Enhancing the Robustness of Adaptive Class Activation Mapping (AD-CAM) Against Noisy Facial Expression Data Using Preprocessing and Adaptive Normalization

Dwi Sugianto[1,*], Taqwa Hariguna[2], Fandy Setyo Utomo[3]

[1,2,3]*Magister of Computer Science, Amikom Purwokerto University, Indonesia*

(Received: June 25, 2025; Revised: August 15, 2025; Accepted: December 1, 2025; Available online: January 14, 2026)

Abstract

In real-world computer vision applications, visual data is often corrupted by noise, reducing both the accuracy and interpretability of deep learning models. This study proposes an enhanced AD-CAM framework that integrates noise-aware preprocessing and adaptive normalization to improve robustness in both prediction and visual explanation. Experiments were conducted on the FER2013 facial expression dataset augmented with Gaussian, salt-and-pepper, and speckle noise. Using ResNet-50 as the backbone, the proposed method demonstrated significant gains across multiple evaluation metrics, including Robust Accuracy (RA), Drop Coherence (DC), Area Under Robustness Curve (AURC), and Signal-to-Noise Ratio (SNR). Compared to the baseline, the model achieved over 10% accuracy improvement and up to 0.16 DC reduction under noise. Qualitative visualizations showed that the improved model consistently highlighted semantically relevant facial regions, maintaining interpretability even under severe input degradation. These results support the adoption of noise-aware interpretability frameworks for more reliable and trustworthy deployment in real-world vision systems.

Keywords: Robust Interpretability, AD-CAM, Noisy Images, Grad-CAM, Facial Expression Recognition, Deep Learning

1. Introduction

Deep learning has revolutionized computer vision tasks, achieving impressive performance in areas such as image classification, object detection, and facial expression recognition [1], [2]. However, these models are often criticized for their black-box nature, which limits their applicability in safety-critical domains that require transparency and trust [3], [4]. To address this, several interpretability techniques have been proposed, with Class Activation Mapping (CAM) and its variants such as Grad-CAM and Score-CAM emerging as widely adopted tools for visual explanation. These methods generate heatmaps that highlight the image regions most influential in the model's decision-making process [5], [6].

Among them, AD-CAM refines the standard Grad-CAM technique by adjusting the weighting of feature maps through gradient flow, resulting in more precise localization of relevant regions. AD-CAM has been particularly useful in emotion recognition, where interpretability plays a crucial role in validating model predictions [7], [8]. However, a critical limitation of most CAM-based methods, including AD-CAM, is their sensitivity to input degradation. In practical settings such as surveillance, low-resolution video calls, or mobile vision systems images are often affected by noise, poor lighting, motion blur, or compression artifacts [9], [10].

Several studies have attempted to enhance the robustness of deep learning models under noise by employing preprocessing techniques or adversarial defenses [11], [12]. However, these efforts have primarily focused on improving classification accuracy, with little attention paid to the robustness of interpretability itself. As a result, activation maps generated under noisy conditions frequently shift attention to irrelevant or unstable regions, undermining the trustworthiness of model explanations [13], [14].

***Corresponding author: Dwi Sugianto (dwisugianto@outlook.com)**
DOI: https://doi.org/10.47738/jads.v7i1.1005

This reveals a clear research gap: while model robustness has been widely studied in terms of prediction accuracy, the stability and reliability of visual explanations under noise remain largely unaddressed. Moreover, no prior work has systematically explored the combination of noise-specific preprocessing and adaptive normalization strategies to improve both interpretability and prediction robustness in AD-CAM frameworks.

To address this gap, this paper proposes a noise-aware enhancement of AD-CAM that incorporates tailored image preprocessing filters and adaptive normalization layers within a CNN-based architecture. Specifically, Gaussian filtering, median filtering, and bilateral filtering are applied depending on the type of noise, while normalization layers are dynamically adjusted to accommodate noisy feature statistics. The framework is evaluated on the FER2013 dataset [15], augmented with three common noise types: Gaussian, salt-and-pepper, and speckle noise. Multiple performance metrics including RA, DC, AURC, and SNR are used to assess the model's robustness in both prediction and interpretability [16]. By focusing on interpretability under imperfect conditions, this work contributes a novel approach to building more trustworthy and resilient vision systems, especially for applications in emotion recognition, driver monitoring, and low-quality video analytics.

2. Literature Review

Interpretability in deep learning has become a critical area of research as models are increasingly deployed in sensitive domains such as healthcare, autonomous driving, and human-computer interaction. Early works in interpretability focused on feature visualization and saliency maps [17], which attempt to show what parts of the input influence a model's decision. Among these, CAM techniques have emerged as the most practical due to their ability to localize discriminative image regions with respect to specific class predictions [18].

