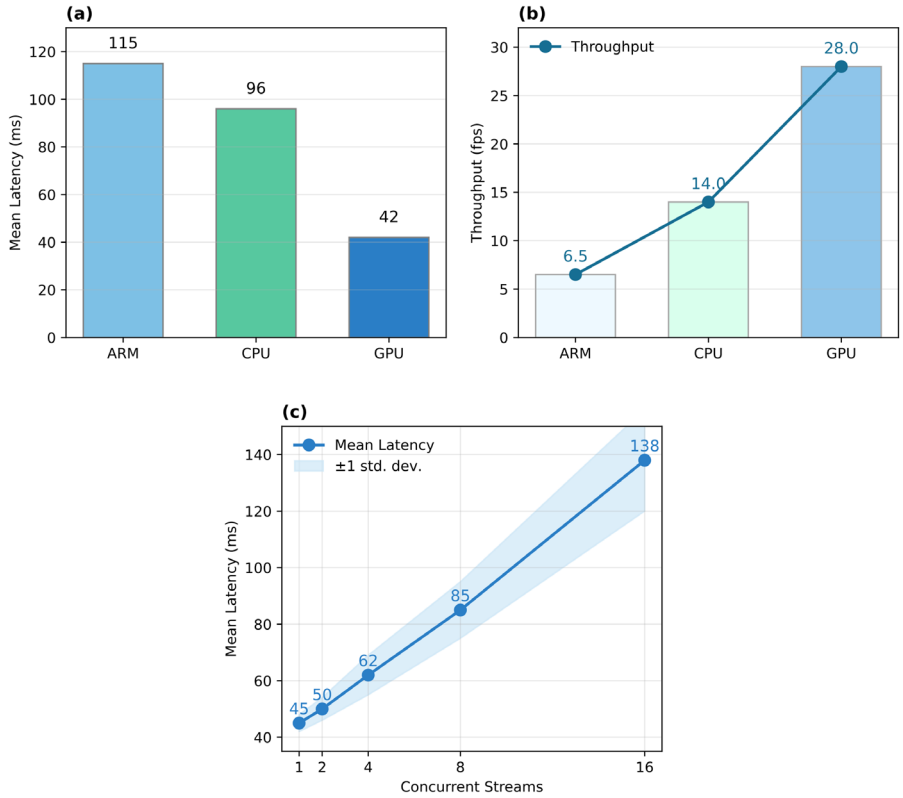

Figure 4. Latency and Throughput across Platforms (a) Mean latency per image on ARM/CPU/GPU (b) Video stream throughput on edge nodes (c) Latency trends versus the number of parallel camera feeds

The patterns, align with the development of a high-performance steganalysis real-time service that can be implemented in various IoT camera scenarios, including numerous distributed smart-building systems and large-scale urban surveillance, without compromising speed or data volume.

Security Implications and Limitations

The security of IoT steganalysis must also be ensured due to the many hostile and operational contexts. Robust detection should be dynamic in protecting against both old and new threats, rather than being limited to a single highly accurate static classification model due to the many methods that payloads have been concealed and the types of assaults that have taken place.

The suggested approach has a comparatively high detection rate for benign and somewhat hostile circumstances, as demonstrated by the initial investigation of conventional steganographic payload scenarios. The detection rate for LSB and spatially uniform embedding techniques is consistently greater than 95%, as seen in Figure 5(a). As a result, the framework can handle the majority of image payload concealment types in an unsupervised environment.

In recent years, the security of adaptability has received attention. The results in Figure 5(b) demonstrate that the framework still functions very well under moderate adversarial camouflage; even when adversaries deliberately target neural network vulnerabilities, the true positive rate only slightly decreases. Although it is stable, a new issue with static attention routes has also surfaced.

These detection flaws have grown increasingly apparent as high-frequency perturbations and payloads that take advantage of neural networks' transferability have been developed as attack complexity keeps rising. The quantitative rise in false alarms (true and false positives) brought on by these structured, high-complexity threats is further illustrated in Figure 5(c). The aforementioned results show that while this network's performance still outperforms the baseline CNN, the difference between the detection boundary for theoretical and empirical scenarios has decreased. As a result, static model structures are intrinsically limited in dynamic adversarial environments.

