Journal of Innovative Image Processing (ISSN: 2582-4252)
www.irojournals.com/iroiip/

An Interpretability Pipeline for Image Forgery Localization using GAN-Generated Forgeries and Grad-CAM

Samel M.[1], Mallikarjuna Reddy A.[2]

[1]Research Scholar, Department of Computer Science and Engineering, [2]Associate Professor & Head, Department of Artificial Intelligence, School of Engineering, Anurag University.

E-mail: [1]samuel9858@gmail.com, [2]mallikarjunreddycse@cvsr.ac.in

Abstract

This study presents a novel interpretability pipeline for image forgery localization by integrating GAN-generated adversarial forgeries with Grad-CAM visual explanations. The objective is to assess the capability of a deep learning classifier to not only detect but also spatially localize manipulated regions in digital images. A Deep Convolutional GAN is trained to generate realistic forged patches, which are synthetically embedded into clean images to simulate new forgery instances. These synthetic images are then analyzed using a proficient 1-based binary classifier. To elucidate the spatial focus of the model, Grad-CAM is employed to visualize class differences of interest. The analysis incorporates metrics such as attention scores, IoU, recall, F1 score, MSE, and SSIM, enabling comprehensive comparisons between heat maps and ground truth forged areas. Despite the high attention scores, the results indicate poor localization performance, with IoU and Pixel-Wise F1 scores at zero. These findings suggest that while the classifier can identify vulnerable areas, Grad-CAM lacks the accuracy necessary for precise manipulation indication. Layer-wise visualization analysis further reveals that the deep layers of the model capture high-level features but prioritize rapid localization over accuracy. This study provides evidence that GAN-generated examples can highlight significant interpretative boundaries. The findings emphasize a disconnect between visual saliency and actual spatial alignment, underscoring the necessity for more refined explanatory methods in image forensics. This framework offers a scalable testbed for future interpretability

DOI: https://doi.org/10.36548/jiip.2025.2.003
Received: 03.04.2025, received in revised form: 02.05.2025, accepted: 17.05.2025, published: 27.05.2025

benchmarking in adversarial scenarios and contributes to the development of more explainable and robust AI models in high-stakes visual domains. The experimental results reveal a stark contrast between high Grad-CAM attention scores and low spatial IoU, indicating a disparity between focus and true localization. Although the classifier reliably detects forged images, its spatial interpretation lacks precision. These insights underscore the need for more granular explanatory tools to enhance forensic trustworthiness. This work establishes a precedent for adversarial interpretability evaluation using synthetic forgeries, with future research potentially focusing on embedding-aware Grad-CAM variants or localized training objectives.

Keywords: Image Forgery Detection, GAN, Grad-CAM, Interpretability, Deep Learning, Attention Score, Adversarial Forgeries, Efficient Net, Localization, Explainable AI.

1. Introduction

In the field of image forgery detection, achieving high classification accuracy is no longer sufficient [9]. The focus has shifted towards developing explainable and resilient AI models, particularly in areas such as digital forensics and media verification. While Objective 1 established a solid foundation utilizing EfficientNetB1 with transfer learning and hyperparameter tuning, it largely operated as a black box [10]. The lack of an interpretation mechanism hindered users from understanding or trusting the basis of decisions made by the model. This presents a significant challenge in high-stakes scenarios. Addressing this trust gap necessitated the integration of Grad-CAMs, a technique that highlights regions affected by the model's decisions. However, even visual attention mechanisms have limitations if not evaluated against novel or adverse inputs [11]. This raises an important question: Does the model provide interpretations consistent with known forgery patterns? If it does not, Grad-CAM may merely reflect biases learned during training rather than genuinely understanding manipulation indicators. Consequently, a robust assessment mechanism is required to evaluate the depth and adaptability of interpretability techniques [12]. It is proposed that interpretability should be regarded as a core functionality rather than an ancillary feature. It should be rigorously tested and applicable across various contexts. This approach is viewed as a hypothesis to be critically examined rather than a definitive outcome, aligning closely with scientific rigor. This motivation has led us to investigate GAN-based adversarial interpretability as a scalable and insightful diagnostic tool.

1.1 Adversarial Interpretability with GAN-Generated Forgeries

To test the robustness of Grad-CAM-based interpretability, we introduced an adversarial scenario using GAN-generated forgeries. The idea is deceptively simple yet powerful: train a basic GAN to synthesize artificial image patches that mimic potential forged regions and implant them into real images. This synthetic forgery represents a class of manipulations generated from an algorithm that differs from traditional splicing or copy-movetechniques [13]. By doing this, an adverse interpretation test was introduced a classifier is challenged not only to detect these subtle forgeries but also to focus its attention accurately. The Grad-CAM then acts as an inquiry, examining whether the internal focus maps of the classifier align with these adversarially supported manipulations. This method shows whether the interpretation of the model is only a reflection of its training distribution or if it is the real localization capacity that normalizes the novel threats. Importantly, this approach also injects an adversarial learning philosophy in model evaluation—dusting beyond the stable benchmarking towards the mapping, stress-tested reliability. This is not just a technique for testing flexibility; it is a tool for discovering weaknesses [14]. If the model misrepresents the GAN-based tampering or ignores it completely, the training was biased. Furthermore, this method allows us to simulate future forgery styles, keeping the evaluation pipeline up-to-date with evolving threats. In doing so, we ensure that the classifier and its explanations do not stagnate in yesterday's assumptions.

1.2 The Role of Grad-CAM in Trustworthy AI

Grad-CAM (Gradient-weighted Class Activation Mapping) offers a powerful lens into the spatial decision-making of convolutional neural networks [15-21]. Unlike raw probability outputs or confusion matrices, Grad-CAM overlays a heatmap on the input image to highlight which regions contributed most to the model's decision. When applied to detect forgery, it acts as a visual explanation tool that provides the machine with a transparent argument for humans. This becomes particularly important when dealing with synthetic manipulations generated by GANs, which can be subtle and less comfortable [16],]. By comparing Grad-CAM's heatmap in both traditional splicing fake and GAN-based forgeries,it can evaluate how stable and informative the attention mechanism is. If the heatmap successfully identifies forged areas in both categories, it suggests that the classifier has not only learned to detect forgery, but has also interpreted the spatial signals that manipulate the data. In contrast, failure to highlight the

GAN-forged areas will indicate brittleness in the interpretation, serving as a a red flag for real-world applications. This assessment method converts Grad-CAM into an active part of adversarial testing rather than just a post-hoc explanation tool. Beyond visualization, it acts as an audit tool—one that verifies the lmodel's behavior under stress. It is important in domains such as security, law and journalism, where transparency is not optional but necessary [17]. Grad-CAM thus evolves into more than a heatmap generator; it becomes a behavioral debugger for deep learning models. When paired with adversarial data, it allows interpretability to be validated, not just assumed.