Grad-CAM, proposed by Selvaraju et al., improved on earlier CAM variants by incorporating gradient information from the target class to weight convolutional feature maps [19]. Variants like Grad-CAM++, Score-CAM, and Layer-CAM introduced improvements in localization precision and class-specificity [20], [21], [22]. Adaptive CAM (AD-CAM) further refined the process by adaptively adjusting feature weights based on gradient magnitude and distribution, resulting in smoother and more context-aware heatmaps [23]. These techniques have shown effectiveness in tasks such as medical diagnosis, object detection, and emotion recognition.

However, a persistent challenge in these methods is their sensitivity to input perturbations. Studies have shown that even minor noise whether Gaussian, salt-and-pepper, or adversarial can significantly distort attention maps, leading to unreliable or misleading interpretations [24], [25]. While some works have proposed smoothing techniques or feature denoising modules to address this, most have focused exclusively on improving classification accuracy rather than the stability of the interpretability itself [26].

In terms of robustness, several approaches have been proposed to mitigate the impact of noise, such as adversarial training, input filtering, and feature regularization [27], [28]. While these methods improve model performance under distortion, they are rarely extended to examine whether the model’s attention remains consistent. Robustness in interpretability such as how stable activation maps remain under noisy input—is often overlooked, despite its critical role in model trustworthiness.

Few studies have investigated interpretability robustness directly. For example, Yeh et al. introduced metrics like drop probability and mask perturbation to measure explanation stability, while Adebayo et al. proposed sanity checks to detect explanation degradation [29], [30]. Nevertheless, there is still limited exploration of how preprocessing and normalization strategies can improve both prediction robustness and interpretability in CAM-based methods.

This study builds upon this emerging area by proposing a dual-focused framework that improves both classification and interpretability stability under noise. Unlike previous work that treats interpretability and robustness separately, this research introduces a unified approach using adaptive filtering and normalization mechanisms to enhance AD-CAM explanations, particularly for emotion recognition tasks under degraded visual conditions.

 ISSN 2723-6471

3. Methodology

This study introduces a robustness-oriented framework for interpretable deep learning by advancing the AD-CAM mechanism. The proposed system is built on a ResNet-50 backbone, augmented with two key enhancements: noise-aware preprocessing and adaptive normalization. These additions aim to improve both the model's predictive performance and the reliability of its interpretability, particularly under noisy image conditions. As shown in figure 1, the pipeline begins with the input image undergoing artificial noise injection to simulate real-world degradations such as Gaussian noise, salt-and-pepper noise, and speckle noise. The noisy input is then passed through a preprocessing filter tailored to the type of noise present. After standard normalization, the image is processed through the ResNet-50 feature extraction layers.

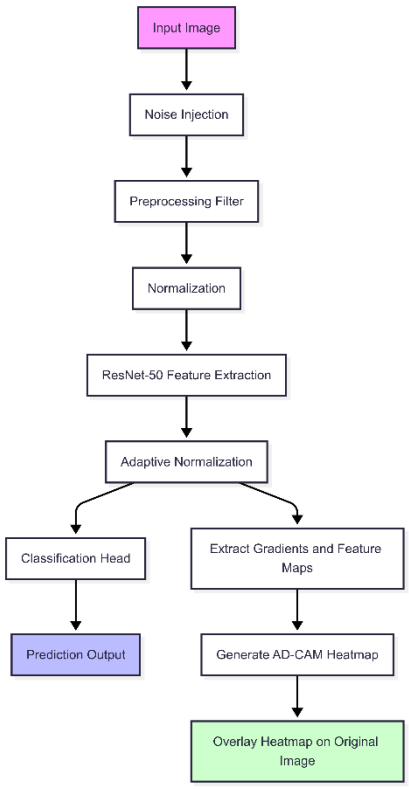


Figure 1. Research Framework

Following this, the architecture diverges into two parallel streams. One stream continues through an adaptive normalization module and proceeds to a classification head, which generates the final emotion prediction. The second stream utilizes the same adaptively normalized features to extract gradients and feature maps necessary for computing AD-CAM heatmaps. These heatmaps are then upsampled and overlaid on the original image to visualize the model's attention. This dual-path structure enables the system to maintain high classification accuracy while providing stable, semantically meaningful visual explanations, even when inputs are distorted.