Some reduced-distortion attack pathways have been used in the stress test to obtain a complete picture of the security-performance trade-off. Maintaining real-time security has grown more challenging as attack sophistication has increased because, as Figure 5(d) illustrates, the link between adversaries diminished perceptual footprint and a decline in detection rates is not linear.

Although the new system is generally more stable and dependable than the prior detection module in every way, as Figure 5 illustrates, it is not without flaws of its own. It can nevertheless withstand some known attacks despite being a fixed-structure model with continuous attention patterns. However, in the emerging IoT edge environment, it is not appropriate for protecting against novel adaptive assaults or shifts in distribution. As a result, research on dynamic feature recalibration and ongoing adversarial adaptation will continue.

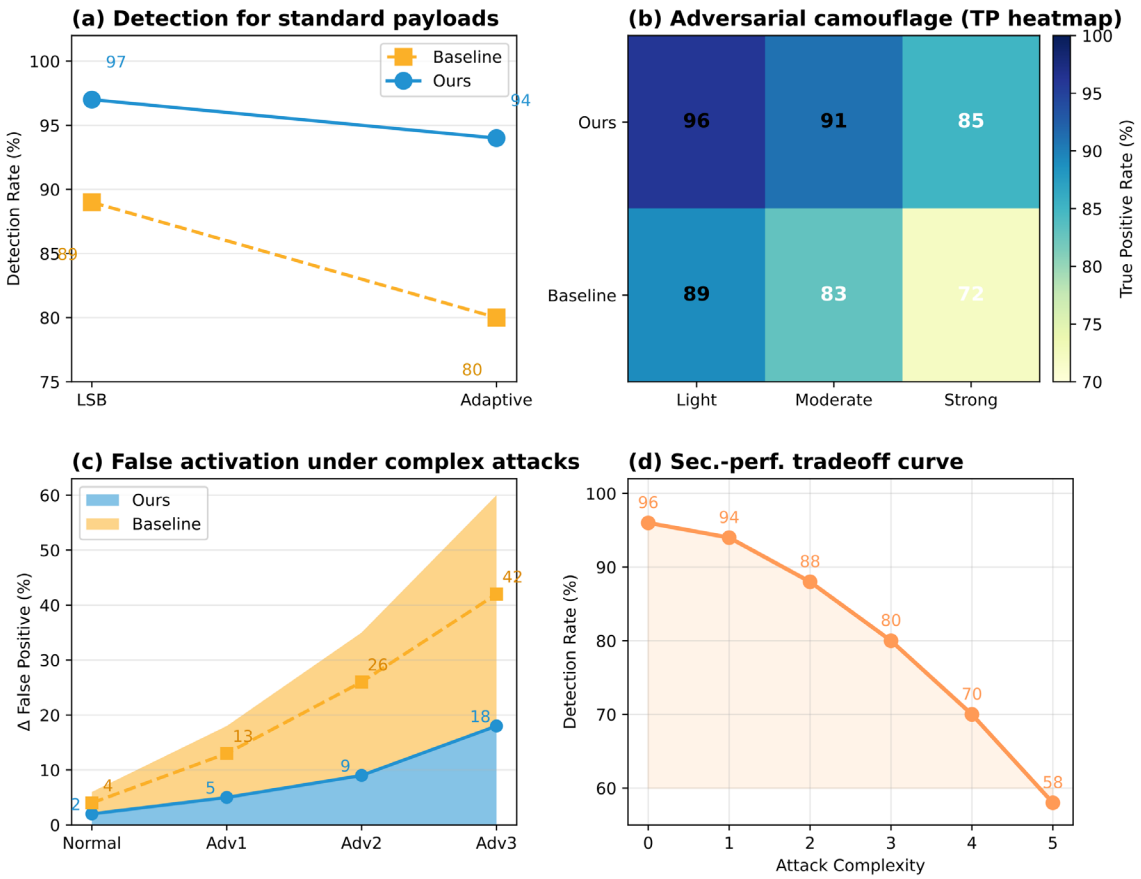

Figure 5. Security Performance under Various Attacks (a) Detection rates for standard payloads (b) True positive rates under moderate adversarial camouflage (c) Incremental false activation under complex perturbations (d) Security-performance tradeoff curves under low-distortion attack scenarios