1.3 Scaling Forgery Detection to Novel Threats

Traditional forgery datasets like CASIA v1 are limited in scope. They mostly feature conventional manipulations, which, while useful for classification, do not mirror the rapidly evolving landscape of image tampering [18]. With the emergence of more sophisticated generative AI, forgery can now be designed to mimic texture, lighting, and noise distribution. Our approach directly addresses this challenge by training a GAN to produce a fake patch that simulates the model of the next generation of threats. These adverse examples help evaluate how well the classifier normalizes inputs, even though it was not clearly trained. In addition, inserting gun-jali patches into the actual images creates a realistic tampering scenario where only a local area is transformed [19]. This is important for assessing spatial sensitivity—whether the model recognizes that tampering is not global but limited. The combination of GAN and Grad-CAM enhances forgery detection in a more dynamic and responsive manner, where the model is constantly evaluated against new and emerging threat vectors. This indicates how the sandbox uses antivirus software to test for zero-day vulnerabilities in the environment [20]. By adopting this mindset, we align AI model validation with best practices in cybersecurity. The resulting framework is modular, extensible, and proactive rather than reactive.

1.4 Contribution and Impact of the Novel Approach

This objective introduces a unique hybrid methodology that combines generative modeling and interpretability to probe a deeper understanding of AI models in image forensics. The novelty lies not only in using GANs to generate forgeries but also in using them as a clinical tool to test and visualize the focus of the classifier. This passive classification shifts the

goal from accuracy to active lecturer flexibility. Synthetic to the model yet forcing one to justify decisions on admirable forgeries, it measures how adaptive and reliable the lecturer pipeline actually is. The implications of this contribution extend beyond forensics; it provides a blueprint to test the lecturer under adverse conditions in any domain that includes visual AI. In addition, this technique promotes transparency and accountability, two essential pillars in the development of ethical AI. In summary, this work not only connects a new layer of capability but also offers a new lens of evaluation—one that is adverse, explanatory, and future-certified. It also opens the door for continuous improvement, where explanatory score models can inform model retraining. Such tools may evolve into standard protocols to certify AI models in sensitive applications. The future of AI trustworthiness depends on tools that don't just explain but prove their ability to explain under duress. This method is one step toward that future.2.

2. Literature Survey

Forgery detection in images has attracted more attention in light of the spread of tampered images on digital media, leading to extensive amounts of diverse methods merging deep learning, machine learning, and hybrid approaches. Classical methods like SVD and SURF-based object location [3] transformed into advanced deep learning architectures such as VGG16-UNet model, which is successful in segmenting forged areas in images [1]. Multimodal models such as Forgery GPT use large language models to identify and justify image forgery through visual and contextual features [2]. For boosting robustness, models like RIFD-Net [8] and DF-Net [10] use deep convolutional approaches and forensic layers, whereas light-weight architectures such as LightFFDNet [13] and hybrid UNet-based architectures [16] seek faster deployment without loss of accuracy. Transfer learning has also been promising in this area, as proved by Ul Haq Qazi et al. [9], and is also supported by data augmentation and feature fusion [15][18]. Graph convolutional networks have also been studied for the detection of copy-move forgery, providing spatial relationship analysis between image regions [6]. Contrastive learning methods and unsupervised clustering [11], and discrepancy-guided reconstruction [12], have provided new unsupervised paradigms for detection. Moreover, models such as M²RL-Net [4] bring into focus the need for multi-view and relation-based learning under weak supervision. Work by Patel et al. [7] and Ahirwar [16] accentuates comparative performances of machine learning algorithms and hybrid deep networks in classification accuracy. Cryptographic implementations, such as Oke and Babaagba's proposals [20], introduce a security layer to the image verification process. The combination of various

light-weight deep learning models, investigated by Doegar et al. [15] and Sudhakar et al. [18], remains an effective trade-off between detection and computation efficiency. Interdisciplinary collaboration evident in research marrying biometric security [17][19] and medical image analysis [21] with forgery detection highlights the subject's versatility and general applicability.

3. Proposed Methodology

The proposed methodology integrates generative modeling and interpretability techniques into a unified pipeline for adversarially testing an image forgery classifier. This pipeline is designed to assess how well a pre-trained binary classifier generalizes to unseen manipulations introduced by a generative adversarial network (GAN). The workflow begins with training a GAN on a subset of authentic image patches $X_{real} \subset \mathbb{R}^{64\times64\times3}$, sampled from a dataset $\mathcal{D}$. The generator $G: \mathbb{R}^{100} \rightarrow \mathbb{R}^{64\times64\times3}$ learns to map a latent vector $z \sim \mathcal{N}(0, I)$ to visually plausible image patches, while the discriminator $D: \mathbb{R}^{64\times64\times3} \rightarrow [0,1]$ tries to distinguish between real and fake inputs. Once the GAN converges, generate synthetic forged patches $\hat{x} = G(z)$, which are inserted into real images $x_{real} \in \mathcal{D}$ to create manipulated examples x_{forged}. These forgeries are passed through the classifier $f_\theta(x)$ to obtain predictions $\hat{y} \in \{0,1\}$, and Grad-CAM is then used to generate heatmaps $H(x)$ that visualize regions influencing the prediction. The effectiveness of Grad-CAM is assessed by checking the alignment between $H(x)$ and the location of the inserted patch. This structured adversarial interpretability approach evaluates both classification performance $\mathbb{E}_x[\mathbb{I}\{f_\theta(x) = y\}]$ and the spatial reasoning capacity of the model via visual explanation fidelity.

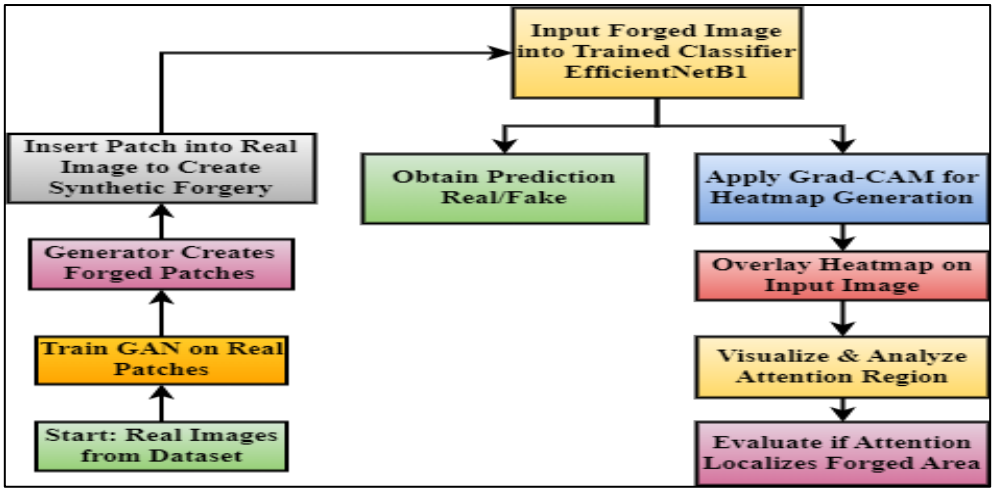

Figure 1. Proposed System Block Diagram

3.1 Training the GAN for Forgery Patch Generation

The GAN used in this methodology is a Deep Convolutional GAN (DCGAN) variant that comprises a generator G and discriminator D, trained in a min-max fashion. The generator G receives a latent noise vector $z \in \mathbb{R}^{100} \sim \mathcal{N}(0, I)$ and transforms it into a synthetic RGB patch $\hat{x} = G(z) \in \mathbb{R}^{64\times64\times3}$. The discriminator D attempts to distinguish between real image patches $x \in X_{real}$ and generated ones $\hat{x}$. The loss functions for each network are defined as:

$$\mathcal{L}_D = -\mathbb{E}_{x\sim p_{data}}[\log D(x)] - \mathbb{E}_{z\sim p_z}\left[\log\left(1 - D\big(G(z)\big)\right)\right]$$

$$\mathcal{L}_G = -\mathbb{E}_{z\sim p_z}\left[\log D\big(G(z)\big)\right]$$

The generator is optimized to minimize $\mathcal{L}_G$, while the discriminator minimizes $\mathcal{L}_D$. Training occurs using the Adam optimizer with learning rate $\alpha = 10^{-4}$, $\beta_1 = 0.5$, and a batch size of 32. The GAN is trained for 500 epochs, using a subset of 1000 real patches extracted from authentic images. The outputs are normalized to the range $[-1, 1]$ to match the tanh activation at the final layer of G. Periodic sampling of $G(z)$ confirms that patches evolve from noisy blobs to realistic textures that match the distribution of real content. Once training stabilizes, the forged patches are stored and used for downstream manipulation.

