CS-FEM: Class-Specific Feature Explanation Method and Its Relation to Grad-CAM*

Ahmet Furkan Ün[1][0009−0003−8768−2170], Dinh Nam Le[1], Jenny Benois-Pineau[1][0000−0003−0659−8894], and Marcos Escudero-Viñolo[2][0000−0002−9156−3428]

[1] Université de Bordeaux, CNRS, LaBRI, UMR 5800, 33400 Talence, France
[2] Escuela Politécnica Superior, Universidad Autónoma de Madrid, 28049 Madrid, Spain

Abstract. Explaning DNN decisions requires balancing visual coherence with class specificity. This paper is a follow-up study based on the previously developed Feature Explanation Method (FEM), which is class-agnostic and thus unable to discriminate between different classes in multi-class scenarios. Conversely, gradient-based methods like Grad-CAM provide discrimination among classes but often suffer from high-frequency noise and instability to input perturbations. In this paper, we propose the Class-Specific Feature Explanation Method (CS-FEM), a novel approach that bridges the gap between statistical robustness and semantic specificity. By integrating global class-specific weights into the statistical thresholding framework of FEM, our method effectively functions as a "semantic noise gate", filtering out irrelevant activations while preserving structural coherence. We provide a mathematical comparison demonstrating that CS-FEM acts as a statistically denoised variant of Grad-CAM. The evaluation of the method is performed across three diverse datasets, CAT2000, SALICON, and MexCulture, assessing methods on plausibility (alignment with human attention), correctness (model faithfulness), and stability (Lipschitz continuity). Results show CS-FEM improves human plausibility and stability over gradient-based baselines ($p < 0.001$). Furthermore, we observe a divergence between plausibility and faithfulness: positive-only weighting aligns better with human attention, while preserving negative signals better reflects the model's true inhibitory mechanisms.

Keywords: Explainable AI · Model Interpretability · Plausibility.

1 Introduction

Deep Neural Networks (DNNs) have achieved outstanding performance across a wide range of tasks, including computer vision. However, their complex and non-linear architecture makes them opaque, hiding the internal logic behind their predictions. As the deployment of deep vision systems accelerates in critical domains such as autonomous driving and medical imaging, the demand for

* Source code is available at https://github.com/ahmtfurknun/CS-FEM

transparency and interpretability, collectively known as eXplainable AI (XAI), has become necessary for enabling model reliability and user trust [2].

To address this opacity, the literature has proposed various saliency map generation techniques that highlight the image regions most influential to the model's decision. These methods are broadly categorized into gradient-based and activation-based approaches. Gradient-based methods, such as vanilla Backpropagation [16] and Integrated Gradients [18], compute the sensitivity of the output with respect to input pixels. While theoretically grounded, these methods often suffer from visual noise and instability [17]. Conversely, activation-based methods, specifically the Class Activation Mapping (CAM) family, including CAM [24], Grad-CAM [14], and Grad-CAM++ [7], aggregate high-level feature maps to produce smoother, more semantically coherent explanations.

Among recent developments, the Feature Explanation Method (FEM) proposed by Fuad et al. [8] has appeared as a promising technique for generating visually plausible saliency maps. By applying statistical thresholding to the feature maps of the final convolutional layer, FEM effectively separates significant activation signals from background noise. This results in more focused explanations compared to raw activation methods.

However, the original FEM formulation suffers from a critical limitation: it is class-agnostic. FEM operates solely on the statistics of the feature extraction layers and ignores the weights of the fully connected (FC) classification layer. Consequently, FEM highlights all salient features in an image regardless of the target class. This lack of class-discriminativity limits its utility in multi-class scenarios where distinguishing between different objects in the same scene is required.

To address this limitation, we investigate whether it is possible to retain FEM's statistical robustness while incorporating class-specific information from the classification layer. In this paper, we propose the Class-Specific Feature Explanation Method (CS-FEM), a novel approach that bridges the gap between the statistical stability of FEM and the class-sensitivity of Grad-CAM. Specifically, we introduce a mechanism to weight the binarized, statistically thresholded FEM maps using class-specific weights from the final decision layer. For architectures using Global Average Pooling (GAP) (e.g., ResNet), we demonstrate that these weights are mathematically equivalent to the global gradients used in Grad-CAM, enabling direct weight extraction without backpropagation. However, unlike Grad-CAM, which aggregates raw activations, our method acts as a semantic noise gate by aggregating only statistically significant feature presence.