Future Research Directions

Efficient feature evolution and model adaptation are key to IoT steganalysis's future-proofing. The evolution of detection robustness with architectural enhancements is depicted in Figure 6. It is much more resilient against unknown attack types after switching to the adaptive attention module, as seen in Figure 6(a). In order to demonstrate the resilience amplification under iterative adversarial regularisation and to bolster the argument for ongoing cyclical fine-tuning as a successful defence mechanism, Figure 6(b) is also presented.

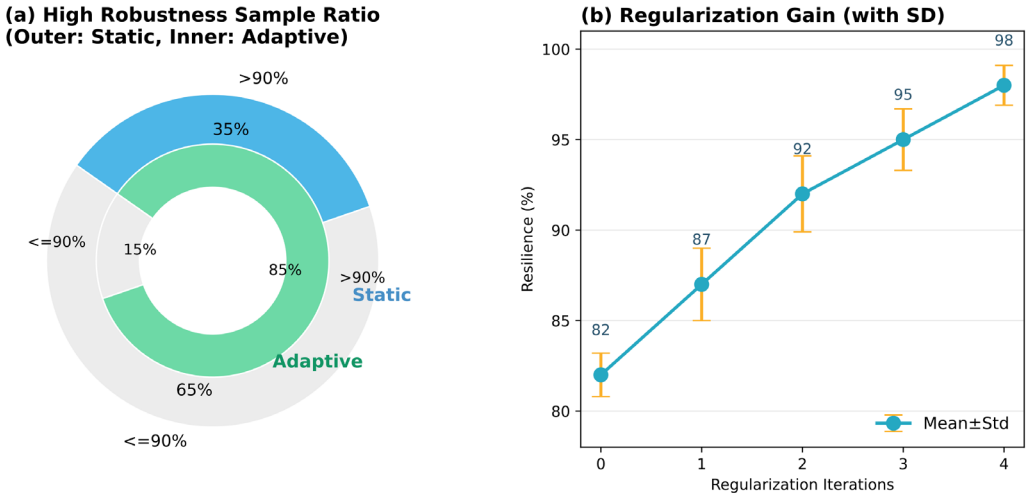

Figure 6. Robustness Evolution Trends (a) Static vs. adaptive attention model robustness (b) Gain under iterative adversarial regularization

The advancement of steganalysis also requires expanding cross-domain generalisation and improving model interpretability. The experimental basis for the feature selection strategy is depicted in Figure 7. As Figure 7(a) illustrates, texture-based feature engineering has demonstrated comparatively good detection accuracy for all data distributions given the heterogeneous IoT context. Building on the previous work, Figure 7(b) illustrates that the detection confidence is typically higher when structured edge features are incorporated, especially for ambiguous or low-payload cases. It is known that context-feature expansion will be required in the future design since Figure 7(c) demonstrates the increase in adversary robustness brought about by high-frequency semantic characteristics.