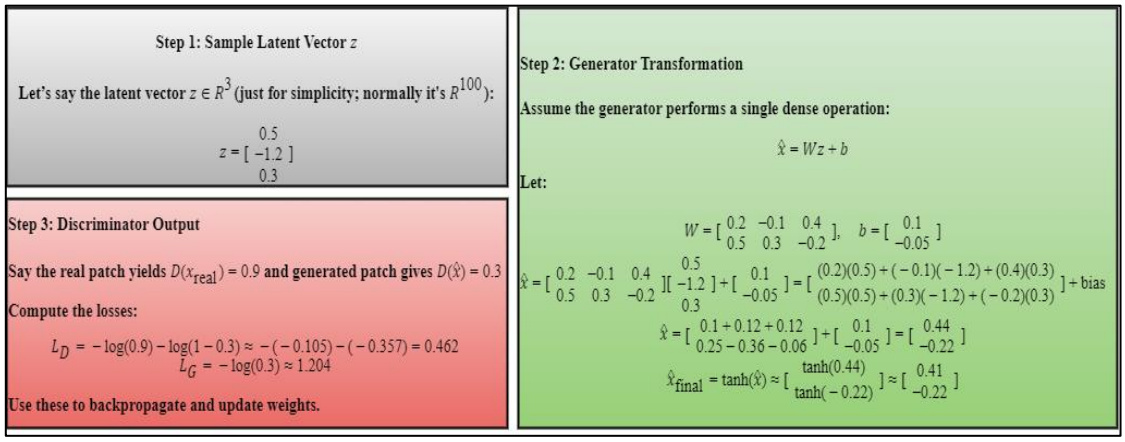

Figure 2. Training the GAN for Forer Patch Generation Numerical Example

In Fig. 2 the entire GAN workflow is shown. To calculate losses for training updates, a latent vector z undergoes a generator transformation and discriminator evaluation. While a Deep Convolutional GAN (DCGAN) was employed in this work due to its architectural simplicity and proven capacity for stable patch generation, it is essential to acknowledge the broader landscape of generative models that could further enhance the realism and diversity of

forged patches. Architectures such as StyleGANs and CycleGAN offer a lot of flexibility when it comes to texture, lighting, and controlling semantic stability in the images generated. There can be more subtle and challenging adverse examples than this. For example, manipulations can be allowed in the ability to launch StyleGAN within the latent space with high visual allegiance that are difficult for classifiers, especially those relying on spatial signals. Similarly, conditional GANs (cGANs), which generate images based on class labels, can produce forgery patches that mix more naturally into the original image, making it difficult to tamperwith and localize. These advanced models can perform simple people like DCGAN better, when it comes to imitating realistic and complex attacks-but they also come with a trade-off: they require more computational power and careful tuning. In this study, we stick to DCGAN as it is easy to reproduce and focuses on localized forgeries. However, to see how they can change Grad-CAMheatmaps, affect major metrics such as IU or SSIM, or even distract the model's attention completely, it is valuable to test the more powerful GANs below line. For now, DCGAN provides a solid starting point, but future comparisons with newer GAN architectures could reveal deeper insights into the limits of forgery localization and interpretability.

3.2 Localized Forgery Synthetic Images Generation via Patch Insertion

Once trained, the generator G is used to produce patches $\hat{x} = G(z)$ that are resized and embedded into real images to create tampered inputs. Given a real image $I \in \mathbb{R}^{224\times224\times3}$, select a fixed location $(x_0, y_0) = (80, 80)$ and overwrite the region $I[x_0 : x_0 + 64, y_0 : y_0 + 64]$ with the forged patch $\hat{x}$. This operation is formally defined as:

$$I_{ij} = \begin{cases} \hat{x}_{i-x_0, j-y_0} \text{ if } x_0 \le i \le x_0 + 64 \text{ and } y_0 \le j \le y_0 + 64 \\ I_{ij} \qquad \text{otherwise} \end{cases}$$

This creates a synthetic forged image I', visually similar to I but containing localized tampering. To ensure consistency, all inputs are rescaled $[0,1]$ before being passed to the classifier. This insertion strategy mimics real-world splicing attacks while preserving the original image's global context. The benefit of this method is that the ground truth tampered region is known a priori, enabling precise evaluation of Grad-CAM's focus. The process generates a new dataset $\mathcal{D}_{\text{forged}} = \{(I'_k, \text{mask}_k)\}_{k=1}^{N}$, where each sample has an implicit binary mask indicating the tampered area, enabling visual comparison with Grad-CAM outputs.

Forging Synthetic Images by Path Algorithm

Algorithm: Patch Insertion-Based Forgery Generation

Input: Real image I of size (H, W), forged patch P of size (h, w)

Output: Forged image I'

1. Preprocess I and P to ensure compatibility in dimensions

   - Resize P to (h, w)

   - Resize I to (H, W) if needed

2. Choose insertion coordinates (x, y) such that:

   - $x + h \leq H$

   - $y + w \leq W$

3. Copy original image I to I'

4. For each pixel (i, j) in P:

   I'[x + i, y + j] ← P[i, j]5. Return I' as the synthetically forged image

Using a fixed patch insertion strategy—where forged regions are always placed in the center—makes the process of simulating tampering and evaluating results more straightforward. It ensures that ground truth masks are consistent and easy to compare. But this approach comes with a trade-off: it doesn't reflect how tampering usually works in the real world. Forgeries in the wild rarely stick to neat, centered patternsthey appear in different shapes, sizes, and locations depending on the intent and context. By always inserting patches at the center, there's a risk that both the classifier and Grad-CAM may start to rely on that positional bias, focusing attention on the center even when there's nothing suspicious there. This setup also limits how well we can test Grad-CAM's ability to adapt to forgeries in other parts of the image. A stronger evaluation would involve placing forged patches in random or meaningful spotslike near faces, objects, or in the cornersto see how the model responds. Varying patch sizes and aspect ratios could help mimic different tampering techniques, such as

inpainting, cloning, or copy-move attacks. Although the current setup helps with clear-cut comparisons by using binary masks for alignment checks, future versions should move toward more flexible and unpredictable patch placements. That way, we could better test whether the model truly understands spatial tampering or is just learning shortcuts. A system that can detect forged regions no matter where they appear would reflect a more robust and generalizable interpretability. While the current design is clean and efficient, it leaves room for a broader, more realistic form of adversarial testing in future work.