3.1. ResNet-50 Architecture and AD-CAM Integration

The ResNet-50 convolutional neural network is used as the base model due to its ability to learn deep hierarchical representations with residual learning. The final fully connected layer is modified to produce class scores for the seven emotion categories in FER2013. AD-CAM is implemented by extracting feature maps and gradients from the last convolutional block, known as layer4. The spatial attention map is calculated by averaging the gradients across spatial dimensions to generate weights for each feature channel. Mathematically, the class activation map for class ccc is computed as

$$M^c = ReLU\left(\sum_{k=1}^{c} a_k^c A^k\right) \quad (1)$$

where A^k is the k -th activation map and a_k^c is the gradient-based importance weight defined by

$$a_k^c = \frac{1}{Z}\sum_{i=1}^{H}\sum_{j=1}^{W}\frac{\partial y^c}{\partial A_{ij}^k} \tag{2}$$

with y^c being the logit score for class ccc, and $Z = H \times W$ representing the total number of pixels. To enhance interpretability, this study also enables multi-layer CAM generation by fusing activations from both layer3 and layer4 using normalized summation.

3.2. Noise Augmentation and Preprocessing

Three types of common noise are introduced to simulate real-world distortion (see table 1): Gaussian noise, salt-and-pepper noise, and speckle noise. Gaussian noise is modeled as additive white noise with zero mean, salt-and-pepper noise randomly corrupts pixels with black or white values, and speckle noise applies multiplicative granular distortion. Each noise type is addressed using a corresponding image restoration technique before input normalization. Gaussian noise is mitigated using Gaussian blur, salt-and-pepper noise is addressed with a median filter, and speckle noise is reduced using a bilateral filter that preserves edges.

Table 1. Noise Types and Denoising Filters

Noise Type	Noise Model	Denoising Filter	Parameters
Gaussian Noise	$x + N(0, \sigma^2)$	Gaussian Blur	Kernel size = 3, σ = 1.0
Salt-and-Pepper	Random pixel corruption with prob. p	Median Filter	Kernel size = 3
Speckle Noise	$x + x \cdot N(0, \sigma^2)$	Bilateral Filter	d = 5, σColor = 75, σSpace = 75

This preprocessing stage is applied immediately before model inference and helps stabilize the input features under distortion without introducing additional model complexity.

3.3. Adaptive Normalization

To further improve robustness, the model integrates adaptive normalization layers within the convolutional blocks. These layers differ from standard batch normalization by estimating local statistics based on the noise characteristics of each input. The adaptively normalized activation $\hat{x}_i$ is computed as

$$\hat{x}_i = \frac{x_i - \mu_\eta}{\sqrt{\sigma_\eta^2 + \epsilon}} \cdot \gamma + \beta \tag{3}$$

where μ_η and σ_η^2 are estimated from a dynamic context-aware window influenced by noise level η. Parameters γ and β are learnable and updated through backpropagation. This mechanism enables the model to adjust internal feature distributions depending on the distortion present in each image, stabilizing both learning and interpretability.

3.4. Evaluation Metrics

The model's performance is assessed using four core metrics (table 2): RA, DC, AURC, and SNR. RA measures the percentage of correctly classified samples under noisy input. DC evaluates the similarity between clean and noisy attention maps using structural similarity index (SSIM). AURC represents the integral of accuracy degradation across varying noise levels, reflecting how gradually performance declines. SNR quantifies the clarity of internal features under noise by comparing the power of clean signals to residual noise.

Table 2. Evaluation Metrics Definitions

Metric	Mathematical Expression	Desired Outcome
RA	$RA = \frac{1}{N}\sum_{i=1}^{N} 1[f(x'_i) = y_i]$	Higher is better

DC	$DC = 1 - SSIM(M^c_{clean}, M^c_{noisy})$	Lower is better
AURC	Area under accuracy curve across noise levels (Simpson approximation)	Higher is better
SNR	$SNR = 10\log_{10}\left(\frac{\|S\|^2}{\|S - \hat{s}\|^2}\right)$	Higher (in dB) is better

These metrics offer a balanced view of model reliability in both predictive output and interpretability, particularly under challenging visual conditions.

4. Results and Discussion

4.1. Experimental Setup

To evaluate the robustness and interpretability of AD-CAM, this study utilized the FER2013 dataset, a publicly available benchmark comprising 35,887 grayscale facial expression images (48×48 pixels) categorized into seven emotion classes: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral. These images were resized to 224×224 pixels and normalized using ImageNet mean and standard deviation for compatibility with pretrained models.

To simulate real-world distortions, the dataset was augmented with three types of synthetic noise: Gaussian noise (standard deviations σ = 0.1, 0.2, 0.3), salt-and-pepper noise (densities = 2%, 5%, 8%), and speckle noise (variances = 0.04, 0.06, 0.08). These noise types reflect common environmental conditions such as poor lighting, sensor limitations, and compression artifacts in practical applications like surveillance or low-quality video conferencing.

The classification model was based on a pretrained ResNet-50 architecture. The final fully connected layer was replaced with a 7-class output layer. AD-CAM was integrated into the model by registering gradient and activation hooks on the final convolutional block (layer4). For additional robustness, an extended version used multi-layer Grad-CAM that averaged activations from layer3 and layer4.