We evaluate our proposed method using a comprehensive set of metrics across three dimensions: plausibility (alignment with human visual attention), correctness (faithfulness to the model's internal decision-making), and stability (robustness against input perturbations). Our experiments on the CAT2000, SALICON, and MexCulture datasets demonstrate that incorporating decision-head weights significantly improves the discriminative power of FEM without sacrificing its visual clarity.

Our main contributions are as follows:

- We propose CS-FEM, an extension of the FEM that incorporates classification-head weights to enable class-specific explanations.
- We provide a mathematical comparison showing that CS-FEM acts as a variant of Grad-CAM, where statistically thresholded binary maps replace raw activations.
- We perform a multi-faceted evaluation across three diverse datasets (CAT2000, SALICON, MexCulture), showing competitive performance on plausibility, correctness, and stability metrics compared to state-of-the-art methods.

2 Related Work

Saliency map generation has become a cornerstone of XAI for computer vision. Existing methods can be broadly categorized into gradient-based, activation-based, and perturbation-based approaches.

2.1 Gradient-based Methods

Gradient-based methods visualize the sensitivity of the network's output with respect to the input image. The seminal work by Simonyan et al. [16] introduced vanilla Backpropagation, which computes the gradients of the target class score with respect to input pixels. While this highlights relevant pixels, the resulting maps are often visually noisy and difficult to interpret. To address this, various improvements have been proposed. Integrated Gradients [18] accumulates gradients along a path from a baseline image to the input, while SmoothGrad [17] reduces noise by averaging gradients over multiple noisy copies of the input image. Despite these improvements, pixel-level gradient methods often focus on high-frequency details rather than semantic regions.

2.2 Activation-based Methods

Activation-based methods aggregate high-level feature maps from final convolutional layers to capture semantic information. CAM [24] achieves this by replacing fully connected layers with a GAP layer, using the learned weights to determine feature importance. However, this imposes strict architectural constraints and requires retraining.

Grad-CAM [14] strictly generalizes CAM by using the gradients of the target class score flowing into the final convolutional layer as importance weights. Crucially, for architectures using GAP, these gradients are mathematically identical to the weights used in CAM. Grad-CAM thus produces identical results on GAP-based architectures while extending applicability to a much broader range of pre-trained models without modification.

Beyond gradient weighting, variants such as Score-CAM [19] eliminate gradient dependence entirely by using forward-pass confidence changes as weights, while Ablation-CAM [13] measures importance through systematic feature removal. Perturbation-based methods [15] take a complementary approach, probing model behavior through randomized input masking. Our work differs by

retaining gradient-equivalent weights while addressing noise through statistical filtering rather than architectural or input changes.

2.3 Feature Explanation Method

While CAM-based methods rely on gradients or class weights, some approaches focus purely on the statistical properties of feature maps. The FEM proposed by Fuad et al. [8] introduces a statistical thresholding mechanism. FEM asserts that significant activation regions can be identified by analyzing the distribution of activation values within the final convolutional layer, independent of the decision head. By binarizing feature maps using statistical thresholds, FEM generates plausible, noise-free saliency maps.

Since its inception, the FEM framework has evolved to address different architectural challenges. Bourroux et al. [5] extended the approach with the Multi-Layer Feature Explanation Method (MLFEM), which aggregates feature maps across multiple network depths to capture both fine-grained and high-level semantics. More recently, Ayyar et al. [3] proposed the Rollout Feature Explanation Method (RFEM), identifying relevant features in Vision Transformers (ViTs) and introducing class-specificity for attention-based architectures. However, for CNNs, standard FEM remains detached from the classification layer, treating all active features as equally important regardless of the target class.

3 Methodology

The proposed approach extends FEM to address its inherent lack of class discriminativity. By explicitly incorporating weights from the classification head, our method produces more precise and semantically meaningful attributions. The remainder of this section outlines the foundational FEM mechanism before detailing our class-specific formulation.