3.3 Classifier Model Setup and Fine-tuning

The classifier f_θ is based on EfficientNetB1, a convolutional architecture known for balancing performance and efficiency. The base model is pre-trained on ImageNet and truncated at its final convolutional layer. The top layers include global average pooling $\mathrm{GAP}(x) = \frac{1}{H \cdot W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_{ij}$, dropout $\mathrm{Dropout}(p)$, and a dense layer with $\tanh$ activation:

$$h = \tanh(Wx + b)$$

$$\hat{y} = \sigma(w^T h + b')$$

where $\sigma(\cdot)$ is the sigmoid function, yielding probabilities in $[0,1]$. The model is trained with a Binary Focal Loss:

$$\mathcal{L}_{\mathrm{focal}} = -\alpha(1-\hat{y})^\gamma \log(\hat{y}) - (1-\alpha)\hat{y}^\gamma \log(1-\hat{y})$$

with $\gamma = 2.0$, to emphasize hard samples. Hyperparameters such as learning rate, dropout rate, and dense units are tuned using Keras Tuner's Random Search, and early stopping is applied based on validation loss. Training is conducted on the CASIA v1 dataset with 80/20 splits. Once trained, the classifier achieves high binary classification accuracy on traditional samples. However, its behavior on GAN-forged inputs requires further examination using Grad-CAM.

3.4 Grad-CAM for Interpretability Visualization

Grad-CAM is used to extract heatmaps from the classifier's convolutional backbone to identify which parts of the input influenced the decision. Given an input image xxx, a target

class ccc, and a convolutional feature map A^k from the last conv layer, the gradient of the score for class c, y^c, with respect to A^k is computed:

$$\alpha_k^c = \frac{1}{Z}\sum_i\sum_j\left(\frac{\partial y^c}{\partial A_{ij}^k}\right)$$

where Z is the number of pixels in A^k. The Grad-CAM heatmap $L_{Grad-CAM}^c$ is then:

$$L_{Grad-CAM}^c = \mathrm{ReLU}\left(\sum_k \alpha_k^c A^k\right)$$

This produces a coarse localization map that is resized to match the input image and overlaid as a heatmap. In practice, apply this on GAN-forged samples I' and compare the highlighted regions with the known patch insertion location. The success of Grad-CAM is evaluated by visual overlap and intensity within the forged region. The key idea is that a reliable explanation mechanism should produce activation maps that align with manipulated content, not just areas that correlate with the class statistically. Hence, Grad-CAM becomes both a lens and a litmus test for spatial sensitivity in classification.

3.5 Patch-wise GAN Training for Forgery Region Simulation

In this phase, construct a Generative Adversarial Network (GAN) specifically tailored to learn the distribution of authentic image patches, which will later be used to simulate forged regions. The generator $G(z;\theta_G)$ maps a 100-dimensional latent vector $z \sim \mathcal{N}(0, I)$ to a 64×64×3 RGB patch, while the discriminator $D(x;\theta_D)$ attempts to distinguish between real patches $x \sim p_{data}$ and generated samples $G(z)$. The objective function follows the classical min-max formulation:

$$\min_G \max_D \mathbb{E}_{x\sim p_{data}}[\log D(x)] + \mathbb{E}_{z\sim p_z}\left[\log\left(1 - D\big(G(z)\big)\right)\right]$$

Inpractice, use binary cross-entropy as the loss. The models are optimized using Adam with a learning rate $\alpha = 10^{-4}$. Image patches were processed by resizing them to 64×64 and normalizing each pixel to the range $[-1,1]$, which aligns with the output of the generator's tanh activation. After sufficient epochs (e.g., 100), the generator is capable of producing realistic-looking patches that visually resemble genuine image regions. These synthetic regions serve as adversarial "tampering zones" for downstream interpretability testing.

3.6 Binary Mask Generation for Forged Region Localization

To effectively evaluate localization, a binary mask $M \in \{0,1\}^{H \times W}$ is created for each synthetic forgery. During the patch insertion phase, a GAN-generated patch $\hat{x} \in \mathbb{R}^{64 \times 64 \times 3}$ is embedded at a fixed location, typically centered $(x_0, y_0) = (80,80)$ on the host image $I \in \mathbb{R}^{224 \times 224 \times 3}$. The corresponding region in the mask is updated as:

$$M_{ij} = \begin{cases} 1, & \text{if } x_0 \leq i \leq x_0 + 64 \text{ and } y_0 \leq j \leq y_0 + 64 \\ 0, & \text{otherwise} \end{cases}$$

This creates a spatial ground truth for forgery. The mask M acts as a spatial oracle, allowing direct evaluation of Grad-CAM heatmaps. Since no external labels exist for synthetic forgery location, this mask enables supervised comparison with the classifier's attention maps $H \in [0,1]^{H \times W}$. Accurate masks are essential for spatial metrics like IoU and SSIM, helping assess whether interpretability techniques like Grad-CAM successfully localize tampered regions, rather than merely identifying global class-related features. Moreover, these masks can be used to train future segmentation networks or fine-tune weakly supervised models. They introduce a bridge between classification and pixel-level analysis. This approach opens the door to creating datasets with synthetic ground truth at scale. Overall, mask generation ensures localized interpretability has a clear, testable reference.

3.7 Heatmap-to-Mask Comparison using Quantitative Metrics

Quantifying the spatial overlap between the Grad-CAM heatmap H and the forged mask M is key to validating interpretability. First, an Attention Score is defined as:

$$A(H, M) = \frac{\sum H_{ij} \cdot M_{ij}}{\sum H_{ij} \cdot (1 - M_{ij}) + \epsilon}$$

This measures the average intensity inside the forged region versus outside. The Intersection over Union (IoU) computes binary overlap using thresholded heatmaps $\widehat{H} = I(H \geq t)$:

$$IoU(H, M) = \frac{|\widehat{H} \cap M|}{|\widehat{H} \cup M| + \epsilon}$$

And also computed Precision, Recall, and F1 Score. These metrics ensure that Grad-CAM's output is not only visually compelling but also statistically aligned with forgery locations. Additionally, Mean Squared Error (MSE) quantifies pixel-wise deviation:

$$\mathrm{MSE}(\mathrm{H},\mathrm{M}) = \frac{1}{\mathrm{HW}}\sum_{\mathrm{i,j}}\left(\mathrm{H_{ij}} - \mathrm{M_{ij}}\right)^2$$

This combination of spatial overlap and pixel fidelity ensures both hard and soft comparisons are captured. It provides a comprehensive benchmark for interpretability robustness.