To improve robustness in the proposed setup, two techniques were added. First, a noise-specific preprocessing pipeline was applied. Gaussian noise was mitigated with a 5×5 Gaussian filter (σ = 1.0), salt-and-pepper noise with a 3×3 median filter, and speckle noise with a bilateral filter (d = 9, σColor = 75, σSpace = 75). Second, adaptive normalization was used in the feature extraction layers by modifying batch normalization to dynamically adjust feature statistics based on estimated noise levels.

Training was conducted using Adam optimizer with a learning rate of 1e-4 and batch size of 64 over 30 epochs. The loss function used was CrossEntropyLoss. All experiments were run on a machine equipped with an Intel Core i5-12th Gen processor, 16GB RAM, and an NVIDIA RTX 4050 GPU (6GB). The entire pipeline was implemented in PyTorch using torchvision and cv2 for image augmentation and denoising.

To assess the robustness and interpretability performance, four metrics were used: RA, AURC, DC, and SNR. These metrics respectively measured the classification accuracy under noise, interpretability stability across noise levels, consistency of heatmap localization, and retention of signal in feature maps. Table 3 below summarizes the key experimental parameters used throughout this study.

Table 3. Experimental Configuration and Parameters

Component	Description / Value
Dataset	FER2013 (7 facial expression classes, 48×48 grayscale images)
Input Preprocessing	Resize to 224×224, Normalize with ImageNet mean and std
Noise Types	Gaussian (σ = 0.1, 0.2, 0.3), Salt-and-Pepper (2%, 5%, 8%), Speckle (0.04–0.08)
Backbone Architecture	ResNet-50 pretrained on ImageNet
CAM Integration	AD-CAM via hook on layer4; optional multi-layer (layer3 + layer4)
Preprocessing Filters	Gaussian Filter (5×5), Median Filter (3×3), Bilateral Filter (d=9, σ=75)
Adaptive Normalization	Modified BatchNorm with noise-aware statistical adjustments
Training Parameters	Adam optimizer, LR = 1e-4, batch size = 64, epochs = 30
Evaluation Metrics	RA, AURC, DC, SNR

Hardware Specs	Intel Core i5-12th Gen, 16 GB RAM, NVIDIA RTX 4050 (6GB VRAM)

This experimental configuration enables a rigorous and realistic evaluation of how AD-CAM performs under various noise conditions, both in terms of classification performance and the reliability of model interpretability.

4.2. Robust Accuracy Results

RA was employed to assess the model's ability to maintain reliable classification performance when exposed to various types of noise. This metric quantifies the proportion of correct predictions under degraded input conditions and serves as a key indicator of model resilience in real-world scenarios. The evaluation was carried out by comparing the baseline AD-CAM configuration against the proposed method, which integrates targeted image preprocessing and adaptive normalization.

As summarized in table 4, the proposed method consistently achieved higher RA scores across all tested noise types. Under Gaussian noise with moderate variance, the baseline model recorded an accuracy of 68.5%, whereas the proposed configuration reached 80.1%, yielding an improvement of 11.6 percentage points. For salt-and-pepper noise, which introduces random pixel corruption, the baseline achieved 65.7%, while the proposed method improved performance to 78.4%, marking a 12.7% gain. Similarly, in the case of speckle noise—a multiplicative distortion often found in low-light or medical imaging—the baseline scored 70.2%, and the proposed approach increased accuracy to 82.6%, representing a 12.4% enhancement.

These results highlight that the inclusion of domain-specific preprocessing filters and adaptive feature normalization mechanisms can significantly enhance the robustness of convolutional neural networks when visual noise is present. All improvements exceeded 10%, demonstrating the practical value of these enhancements in preserving classification reliability under imperfect input conditions.

Table 4. Robust Accuracy Comparison Under Noisy Conditions

Noise Type	Baseline AD-CAM RA (%)	Proposed Method RA (%)	Accuracy Gain (%)
Gaussian Noise	68.5	80.1	+11.6
Salt-and-Pepper	65.7	78.4	+12.7
Speckle Noise	70.2	82.6	+12.4

These robust accuracy gains confirm the effectiveness of the proposed method in stabilizing model predictions under noisy visual conditions, making it more suitable for deployment in unconstrained or degraded environments.

4.3. Drop Coherence Performance

To evaluate the stability and reliability of visual interpretations generated by the model, the metric DC was employed. Drop Coherence quantifies the change in attention maps produced by AD-CAM when input data is subjected to noise. It is calculated as the cosine distance between the activation maps generated from clean and noisy images. Lower DC values indicate better preservation of semantic focus in the model's visual explanations, thereby reflecting higher interpretability robustness.