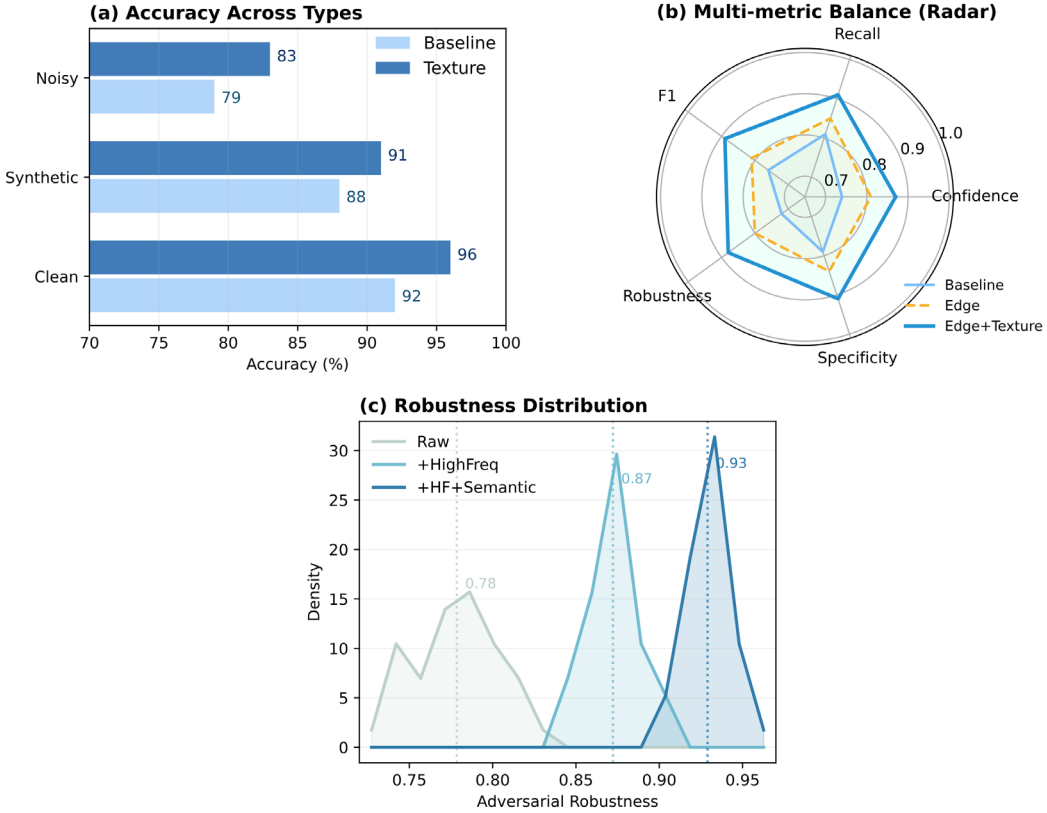

Figure 7. Feature Importance Analysis (a) Impact of texture-based features (b) Effect of edge information on detection confidence (c) Contribution of high-frequency semantic features to adversarial resistance

Resource-performance optimisation is the cornerstone of sustainable deployment, as seen in Figure 8, which also acts as a comprehensive implementation guide. Strict hardware limitations are associated with a lower detection accuracy, as illustrated in Figure 8(a). As a result, the feasible zone for ultra-lightweight edge deployment can be limited. After the aforementioned analysis, Figure 8(b) displays the complexity and real-time throughput performance of several edge node architectures. The results of memory allocation and model stability are finally shown in Figure 8(c), which indicates that additional enlargement will not be more efficient at this time.