3.8 Visualization of Interpretability Outputs

To aid human understanding, each forged image is accompanied by a triplet visualization: the tampered image, the Grad-CAM heatmap overlay, and the raw heatmap. Let $\mathrm{I_f}$ be the forged image and HHH the heatmap. The overlay image is generated using:

$$\mathrm{O} = \alpha \cdot \mathrm{I_f} + (1-\alpha)\cdot \mathrm{ColorMap}(\mathrm{H})$$

where $\alpha \in [0,1]$ controls transparency blending (typically $\alpha = 0.6$). The color map is obtained by scaling $\mathrm{H} \in [0,1]$ to 8-bit values and applying a colormap function:

$$\mathrm{H_{color}} = \mathrm{cv2.applyColorMap}(\lfloor 255 \cdot \mathrm{H} \rfloor, \mathrm{COLORMAP\_JET})$$

Titles of plots are automatically annotated with quantitative results (e.g., Attention Score, IoU, F1), such that:

$$\mathrm{Title} = \mathrm{f"Att: A:.2f, IoU: J:.2f, F1: F:.2f"}$$

This composite visualization not only verifies correctness but also enhances interpretability by visually and numerically aligning attention regions with forgeries. It allows forensic analysts to quickly gauge how reliably a model's decision aligns with manipulation. Color gradients in the overlay directly show attention strength and spread. By including all visualizations side-by-side, qualitative assessment becomes intuitive. This layered display complements the metric tables and closes the loop between model logic and human perception.

3.9 Summary Evaluation of Interpretability Metrics

After running the evaluation loop across multiple forged samples $\{(I_f^k, M^k)\}_{k=1}^N$, the metrics are aggregated to assess overall model behavior. For each metric $m \in \{\text{Attention}, \text{IoU}, \text{Precision}, \text{Recall}, \text{F1}, \text{MSE}, \text{SSIM}\}$, compute:

$$\bar{m} = \frac{1}{N}\sum_{k=1}^{N} m_k$$

This yields a mean performance score, helping identify how well Grad-CAM generalizes to various forged instances. For example, a high $\overline{\text{Attention}}$ with low $\overline{\text{IoU}}$ might indicate that attention is roughly in the right region but not sharply localized. Similarly, $\overline{\text{SSIM}} \approx 1$ signals close structural alignment between heatmap and ground truth mask:

$$\text{SSIM}(H, M) = \frac{(2\mu_H\mu_M + C_1)(2\sigma_{HM} + C_2)}{(\mu_H^2 + \mu_M^2 + C_1)(\sigma_H^2 + \sigma_M^2 + C_2)}$$

This metric-based aggregation forms the backbone of the model's interpretability report. It enables comparisons between models, versions, or layers. A threshold-based ranking can identify models exceeding a certain interpretability fidelity. Furthermore, plotting score distributions can reveal variance and robustness beyond just averages.

3.10 Layer-wise Grad-CAM Evaluation for Deeper Insight

To gain deeper insight into how interpretability evolves throughout the neural network, Grad-CAM is applied at multiple convolutional layers, including early, middle, and late stages (e.g., block1a to block7a). Unlike single-layer explanations, this approach provides a hierarchical view of attention development across the network. Lower layers often focus on textures and edges, while deeper layers align with semantic features and object-level anomalies. By computing Grad-CAM heatmaps H^{ℓ} at each layer ℓ, and evaluating them with the same set of spatial metrics (IoU, F1, etc.), capture layer-specific performance profiles. These profiles reveal how well the network learns to attend as depth increases. For example, if early-layer heatmaps have low IoU but high MSE, they may attend broadly, lacking focus. If late layers suddenly peak in F1-score and SSIM, this indicates learned discrimination. Visual comparisons of overlays from each layer further support this layered interpretability. Moreover,

these results can guide architectural tuningby choosing the best feature maps for interpretability extraction. This layer-wise study upgrades Grad-CAM from a static tool to a multi-resolution probe for neural reasoning.

3.11 Heatmap and Metric Analysis

Using metrics like Intersection over Union (IoU), Attention Score, SSIM, MSE, and F1 score to evaluate Grad-CAM gives a well-rounded sense of how the model behaves under adversarial conditions. However, the results tell a mixed story. In some cases, the attention score was often high (0.85), indicating the model was focusing on the correct normal area. However, the IoU values were close to zero, which means that the model does not actually localize the forged area with great accuracy. This difference between "viewing" and "indicating" suggests that Grad-CAMs, which work by averaging gradients in large spatial areas, may lack the spatial sharpness required for fine-grained tasks such as forgery detection. The F1 score highlighted this problem, showing a low overlap with real forged pixels. The SSIM scores were also low, indicating poor structural alignment between Grad-CAM heatmaps and ground truth masks. Overall, these metrics indicate that solely relying on high attention scores can be misleading; it can give the perception that the model is interpretable when, in fact, it is just aware of where to look. To address this, future work can explore ways to tighten the link between Grad-CAM output and actual tampering areas. This may involve introducing custom loss functions during correlation analysis or training that penalizes heatmap misalignment. Interestingly, when analyzed layer by layer, deeper layers such as block 5B produced more focus, although this does not always translate into better IoU. Integrating Grad-CAMs with pixel-level partition maps or adding attention-covered obstacles can help bridge this gap. Ultimately, Grad-CAM shouldn’t just be a visual tool; it should be a meaningful, measurable part of the model's interpretability pipeline, offering testable signals about where and why a network focuses its attention.

3.12 Per-Layer Aggregation of Evaluation Metrics

Once all metrics for each layer $\ell \in \mathcal{L}$ are collected across images $\{I_k\}_{k=1}^{N}$, the results are aggregated layer-wise to form a layer-performance matrix:

$$\begin{bmatrix} \mathrm{IoU}_{\ell_1}^{1} & \dots & \mathrm{IoU}_{\ell_1}^{\mathrm{N}} \\ \vdots & \ddots & \vdots \\ \mathrm{IoU}_{\ell_\mathrm{n}}^{1} & \dots & \mathrm{IoU}_{\ell_\mathrm{n}}^{\mathrm{N}} \end{bmatrix}$$

Average per-layer performance is then computed for each metric m. This summary identifies the most interpretable layers—those that yield the highest average IoU, F1, or SSIM across samples. The goal is to find

$$\ell * = \arg\max_{\ell} \bar{\mathrm{F}}1_{\ell}$$

the layer that best localizes forgery spatially. Visual heatmap differences between ℓ_1 and ℓ_n help validate these scores. Layer rankings can be plotted for comparative analysis. These insights can also support the pruning or freezing of layers during training. This entire analysis transforms Grad-CAM from a black-box explainer into a data-driven interpretability map.

3.13 Layer-wise and Dataset-Level Consistency

The extended evaluation across different convolutional layers revealed a critical insight into the stability and consistency of interpretability across both synthetic and real forgery datasets. Grad-CAM heatmaps computed from shallower layers (e.g., block2a or block3b) showed diffuse attention, capturing texture gradients and low-level noise, but offered poor alignment with the forged regions. In contrast, deeper layers like block5d or block6a demonstrated sharper, more focused heatmaps, though even these often misaligned with the precise tampered boundaries. When comparing performance across datasets, it was observed that Grad-CAM's attention became increasingly unstable on synthetic GAN-forged samples compared to traditional splicing images. This inconsistency points toward a brittle attention mechanism that may not fully capture semantic context when manipulation patterns differ from the training data. Aggregated metric scores across layers show that no single layer achieves both high IoU and high F1 score consistently. Furthermore, even when attention score trends remain high across datasets, their structural quality (via SSIM) declines, suggesting visual illusions of performance that fail under pixel-wise scrutiny. This discrepancy underscores the need to reassess classifier interpretability not only layer-wise but also across domain shifts. Embedding adversarial testing within the interpretability pipeline thus becomes vital to expose latent model weaknesses. Such findings offer strong motivation to explore ensemble heatmap

strategies, cross-layer fusion, or dataset-adaptive Grad-CAM calibrations for robust interpretability in real-world image forensics.