3.1 FEM Mechanism

FEM [8] identifies semantically relevant image regions by applying a statistical filter to the activations of the final convolutional layer. Assuming that feature map activations approximate a Gaussian distribution, FEM binarizes each feature map $f_i(x, y)$ using a threshold based on its mean μ_i and standard deviation σ_i:

$$b_i(x, y) = \begin{cases} 1, & \text{if } f_i(x, y) \geq \mu_i + k\sigma_i \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where k is a hyperparameter (default $k = 2$) controlling the strictness of the filter. Here, μ_i and σ_i represent the scalar mean and standard deviation computed over the spatial dimensions (x, y) of the i-th feature map. While empirical activations may deviate from strict Gaussianity in practice, this assumption has

Fig. 1. An example from the CAT2000 dataset showing the discrepancy between the activation map generated by standard FEM and the ground-truth Gaze Fixation Density Map (GFDM).

been shown to be a highly effective, standardized heuristic to isolate salient regions [8]. The final feature importance map $s(x, y)$ is generated by weighting each binary map by its mean activation μ_i (which corresponds to the output of GAP in ResNet architectures) and summing across all D feature maps:

$$s(x, y) = \sum_{i=1}^{D} b_i(x, y) \cdot \mu_i \tag{2}$$

This aggregated map is then upsampled to the input image resolution via linear interpolation and min-max normalized to the range $[0, 1]$ to produce the final heatmap.

FEM assumes activation magnitude dictates importance, completely ignoring the classification head. This causes poor alignment with human attention as illustrated in Fig. 1. Figure 2 illustrates two specific failure modes of this class-agnostic approach:

- Highly activated features often dominate the explanation despite being effectively ignored by the classifier (FC weight ≈ 0).
- Crucial predictive features (high positive FC weights) are frequently omitted because their activation magnitude falls below the statistical threshold.

These critical misalignments motivate the integration of class-specific decision weights.

3.2 Proposed Method: Class Specific Feature Explanation Method

To address the explained limitation of FEM, we propose the following modification. Instead of weighting features solely by their mean activations, we introduce a weighting mechanism based on class-specific importance scores from the FC layer.

Given an input image, let c be the target class we wish to explain. We extract the weights $\mathbf{w}^c \in \mathbb{R}^D$ that connect the feature maps to the target-class output node. For CNN architectures using global average pooling (GAP), the

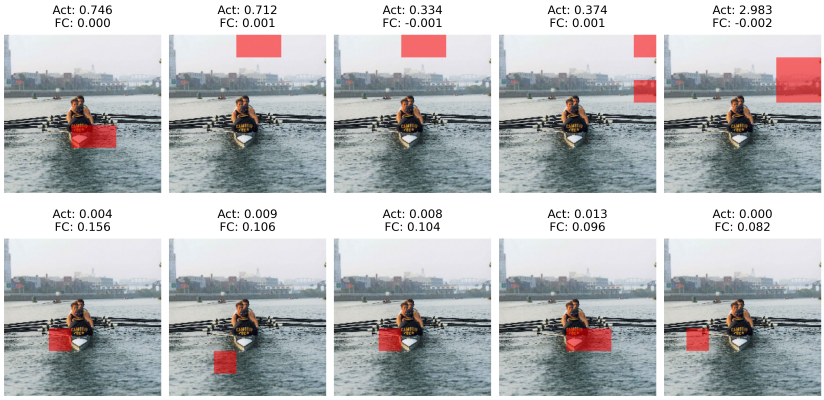

Fig. 2. Visual comparison of standard FEM failure modes on the CAT2000 dataset, highlighting the disconnect between activation magnitude and classification relevance. Top: Feature maps with high activations but near-zero FC weights. Bottom: Feature maps with low activations but high positive FC weights

classification score S^c is a linear combination of the spatially pooled features in channels:

$$S^c = \sum_{i=1}^{D} w_i^c \left(\frac{1}{Z} \sum_{x,y} f_i(x,y) \right) \tag{3}$$

where w_i^c corresponds to the weight for the i-th feature channel (map) for class c and Z is the number of pixels in the feature map.

We retain the statistical robustness of FEM by using the same binarization process defined in Equation 1 to obtain the binary maps $b_i(x,y)$. These binary maps isolate the statistically significant regions of each feature map.

We investigate two aggregation strategies based on different hypotheses about what constitutes a good explanation. The first hypothesis is that explanations should reflect the full decision mechanism, including inhibitory signals that reduce class confidence. The second hypothesis is that human-interpretable explanations should focus only on supporting evidence, as negative contributions may confuse rather than clarify. These lead to two variants:

Combined Contribution (Comb) This variant uses both positive and negative weights. A negative weight w_i^c implies that the presence of feature i reduces the confidence for class c. By including these, we allow inhibitory features to cancel out regions of the map that might otherwise be highlighted. We apply a ReLU activation at the end to ensure the final map only shows positive evidence:

$$s_{comb}(x,y) = ReLU \left(\sum_{i=1}^{D} w_i^c \cdot \mu_i \cdot b_i(x,y) \right) \tag{4}$$