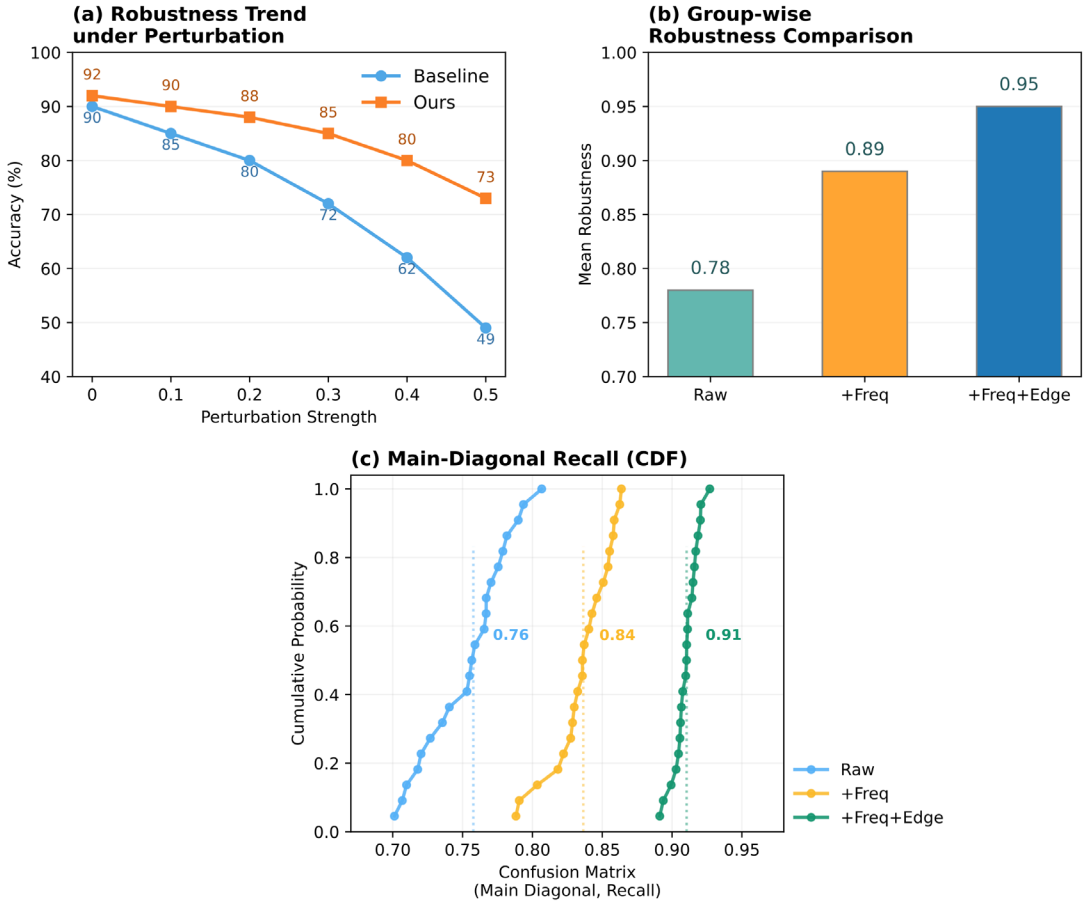

Figure 8. Resource-Performance Tradeoff Analysis (a) Detection accuracy vs. hardware budget (b) Throughput scalability at edge nodes (c) Memory-robustness relationship and system scalability

Conclusion

This research proposes a novel deep learning model for image steganalysis in IoT camera networks that exhibits good accuracy and practical application. The novel system has consistently beaten other simple and conventional deep models in situations including domain shift and sophisticated obfuscation strategies by incorporating an adaptive attention mechanism and adversarial regulation. The findings demonstrate that multi-scale, semantically-aware feature extraction is more dependable, explainable, and better at making decisions under a variety of real-world conditions.

According to experiments, the structure performs noticeably better than conventional steganography when it comes to managing resource-constrained edge deployment and dynamic hostile interference. Analysis has also revealed that attention-based design and interpretable feature selection are necessary to maintain the model's accuracy and universality in various IoT scenarios. The system's defences may need to be strengthened because of this long-term vulnerability to sophisticated persistent threats.

Future research will focus on real-time meta-adaptation, ongoing learning, and effective scaling for the growth of threat models and IoT device ecosystems. To close the gap between laboratory benchmarks and field

application, expand the framework to include online feature recalibration and improve the coupling of resources and performance. Building a high-resilience, large-scale, and reliable steganalysis platform for the upcoming generation of intelligent visual sensor networks requires following the aforementioned guidelines.