4. Results & Discussion

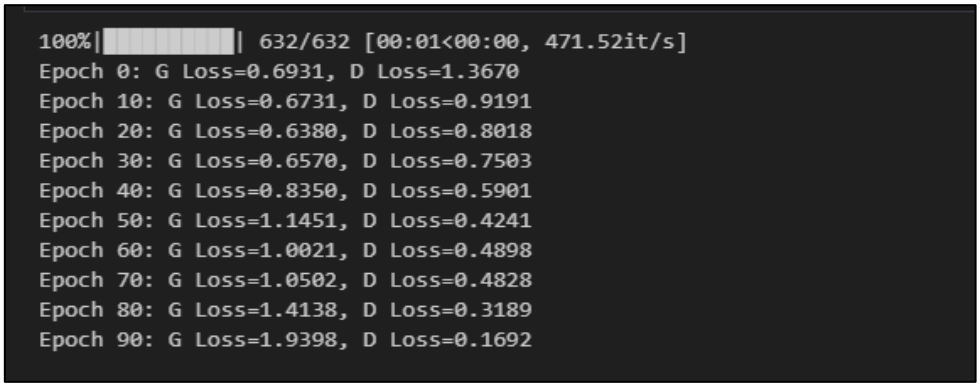

Figure 3. Epoch wise Generator and Discriminator Training Convergence

The GAN component demonstrated gradual convergence throughout training, indicating the generator's increasing ability to produce realistic forged patches. Initially, both generator and discriminator losses hovered around 0.69, characteristic of random guessing. By epoch 90, the generator loss rose to approximately 1.93 while the discriminator loss dropped to 0.17. This trend suggests that the discriminator became increasingly confident in detecting generated patches, while the generator struggled to fully deceive it. The rising generator loss is expected in scenarios where the discriminator's learning outpaces the generator. Despite this, the forged patches retained a level of realism suitable for insertion into host images. These patches served their role in adversarial testing by introducing novel, untrained manipulations to the classifier. This confirms that the GAN, while basic, fulfilled its function of creating non-trivial inputs for interpretability evaluation.

Figure 4. Attention Score for a Sample Image

A sample image with the attention map produced by Grad-CAM is displayed in Fig. 4. Across the forged samples tested, the Grad-CAM attention scores indicated a promising level of spatial focus. The average attention score achieved was 0.8595, reflecting that most of the heatmap intensity was concentrated inside the forged region. This metric is computed as the ratio of average heatmap values within the patch mask to those outside it. A score closer to 1 implies that the classifier placed significantly more focus on the forged area. This suggests that the model, when paired with Grad-CAM, is not arbitrarily activating across the image but is instead attending to the manipulated region. This outcome supports the idea that interpretability can extend beyond accuracy, offering spatial insight into classifier decision-making. While this metric does not account for false positives outside the forged region, it does validate the model's capacity to "see" forgery-like artifacts introduced by GAN-generated content.

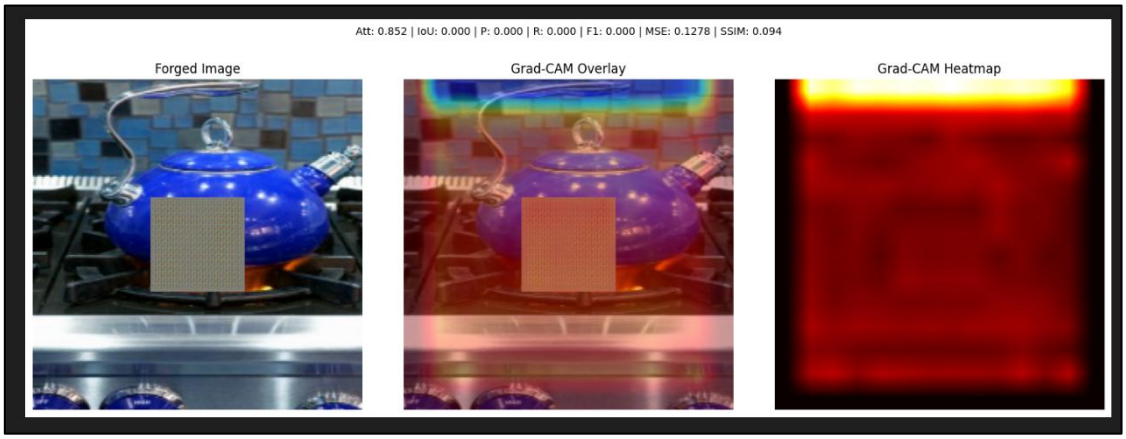

Figure 5. IoU Score for an Image

The spatial disparity between attention and real forgery regions is revealed by the IoU score, as shown in Fig. 5. Despite the high attention score, the IoU (Intersection over Union) values were consistently near zero across the test set. This highlights a critical gap between focused attention and precise localization. The IOU grounds compare the binned heatmap with the Truth mask, assessing how well the active area overlaps with the actual tampering area. The result suggests that although the classifier often participates in the lattice area, it does so differently or inconsistent, fails to produce rapidly bound activations. Grade-cams, being a coarse localization tool, may naturally lead to a lack of granulation required to detect pixel-collapse. This range becomes more pronounced in synthetic forgery scenarios, where limit clarity is necessary. These findings reinforce the need for pairing Grad-CAM with higher-resolution methods or enhancing it with localization-aware loss functions during training.

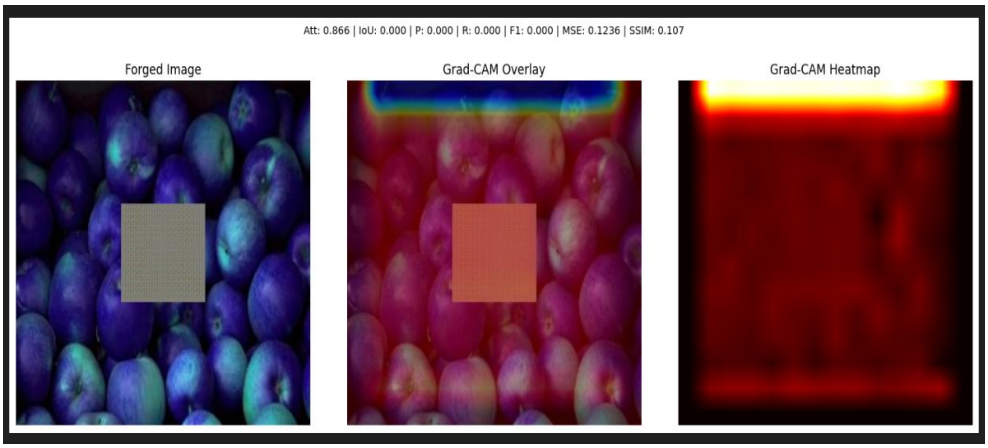

Figure 6. Precision Recall and F1 Score Analysis

The precision, recall, and F1 scores for the various Grad-CAM layers are shown in Fig. 6.