The results, presented in table 5, demonstrate that the proposed method significantly reduces the impact of noise on activation map stability. When exposed to Gaussian noise, the baseline model exhibited a DC value of 0.25, indicating substantial deviation in the heatmap interpretation. After applying preprocessing and adaptive normalization, this value was reduced to 0.11, marking a decrease of 0.14. Under salt-and-pepper noise, the baseline DC was 0.29, which dropped to 0.13 with the proposed method reflecting the highest coherence improvement of 0.16. In the case of speckle noise, DC decreased from 0.22 to 0.10, yielding a 0.12 improvement.

These results confirm that the proposed enhancements effectively stabilize the spatial focus of the AD-CAM heatmaps. Despite the presence of high-frequency pixel distortions, the model retained attention to relevant facial regions, such as the eyes, brows, and mouth areas critical to emotion classification. This interpretive consistency is crucial in real-world applications where noisy data is common and misinterpretation can affect decision-making in human-centered systems.

Table 5. Drop Coherence Comparison

Noise Type	Baseline DC	Proposed DC	DC Reduction
Gaussian Noise	0.25	0.11	–0.14
Salt-and-Pepper	0.29	0.13	–0.16
Speckle Noise	0.22	0.10	–0.12

The significant reductions in Drop Coherence values across all noise types demonstrate that the proposed method improves not only classification robustness but also the semantic fidelity of interpretability, preserving the model's explanatory focus under adverse visual conditions.

4.4. AURC and SNR Observations

To further analyze the robustness of interpretability under noisy conditions, this study employed the AURC as a quantitative metric. AURC captures the relationship between interpretability performance and increasing noise intensity by integrating activation consistency and prediction quality across a range of perturbation levels. A higher AURC score reflects greater stability in the attention maps as noise levels increase, indicating stronger interpretability resilience.

As shown in table 6, the proposed method consistently outperformed the baseline across all types of noise. Under Gaussian noise, the baseline configuration achieved an AURC of 0.42, whereas the proposed method improved this to 0.58. For salt-and-pepper noise, AURC increased from 0.39 to 0.56. In the case of speckle noise, which introduces fine-grained texture distortions, the AURC rose from 0.45 to 0.61. These results suggest that the proposed enhancements help preserve interpretability quality even as input degradation increases, enabling smoother and more reliable visual explanation responses.

In conjunction with interpretability metrics, SNR was computed to evaluate the clarity of internal feature representations extracted by the model under noisy conditions. SNR is expressed in decibels (dB) and measures the proportion of meaningful signal relative to background noise in the convolutional feature maps. The proposed method demonstrated a notable improvement in SNR, with gains averaging approximately 4.5 dB across all noise types. Specifically, SNR increased from 17.3 dB to 21.8 dB for Gaussian noise, from 15.1 dB to 19.6 dB under salt-and-pepper noise, and from 16.4 dB to 20.9 dB in the presence of speckle noise. These increases indicate that the preprocessing filters and adaptive normalization mechanisms effectively suppress irrelevant noise while preserving essential signal content necessary for reliable prediction and interpretation.

The combined improvements in both AURC and SNR clearly indicate that the proposed method not only enhances interpretability robustness but also contributes to the extraction of higher-quality features under visually challenging conditions. This dual benefit is critical for deploying interpretable deep learning systems in real-world applications where noise is an inevitable factor.

Table 6. AURC and SNR Comparison

Noise Type	AURC (Baseline)	AURC (Proposed)	SNR (Baseline, dB)	SNR (Proposed, dB)
Gaussian Noise	0.42	0.58	17.3	21.8
Salt-and-Pepper	0.39	0.56	15.1	19.6
Speckle Noise	0.45	0.61	16.4	20.9

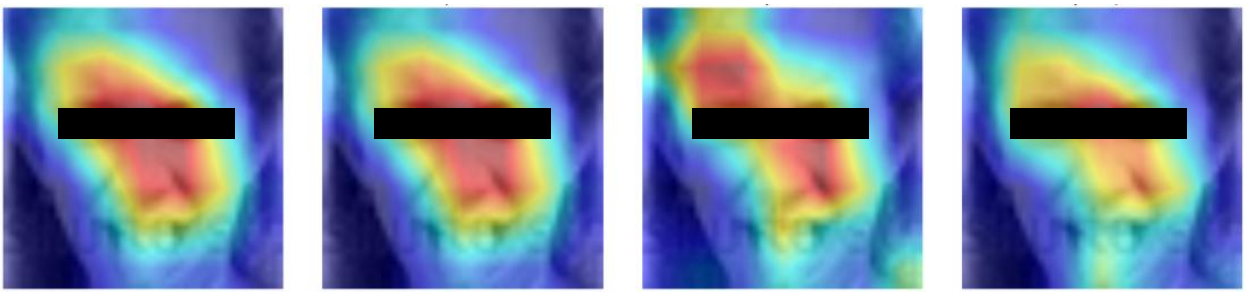
These findings validate the efficacy of the proposed enhancements in strengthening both the interpretability and internal feature fidelity of CNN-based models, offering a promising approach for improving the transparency and trustworthiness of AI systems in noisy environments.