Author Contributions

Agnieszka Szymanski contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision. Weronika Czarnecki contributes to methodology, software, validation, analysis, investigation. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] Zhou, Y., Wang, N., Hong, X., Peng, Y., & Shao, S. (2025). Deep Learning-Based Image Steganography with Latent Space Embedding and Smart Decoder Selection. Entropy, 27(12), 1223. https://doi.org/10.3390/e27121223

[2] Chen, K., Nie, M., Coatrieux, J. L., Chen, Y., & Xie, S. (2025). Airs-Net: Adversarial-improved reversible steganography network for CT images in the Internet of Medical Things and telemedicine. IEEE Journal of Biomedical and Health Informatics. https://doi.org/10.1109/JBHI.2025.3602272

[3] Tan, Y., Xiang, X., Qin, J., & Tan, Y. (2025). Robust Coverless Image Steganography Against Geometric Attacks Via Deep Unsupervised Hashing. IEEE Transactions on Dependable and Secure Computing. https://doi.org/10.1109/TDSC.2025.3626810

[4] Driss, M., Berriche, L., Atitallah, S. B., & Rekik, S. (2025). Steganography in IoT: A comprehensive survey on approaches, challenges, and future directions. IEEE Access. https://doi.org/10.1109/ACCESS.2025.3564120

[5] Zeng, L., Yang, N., Li, X., Chen, A., Jing, H., & Zhang, J. (2023). Advanced image steganography using a U-Net-based architecture with multi-scale fusion and perceptual loss. Electronics, 12(18), 3808. https://doi.org/10.3390/electronics12183808

[6] Ali, S., & Anwer, F. (2025). A Novel Lightweight Framework for Secure and Efficient IoT Communication Using Chaotic Cryptography and Adaptive Steganography. IEEE Transactions on Dependable and Secure Computing. https://doi.org/10.1109/TDSC.2025.3648316

[7] Zhang, S., Li, H., Li, L., Lu, J., & Zuo, Z. (2022). A high-capacity steganography algorithm based on adaptive frequency channel attention networks. Sensors, 22(20), 7844. https://doi.org/10.3390/s22207844

[8] Sharma, V. K., Mir, R. N., & Rout, R. K. (2023). Towards secured image steganography based on content-adaptive adversarial perturbation. Computers and Electrical Engineering, 105, 108484. https://doi.org/10.1016/j.compeleceng.2022.108484

[9] Liang, J., Xie, W., Wu, H., Zhao, J., & Song, X. (2025). High-security image steganography integrating multi-scale feature fusion with residual attention mechanism. Neurocomputing, 632, 129838. https://doi.org/10.1016/j.neucom.2025.129838

[10] Yang, Z., Luo, Y., Yang, J., Xu, X., Zhang, R., & Huang, Y. (2025). Class-aware adversarial unsupervised domain adaptation for linguistic steganalysis. IEEE Transactions on Information Forensics and Security. https://doi.org/10.1109/TIFS.2025.3569409

[11] Al-Rawashdeh, R., Rahman, M. M., & Niazi, M. (2025). Robust image steganography approach based on edge detection combined with cnn algorithm. IEEE Access. https://doi.org/10.1109/ACCESS.2025.3582159

[12] Mikhail, D. Y., Hawezi, R. S., & Kareem, S. W. (2023). An ensemble transfer learning model for detecting stego images. Applied Sciences, 13(12), 7021. https://doi.org/10.3390/app13127021

[13] Liang, J. (2025). Research on encryption algorithm and embedded system optimization strategy based on IoT security. Journal of Cyber Security and Mobility, 14(1), 229-257. https://doi.org/10.13052/jcsm2245-1439.14110

[14] Alshdaifat, E. A., Alshdaifat, D. A., Alsarhan, A., Hussein, F., & El-Salhi, S. M. D. F. S. (2021). The effect of preprocessing techniques, applied to numeric features, on classification algorithms' performance. Data, 6(2), 11. https://doi.org/10.3390/data6020011

[15] Kim, H., Park, H., & Cho, Y. (2025). Performance Comparison of Adversarial Example Attacks Against CNN-Based Image Steganalysis Models. Electronics, 14(22), 4422. https://doi.org/10.3390/electronics14224422

[16] Younis, Y. M., Mstafa, R. J., & Shamal, A. D. (2025). AttenHideNet: A novel deep learning-based image steganography method using a lightweight U-net with soft attention. Applied Soft Computing, 113583. https://doi.org/10.1016/j.asoc.2025.113583