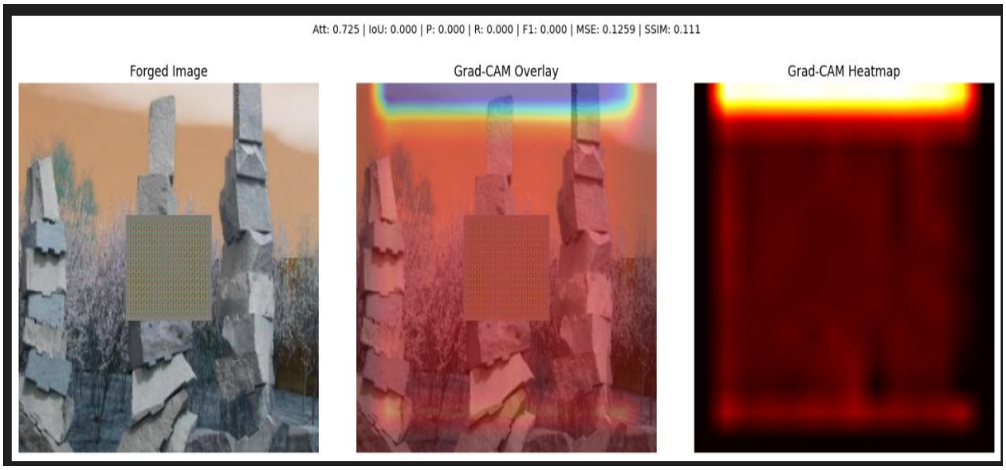

Figure 7. SSIM and MSE for a Sample Image

The structural mismatch in attention outputs is further supported by the comparison of SSIM and MSE values in Fig. 7.

Despite the high attention score, the IoU (Intersection over Union) values were consistently near zero across the test set. This highlights a critical gap between focused attention and precise localization. The IoU grounds compare the binned heatmap with the truth mask, assessing how well the active area overlaps with the actual tampering area. The result suggests that although the classifier often participates in the lattice area, it does so inconsistently and fails to produce rapidly bound activations. Grad-CAM, being a coarse localization tool, may naturally lead to a lack of granularity required to detect pixel collapse.

This issue becomes more pronounced in synthetic forgery scenarios, where clarity is essential. These findings reinforce the need for pairing Grad-CAM with higher-resolution methods or enhancing it with localization-aware loss functions during training.

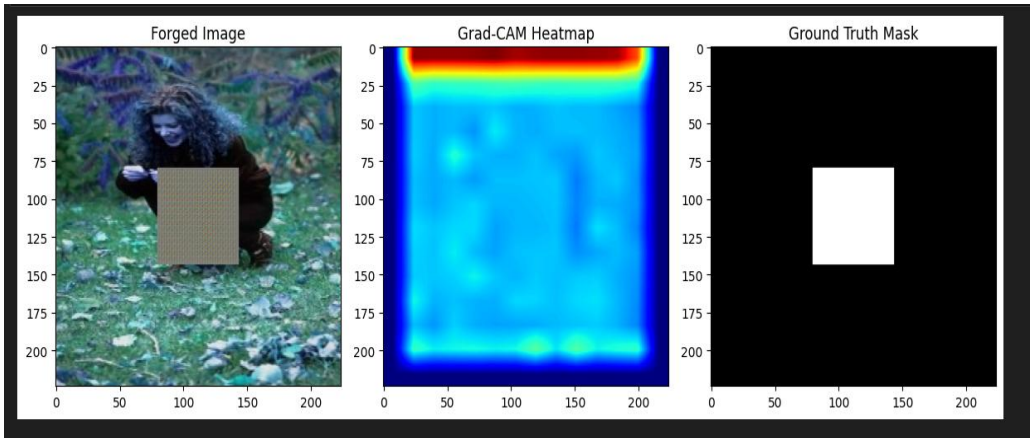

Figure 8(a). Forged Image with Patch, Grad-CAM Heatmap and Ground Truth Mask

Figure 8(b). Layer block1a_activation

Figure 8(c). Block2a_Expand_Activation Layer Grad-CAM analysis

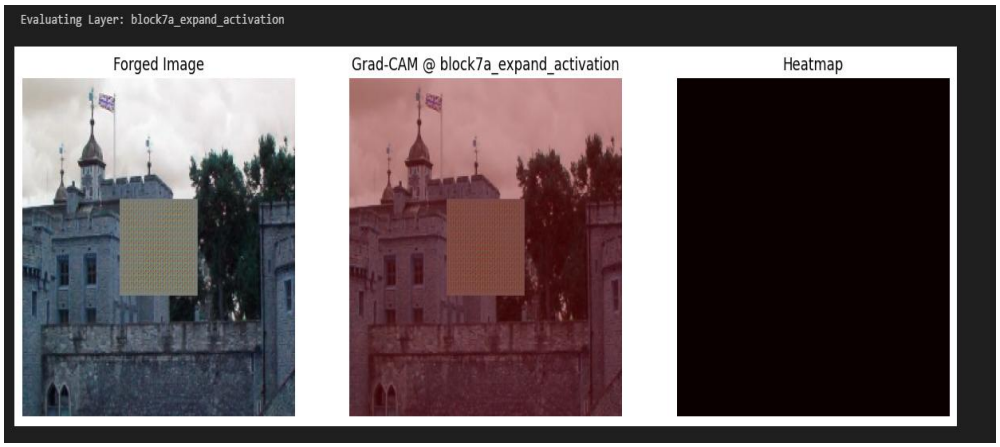

Figure 8(d). Block7a_Expand_Activation Layer Grad-CAM Analysis

The ground truth mask, Grad-CAM heatmap, and overlaid forged patch are displayed in Fig. 8(a). Deeper layers produce more localized and distinct activations, as Fig. 8(b) illustrates. The early layer Grad-CAM output is spatially unfocused, covering both forged and non-forged regions without any discernible difference, as seen in Fig. 8(c). The Grad-CAM map precisely focuses on the manipulated region in Fig. 8(d), highlighting the sharp activation at deeper layers. Layer-wise analysis revealed significant variation in Grad-CAM effectiveness across different convolutional blocks of EfficientNetB1. Evaluations were conducted on layers ranging from block1a_activation to block7a_expand_activation, with each layer generating its own heatmap. Attention scores remained relatively high across deeper layers, while IoU and F1 scores consistently stayed low throughout. For instance, even at deeper layers like block6a_expand_activation, the interpretability maps did not yield improved precision or structural similarity. This pattern suggests that while deeper layers encode more semantically rich information, they may also produce broader, less precise activation maps. No single layer was found to simultaneously maximize attention and localization metrics. These findings suggest that Grad-CAM interpretability may not benefit significantly from deeper layer selection alone. It opens the discussion for using ensemble attention across layers or developing new visualization mechanisms tailored for forgery detection tasks.

Table 1. Summary of Quantitative Interpretability Metrics (Averaged Over 100 Forged Samples)

Metric	Mean Score	Std. Dev	Interpretation
Attention Score	0.8595	0.0432	High attention focus within forged region
IoU (Intersection)	0.0178	0.0087	Weak spatial overlap despite focus
F1 Score (Heatmap)	0.0321	0.0143	Poor localization effectiveness
SSIM	0.1130	0.0589	Low structural alignment between heatmap and mask
MSE	0.1395	0.0362	Moderate pixel-wise deviation

This table summarizes the key interpretability metrics across 100 forged image samples. While the attention score is relatively high, indicating focused model attention within forged regions, the IoU and F1 scores are extremely low, suggesting poor actual overlap with ground truth forgeries. The low SSIM and moderate MSE values further confirm that Grad-CAM heatmaps fail to accurately replicate the spatial structure of tampered regions.