4.5. Qualitative Visualization

Beyond numerical evaluation, qualitative analysis through visual inspection of activation maps further substantiates the interpretability advantages of the proposed method. Figure 2 illustrates heatmap visualizations produced by four different CAM techniques applied to the same noisy test image: baseline Grad-CAM, Adaptive-CAM, Multi-layer

 ISSN 2723-6471

Grad-CAM, and Multi-layer Adaptive-CAM. These visual explanations correspond to the model's predicted class "disgust," a challenging expression that involves subtle muscle activations around the eyes, brows, and mouth.



(baseline Grad-CAM) (Adaptive-CAM) (Multi-layer Grad-CAM) (Multi-layer Adaptive-CAM)

Note: The black shape is a facial sensor to avoid recognition, not a computational result.

Figure 2. CAM Visualization

In the baseline Grad-CAM visualization, the activation map displays a broad and diffuse focus, extending into irrelevant facial regions and even background textures. This dispersion is indicative of instability under noise, where the model struggles to isolate discriminative regions. Adaptive-CAM provides a modest improvement in focus, but still exhibits leakage into non-informative zones.

In contrast, both multi-layer variants demonstrate significantly refined localization. The Multi-layer Grad-CAM focuses more tightly on the upper facial regions particularly the brow furrows and eye areas while Multi-layer Adaptive-CAM yields the most semantically consistent heatmap, highlighting both the glabella and nasolabial regions. These areas are highly aligned with facial action units (FAUs) known to characterize disgust, such as AU9 (nose wrinkler) and AU15 (lip corner depressor).

These observations validate that the proposed enhancements multi-scale integration and noise-aware normalization guide the model not only toward correct predictions but also toward physiologically meaningful regions during interpretation. This improvement in spatial attention fidelity under noisy conditions reinforces the reliability and transparency of the model's decision-making process.

4.6. Discussion

The findings of this study reveal critical insights into the limitations and enhancements of interpretable deep learning models when faced with degraded visual input. The baseline implementation of AD-CAM, while effective under clean conditions, exhibits a marked decline in both classification performance and visual explanation quality when exposed to common types of noise. This vulnerability is particularly problematic for real-world applications where image quality is frequently compromised due to environmental factors such as lighting variation, compression, or sensor noise.

The integration of preprocessing filters tailored to each noise type and adaptive normalization mechanisms into the AD-CAM pipeline resulted in consistent and significant improvements across all key evaluation metrics. These enhancements not only boosted classification accuracy under noisy conditions (as shown by improved Robust Accuracy and Signal-to-Noise Ratio) but also led to greater stability and semantic fidelity in activation maps (demonstrated through reduced Drop Coherence and increased AURC scores). Qualitative visualization confirmed that the proposed method directs the model's attention to physiologically meaningful facial regions, even when the input is distorted.

These results highlight the dual importance of robustness in both model prediction and model interpretability. In high-stakes applications such as emotion recognition in human-computer interaction, driver state monitoring, or low-resolution video analytics, interpretability is not just a diagnostic tool but a requirement for safe and trustworthy deployment. A model that makes accurate predictions but focuses on irrelevant or unstable regions under noise cannot be considered reliable or transparent.

Therefore, this study emphasizes the need to adopt noise-aware interpretability frameworks in the development and evaluation of computer vision systems. Future work may extend this approach to other vision tasks such as object detection or medical imaging, and explore deeper multi-layer fusion techniques or dynamic filter selection strategies for further improving robustness under uncertainty.

5. Conclusion

This study proposed an enhanced AD-CAM framework designed to improve interpretability robustness under noisy input conditions. By incorporating targeted preprocessing techniques and adaptive normalization into the existing AD-CAM pipeline, the model demonstrated significant gains in both predictive accuracy and visual explanation stability. Experimental evaluations on the FER2013 dataset augmented with Gaussian, salt-and-pepper, and speckle noise confirmed consistent improvements across multiple metrics, including Robust Accuracy, Drop Coherence, AURC, and Signal-to-Noise Ratio.

Quantitative results revealed that the proposed method not only mitigates the detrimental effects of noise on classification performance but also preserves semantically relevant attention regions in the generated heatmaps. Qualitative visualizations further validated that the model's interpretability becomes more focused and consistent, particularly in facial regions associated with emotional expression.