[17] Zheng, M., Law, N. F., & Siu, W. C. (2025). Unveiling image source: Instance-level camera device linking via context-aware deep Siamese network. Expert Systems with Applications, 262, 125617. https://doi.org/10.1016/j.eswa.2024.125617

[18] Zhang, D., Ren, L., Shafiq, M., & Gu, Z. (2022). A lightweight privacy-preserving system for the security of remote sensing images on iot. Remote Sensing, 14(24), 6371. https://doi.org/10.3390/rs14246371

[19] Shehab, D. A., & Alhaddad, M. J. (2022). Comprehensive survey of multimedia steganalysis: Techniques, evaluations, and trends in future research. Symmetry, 14(1), 117. https://doi.org/10.3390/sym14010117

[20] Chen, K., Zhou, Z., Li, Y., Ji, X., Wu, J., Coatrieux, J. L., ... & Coatrieux, G. (2023). RED-Net: Residual and enhanced discriminative network for image steganalysis in the internet of medical things and telemedicine. IEEE Journal of Biomedical and Health Informatics, 28(3), 1611-1622. https://doi.org/10.1109/JBHI.2023.3316468

[21] Liu, Z., Kang, Y., & Liu, X. (2026). AISM: Adversarial image steganography model for defending unauthorized recognition. Journal of Information Security and Applications, 99, 104453. https://doi.org/10.1016/j.jisa.2026.104453

[22] Jan, A., Parah, S. A., Malik, B. A., & Rashid, M. (2021). Secure data transmission in IoTs based on CLoG edge detection. Future Generation Computer Systems, 121, 59-73. https://doi.org/10.1016/j.future.2021.03.005

[23] Al–Janabi, H. A. H., & Al–Ta'i, Z. T. M. (2025, May). Improvement of Video Steganography Using Deep Learning: A Multiscale Attention Mechanism. In 2025 3rd International Conference on Business Analytics for Technology and Security (ICBATS) (pp. 01-09). IEEE. https://doi.org/10.1109/ICBATS66542.2025.11258411

[24] Yuan, K., Yang, Y., Zhang, Z., & Wen, J. (2024). Multi-task few-shot text steganalysis based on context-attentive prototypes. Expert Systems with Applications, 249, 123437. https://doi.org/10.1016/j.eswa.2024.123437

[25] Yang, H., He, H., Zhang, W., & Cao, X. (2020). FedSteg: A federated transfer learning framework for secure image steganalysis. IEEE Transactions on Network Science and Engineering, 8(2), 1084-1094. https://doi.org/10.1109/TNSE.2020.2996612

[26] Chahar, N. K., Dhaka, A., Nandal, A., & Kumar, V. (2025, June). An explainable deep learning framework for usable and secure image steganography. In 2025 International Conference on Electronics, AI and Computing (EAIC) (pp. 1-6). IEEE. https://doi.org/10.1109/EAIC66483.2025.11101501

[27] Hassaballah, M., Hameed, M. A., Awad, A. I., & Muhammad, K. (2021). A novel image steganography method for industrial internet of things security. IEEE Transactions on Industrial Informatics, 17(11), 7743-7751. https://doi.org/10.1109/TII.2021.3053595

[28] Fu, G., Peng, Y., Hu, J., & Hao, G. (2025, November). A Systematic Review of Deep Learning-Based Image Steganography: Paradigms, Progress, and Prospects. In 2025 4th International Conference on Image Processing, Computer Vision and Machine Learning (ICICML) (pp. 307-312). IEEE. https://doi.org/10.1109/ICICML67980.2025.11333424

[29] Zhan, S., Huang, L., Luo, G., Zheng, S., Gao, Z., & Chao, H. C. (2025). A review on federated learning architectures for privacy-preserving AI: Lightweight and secure cloud–edge–end collaboration. Electronics, 14(13), 2512. https://doi.org/10.3390/electronics14132512

[30] Hassaballah, M., Hameed, M. A., Awad, A. I., & Muhammad, K. (2021). A novel image steganography method for industrial internet of things security. IEEE Transactions on Industrial Informatics, 17(11), 7743-7751. https://doi.org/10.1109/TII.2021.3053595