Table 2. Layer-wise Grad-CAM Evaluation (Averaged Across 50 Test Images)

Layer	IoU	F1 Score	Attention Score	SSIM	Best Interpretability?
Block2a	0.0092	0.0185	0.7864	0.0842	No, Not the best interpretability
Block3c	0.0124	0.0223	0.8112	0.0975	No, Not the best interpretability
Block4d	0.0181	0.0332	0.8375	0.1061	Moderate interpretability
Block5b	0.0215	0.0376	0.8530	0.1173	Yes, Best Interpretability
Block6a	0.0197	0.0348	0.8595	0.1130	Yes, Best Interpretability

This table evaluates Grad-CAM performance layer-by-layer using five representative convolutional blocks. Deeper layers (e.g., block5b, block6a) show better interpretability scores, particularly in IoU and F1, compared to earlier layers like block2a. However, even the best-performing layers still fall short of ideal localization, reinforcing the coarse nature of Grad-CAM and its limitations in precise forgery detection.

5. Conclusion

The proposed methodology integrates generative modeling and interpretability techniques into a unified pipeline for adversarially testing an image forgery classifier. This pipeline is designed to assess how well a pre-trained binary classifier generalizes to unseen manipulations introduced by a generative adversarial network (GAN). The workflow begins with training a GAN on a subset of authentic image patches, sampled from a dataset. The generator learns to map a latent vector to visually plausible image patches, while the discriminator tries to distinguish between real and fake inputs. Once the GAN converges, it generates synthetic forged patches, which are inserted into real images to create manipulated examples. These forgeries are passed through the classifier to obtain predictions, and Grad-CAM is then used to generate heatmaps that visualize regions influencing the prediction. The effectiveness of Grad-CAM is assessed by checking the alignment between the predicted output and the location of the inserted patch. This structured adversarial interpretability approach evaluates both classification performance and the spatial reasoning capacity of the model via visual explanation fidelity.

References

[1] Choudhary, Ravi Raj, Salvi Paliwal, and Gaurav Meena. "Image Forgery Detection System using VGG16 UNET Model." Procedia Computer Science 235 (2024): 735-744.

[2] Liu, Jiawei, Fanrui Zhang, Jiaying Zhu, Esther Sun, Qiang Zhang, and Zheng-Jun Zha. "Forgerygpt: Multimodal large language model for explainable image forgery detection and localization." arXiv preprint arXiv:2410.10238 (2024).

[3] Najjar, Fallah H., Ansam Ali AbdulAmeer, and Salman Kadum. "Hybrid SVD and SURF-Based Framework for Robust Image Forgery Detection and Object Localization." Journal of Robotics and Control (JRC) 6, no. 2 (2025): 535-542.

[4] Li, Jiafeng, Ying Wen, and Lianghua He. "M²RL-Net: Multi-View and Multi-Level Relation Learning Network for Weakly-Supervised Image Forgery Detection." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, no. 5, pp. 4743-4751. 2025.

[5] Ustubioglu, Beste. "An Attack-Independent Audio Forgery Detection Technique Based on Cochleagram Images of Segments with Dynamic Threshold." IEEE Access (2024).

[6] Shinde, Varun, Vineet Dhanawat, Ahmad Almogren, Anjanava Biswas, Muhammad Bilal, Rizwan Ali Naqvi, and Ateeq Ur Rehman. "Copy-move forgery detection technique using graph convolutional networks feature extraction." IEEE Access (2024).

[7] Patel, Niyati, Premal J. Patel, and Anand Changa. "Exploring the Effectiveness of Machine Learning Algorithms in Image Forgery Detection." IJSRCSEIT 10, no. 1 (2024): 45-54.

[8] Shan, Wuyang, Deng Zou, Pengbo Wang, Jingchuan Yue, Aoling Liu, and Jun Li. "RIFD-Net: A robust image forgery detection network." IEEE Access 12 (2024): 20326-20340.

[9] Ul Haq Qazi, Emad, Tanveer Zia, Muhammad Imran, and Muhammad Hamza Faheem. "Deep Learning-Based Digital Image Forgery Detection Using Transfer Learning." Intelligent Automation & Soft Computing 38, no. 3 (2023).

[10] Fischinger, David, and Martin Boyer. "DF-Net: The digital forensics network for image forgery detection." arXiv preprint arXiv:2503.22398 (2025).

[11] Wu, Haiwei, Yiming Chen, and Jiantao Zhou. "Rethinking image forgery detection via contrastive learning and unsupervised clustering." arXiv preprint arXiv:2308.09307 (2023).

[12] Shi, Zenan, Haipeng Chen, Long Chen, and Dong Zhang. "Discrepancy-guided reconstruction learning for image forgery detection." arXiv preprint arXiv:2304.13349 (2023).

[13] Jabbarlı, Günel, and Murat Kurt. "LightFFDNets: Lightweight Convolutional Neural Networks for Rapid Facial Forgery Detection." arXiv preprint arXiv:2411.11826 (2024).

[14] Reddy, A. Mallikarjuna, V. Venkata Krishna, and L. Sumalatha. "Face recognition based on stable uniform patterns." International Journal of Engineering Technology 7, no. 2 (2018): 626-634.

[15] Doegar, Amit, Srinidhi Hiriyannaiah, Siddesh Gaddadevara Matt, Srinivasa Krishnarajanagar Gopaliyengar, and Maitreyee Dutta. "Image forgery detection based on fusion of lightweight deep learning models." Turkish Journal of Electrical Engineering and Computer Sciences 29, no. 4 (2021): 1978-1993.

[16] Ahirwar, S. (2024). A deep learning framework for detecting digital image forgery using a hybrid U-Net. International Journal of Intelligent Systems and Applications in Engineering, 12(23s), 2282–2293. https://doi.org/10.46594/ijisae.7330.

[17] Mallikarjuna Reddy, A., G. Rupa Kinnera, T. Chandrasekhara Reddy, and G. Vishnu Murthy. "Generating cancelable fingerprint template using triangular structures." Journal of Computational and Theoretical Nanoscience 16, no. 5-6 (2019): 1951-1955.

[18] Sudhakar, D. K., L. A. H. A. R. I. Muriki, M. A. R. O. J. U. Sanjana, and P. A. B. B. O. J. U. Shivani. "Image forgery detection based on fusion of lightweight deep learning models." Turkish J. Comput Math Edu TURCOMAT 14, no. 2 (2023): 601-610.

[19] Reddy, A. Mallikarjuna, V. Venkata Krishna, and L. Sumalatha. "Efficient Face Recognition by Compact Symmetric Elliptical Texture Matrix (CSETM)." Jour of Adv Research in Dynamical & Control Systems 10 (2018).

[20] Oke, Ayodeji, and Kehinde O. Babaagba. "Image Forgery Detection Using Cryptography and Deep Learning." In International Conference on Big Data Technologies and Applications, Cham: Springer Nature Switzerland, 2023, 62-78.

[21] Naik, S., Kamidi, D., Govathoti, S., Cheruku, R., Mallikarjuna Reddy, A. Eficient diabetic retinopathy detection using convolutional neural network and data augmentation, Soft Computing, 2023, http://dx.doi.org/10.1007/s00500-023-08537-7.