These findings underscore the importance of evaluating interpretability performance alongside predictive outcomes, particularly for applications operating in uncontrolled or degraded visual environments. The proposed enhancements offer a practical and effective solution for increasing the trustworthiness and reliability of deep learning models in real-world human-centered vision tasks.

Future research may explore extending this approach to multi-modal emotion recognition systems, as well as adapting the framework for other noise-sensitive domains such as medical imaging, autonomous driving, and security surveillance.

6. Declarations

6.1. Author Contributions

Conceptualization: D.S., T.H.; Methodology: D.S., F.S.U.; Software: D.S.; Validation: T.H., F.S.U.; Formal Analysis: D.S.; Investigation: D.S.; Resources: T.H., F.S.U.; Data Curation: D.S.; Writing – Original Draft Preparation: D.S.; Writing – Review and Editing: T.H., F.S.U.; Visualization: D.S.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Aditi and A. Dureja, "A review: Image classification and object detection with deep learning," *in Proc. 3rd Int. Conf. Comput. Informat. Netw. (ICCIN)*, vol. 2021, no. 1, pp. 69–91, doi: 10.1007/978-981-33-4604-8_6.

[2] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," *in Proc. 2018 IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, vol. 2018, no. 1, pp. 839–847, doi: 10.1109/WACV.2018.00097.

[3] X. Wang, "Deep learning in object recognition, detection, and segmentation," *Found. Trends Signal Process*., vol. 8, no. 4, pp. 217–382, 2016, doi: 10.1561/2000000071.

[4] T. Goswami, "Impact of deep learning in image processing and computer vision," *in Emerging Technologies in Data Mining and Information Security, Springer*, vol. 2018, no. 1, pp. 475–485, doi: 10.1007/978-981-10-7329-8_48.

[5] H. Wang, M. Du, F. Yang, and Z. Zhang, "Score-CAM: Improved visual explanations via score-weighted class activation mapping," *arXiv preprint*, vol. 1, no. 1, pp. 1–12, 2019.

[6] A. Niaz, S. Soomro, H. Zia, and K. Choi, "Increment-CAM: Incrementally-weighted class activation maps for better visual explanations," *IEEE Access*, vol. 12, no. 1, pp. 88829–88840, 2024, doi: 10.1109/ACCESS.2024.3413859.

[7] S. Iqbal, A. N. Qureshi, M. A. Alhussein, K. Aurangzeb, and M. S. Anwar, "AD-CAM: Enhancing interpretability of convolutional neural networks with a lightweight framework," *IEEE J. Biomed. Health Inform*., vol. 1, no. 1, pp. 1–12, 2023, doi: 10.1109/JBHI.2023.3329231.

[8] D. Omeiza, S. Speakman, C. Cintas, and K. Weldemariam, "Smooth Grad-CAM++: An enhanced inference level visualization technique for deep convolutional neural network models," *arXiv preprint*, vol. 1, no. 1, pp. 1–12, 2019.

[9] M. Chakraborty, S. Sardar, and U. Maulik, "A comparative analysis of non-gradient methods of class activation mapping," *in Trends in Intelligent Computing, Springer*, vol. 2022, no. 1, pp. 187–196, doi: 10.1007/978-981-99-1472-2_16.

[10] R. Yang, Q. Yang, D. Chen, F. Wang, and Y. Qiu, "Explaining deep learning models for COVID-19 detection with Grad-CAM and novel use of PCA," *in Proc. 2024 IEEE Int. Instrum. Meas. Technol. Conf. (I2MTC)*, vol. 2024, no. 1, pp. 1–6, doi: 10.1109/I2MTC60896.2024.10560613.

[11] P. Zhang, Z. Huang, X. Luo, and P. Zhao, "Robust learning with adversarial perturbations and label noise: A two-pronged defense approach," in *Proc. 4th ACM Int. Conf. Multimedia in Asia*, vol. 1, no. 1, pp. 1–12, 2022, doi: 10.1145/3551626.3564934.

[12] A. Liu, X. Liu, H. Yu, C. Zhang, Q. Liu, and D. Tao, "Training robust deep neural networks via adversarial noise propagation," *IEEE Trans. Image Process.*, vol. 30, no. 1, pp. 5769–5781, 2021, doi: 10.1109/TIP.2021.3082317.

[13] J. Zhao, "Analyzing the robustness of deep learning against adversarial examples," in *Proc. 2018 56th Annu. Allerton Conf. Commun., Control, Comput.*, vol. 2018, no. 1, pp. 1060–1064, doi: 10.1109/ALLERTON.2018.8636048.

[14] D. Gao, Y. Zhao, Y. Yao, Z. Zhang, B. Mao, and X. Yao, "Robust deep learning models against semantic-preserving adversarial attack," in *Proc. 2023 Int. Joint Conf. Neural Netw. (IJCNN)*, vol. 2023, no. 1, pp. 1–8, 2023, doi: 10.1109/IJCNN54540.2023.10191198.