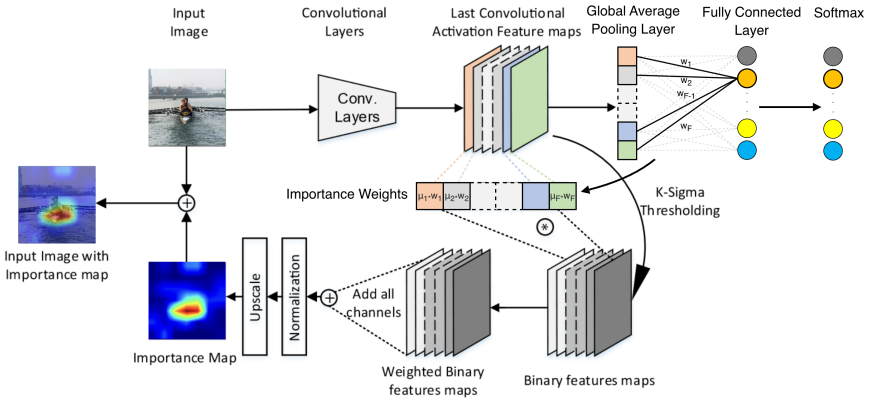

Fig. 3. Overview of the CS-FEM

Positive-Only Contribution (Pos) In this variant, we hypothesize that for saliency generation, we should focus strictly on features that positively support the class decision, avoiding the cancellation effects of inhibitory features. We filter the weights to retain only those where $w_i^c > 0$. The aggregation becomes:

$$s_{pos}(x,y) = \sum_{i \in \{j | w_j^c > 0\}} w_i^c \cdot \mu_i \cdot b_i(x,y) \tag{5}$$

Finally, the resulting map (from either strategy) is upsampled to the input resolution and normalized using the min-max normalization.

Figure 3 illustrates the overall architecture of our proposed CS-FEM, detailing the extraction of feature maps, the computation of class-specific weights, and the final weighted aggregation process.

3.3 Relation to Grad-CAM

This section provides a mathematical comparison demonstrating that the proposed CS-FEM can be interpreted as a variant of Grad-CAM [14].

Grad-CAM Formulation Grad-CAM generates visual explanations by computing the gradients of the target class score S^c with respect to the feature maps f_i of the last convolutional layer. The importance weights α_i^c are obtained by global average pooling of these gradients:

$$\alpha_i^c = \frac{1}{Z} \sum_{x,y} \frac{\partial S^c}{\partial f_i(x,y)} \tag{6}$$

Ideally, for architectures involving a GAP layer (such as the ResNet models used in our experiments), it has been mathematically shown [14] that these gradient-based weights are equivalent to the weights of the fully connected layer. Thus, $\alpha_i^c = \frac{1}{Z} w_i^c$.

The final Grad-CAM heatmap is computed as a weighted combination of the *raw* activation maps, followed by a ReLU activation:

$$\begin{aligned} L^c_{Grad-CAM}(x,y) &= ReLU\left(\sum_{i=1}^{D} \frac{1}{Z} w_i^c \cdot f_i(x,y)\right) \\ &\propto ReLU\left(\sum_{i=1}^{D} w_i^c \cdot f_i(x,y)\right) \end{aligned} \tag{7}$$

Mathematical Comparison By comparing Equation 7 with our proposed aggregation in CS-FEM, we can identify the fundamental difference in how feature significance is handled.

Grad-CAM operates directly on the raw feature map $f_i(x,y)$. This means that all activations within a map, including low-magnitude background noise, are carried forward into the final explanation, scaled only by the class weight w_i^c. If w_i^c is large, even insignificant noise in f_i can become visible in the final heatmap.

In contrast, CS-FEM replaces the raw feature map $f_i(x,y)$ with the statistically filtered term $\mu_i \cdot b_i(x,y)$, weighted by w_i^c. This binarization is a heuristic: it discards the per-pixel magnitude that Grad-CAM retains in $w_i^c \cdot f_i(x,y)$, trading strict mathematical faithfulness for visual clarity. In exchange, it acts as a noise gate. Activations below $\mu_i + k\sigma_i$ are zeroed, so only structurally significant patterns are preserved and high-frequency noise does not accumulate.