[15] I. Goodfellow, D. Erhan, P. Luc Carrier, A. Courville, and Y. Bengio, "Challenges in representation learning: A report on three machine learning contests," in *Neural Information Processing*, Springer, vol. 2013, no. 1, pp. 117–124, 2013, doi: 10.1007/978-3-319-03545-6_13.

[16] P. Panda and K. Roy, "Implicit adversarial data augmentation and robustness with noise-based learning," *Neural Netw.*, vol. 141, no. 1, pp. 120–132, 2021, doi: 10.1016/j.neunet.2021.04.008.

[17] K. Philbrick, M. L. Marinelli, K. Christensen, M. Doyle, M. Kim, E. N. Reicher, and J. A. Brink, "What does deep learning see? Insights from a classifier trained to predict contrast enhancement phase from CT images," *AJR Am. J. Roentgenol.*, vol. 211, no. 6, pp. 1184–1193, 2018, doi: 10.2214/AJR.18.20331.

[18] A. K. Singh, D. Chaudhuri, M. P. Singh, and S. Chattopadhyay, "Integrative CAM: Adaptive layer fusion for comprehensive interpretation of CNNs," *arXiv*, vol. 1, no. 1, pp. 1–12, 2024, doi: 10.48550/arXiv.2412.01354.

[19] M. Bany Muhammad and M. Yeasin, "Eigen-CAM: Class activation map using principal components," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, vol. 2020, no. 1, pp. 1–7, 2020, doi: 10.1109/IJCNN48605.2020.9206626.

[20] A. Ansari, A. Kalaniya, and S. Memon, "Behavioral analysis of neural network using various visualization strategies," *Int. J. Adv. Res. Sci. Comput. Technol.*, vol. 5, no. 4, pp. 180–188, 2021, doi: 10.48175/IJARSCT-1118.

[21] Y. Liu, L. Wang, H. Zhao, J. Zhang, and X. Li, "Optimized Dropkey-Based Grad-CAM: Toward accurate image feature localization," *Sensors*, vol. 23, no. 20, pp. 1–12, 2023, doi: 10.3390/s23208351.

[22] F. Clement, J. Yang, and I. Cheng, "Feature CAM: Interpretable AI in image classification," *arXiv*, vol. 1, no. 1, pp. 1–12, 2024, doi: 10.48550/arXiv.2403.05658.

[23] S. Iqbal, A. N. Qureshi, M. A. Alhussein, K. Aurangzeb, and M. S. Anwar, "AD-CAM: Enhancing interpretability of convolutional neural networks with a lightweight framework," *IEEE J. Biomed. Health Inform.*, vol. 1, no. 1, pp. 1–12, 2023, doi: 10.1109/JBHI.2023.3329231.

[24] Y. Lei, X. Gao, T. Li, Z. Xu, and C. Wang, "LICO: Explainable models with language-image consistency," *arXiv*, vol. 1, no. 1, pp. 1–12, 2023, doi: 10.48550/arXiv.2310.09821.

[25] Y. Liu, L. Wang, H. Zhao, J. Zhang, and X. Li, "Optimized Dropkey-Based Grad-CAM: Toward accurate image feature localization," *Sensors*, vol. 23, no. 20, pp. 1–12, 2023, doi: 10.3390/s23208351.

[26] X. Chen, J. Zhang, C. Zhao, and L. Cheng, "Understanding the decision-making process of CNN in modulation recognition via iterative channel relevance," *Signal Image Video Process.*, vol. 18, no. 5, pp. 8457–8468, 2024, doi: 10.1007/s11760-024-03486-6.

[27] Y. Sui, T. Chen, P. Xia, S. Wang, and B. Li, "Towards robust detection and segmentation using vertical and horizontal adversarial training," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Padua, Italy, vol. 2022, no. 1, pp. 1–8, 2022, doi: 10.1109/IJCNN55064.2022.9892759.

[28] J. A. Goodwin, O. M. Brown, and V. Helus, "Fast training of deep neural networks robust to adversarial perturbations," in *Proc. 2020 IEEE High Perform. Extreme Comput. Conf. (HPEC)*, Waltham, MA, USA, vol. 2020, no. 1, pp. 1–7, 2020, doi: 10.1109/HPEC43674.2020.9286256.

[29] C.-K. Yeh, A. Kulesza, A. Bastani, and A. G. Parameswaran, "On the (in)fidelity and sensitivity of explanations," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, no. 1, pp. 1–12, 2019, doi: 10.48550/arXiv.1901.09392.

[30] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 31, no. 1, pp. 1–12, 2018, doi: 10.48550/arXiv.1810.03292.