Generalization to non-GAP Architectures Our current formulation assumes GAP architectures. Extension to non-GAP models is straightforward: the scalar weight w_i^c in our equations can be replaced by the global average gradient α_i^c, consistent with the Grad-CAM generalization. This allows CS-FEM to be applied to a broader class of models, although we leave the empirical evaluation of gradient-based weights for future work.

4 Experiments and Results

This section presents a comprehensive evaluation of our proposed Class-Specific FEM compared with the original FEM and Grad-CAM, using both quantitative metrics and qualitative analysis.

4.1 Experimental Setup

Datasets We use three datasets that cover different visual domains and tasks:

- **MexCulture [11]:** Focuses on classifying Mexican cultural heritage sites into three architectural styles. It contains 12,000 training, 4,000 validation, and 4,000 testing images. A subset of 284 images includes Gaze Fixation Density Maps (GFDM) collected via psychovisual experiments, allowing alignment evaluation between model explanations and human attention.
- **SALICON [10]:** Contains 10,000 training, 5,000 validation, and 5,000 testing images from MS-COCO. We formulate this as an 80-class multi-label classification task. Saliency maps were generated using a mouse-contingent paradigm correlating with eye-tracking data.
- **CAT2000 [4]:** A benchmark providing high-quality ground-truth eye-tracking data. It includes 2,000 images across 20 distinct categories, which is used to define a standard 20-class classification task.

Model and Training For all experiments, we used a ResNet-50 [9] architecture pre-trained on ImageNet by replacing the final layer with a single linear fully connected classification head. Models were fine-tuned on the respective training splits for a maximum of 50 epochs with a batch size of 32. Optimization used the Adam optimizer with an initial learning rate of 1×10^{-5}. A learning rate scheduler was applied to reduce the learning rate by a factor of 0.5 if validation accuracy did not improve for 3 epochs. Early stopping with a patience of 10 epochs was used based on validation accuracy to retain the best weights. Following the original FEM formulation [8], we set $k = 2$, corresponding to approximately the top 2.5% of activations under Gaussianity.

4.2 Evaluation Metrics

To provide a comprehensive assessment of our proposed method, we use a diverse set of metrics categorized into three categories: plausibility, correctness, and stability.

Ground Truth and Plausibility: While standard XAI metrics often measure faithfulness to the model, they fail to assess *plausibility*, whether the explanation makes sense to a human observer. To evaluate this, we use datasets that provide GFDMs, in which sparse eye-tracking coordinates are converted into continuous ground-truth maps using the standard Gaussian smoothing technique proposed by Wooding [21]. To quantify the alignment between our generated explanations and these GFDMs, we compute four standard XAI plausibility metrics [20]: Similarity (SIM), Pearson Correlation Coefficient (PCC), AUC-Judd (AUC-J), and Normalized Scanpath Saliency (NSS). For all plausibility metrics, higher values indicate better alignment with human attention.

Correctness Metrics: To assess the *faithfulness* of the explanation to the model's internal decision-making, we use perturbation-based metrics. We calculate Deletion Area Under Curve (DAUC) and Insertion Area Under Curve (IAUC) scores [12], which track model prediction scores as pixels are progressively removed or added based on their saliency rank. We also compute Average

Table 1. Quantitative evaluation of XAI methods on the CAT2000 dataset across plausibility, correctness, and stability metrics. Best results are highlighted in **bold**. The arrows (↑ / ↓) indicate whether higher or lower values are desirable.

Category	Metric	Baseline	Grad-CAM	FEM	CS-FEM (Pos)	CS-FEM (Comb)
Plausibility	SIM ↑	0.597 ± 0.073	0.390 ± 0.114	0.445 ± 0.073	$\mathbf{0.452 \pm 0.078}$	0.398 ± 0.106
	PCC ↑	0.699 ± 0.117	0.248 ± 0.248	0.386 ± 0.191	$\mathbf{0.392 \pm 0.195}$	0.265 ± 0.239
	AUC-J ↑	0.930 ± 0.058	0.686 ± 0.214	$\mathbf{0.808 \pm 0.136}$	0.806 ± 0.141	0.702 ± 0.199
	NSS ↑	1.820 ± 0.300	0.718 ± 0.856	$\mathbf{1.187 \pm 0.713}$	$\mathbf{1.187 \pm 0.727}$	0.786 ± 0.869
Correctness	IAUC ↑	0.641 ± 0.243	$\mathbf{0.669 \pm 0.241}$	0.643 ± 0.242	0.660 ± 0.238	0.667 ± 0.240
	DAUC ↓	0.543 ± 0.270	$\mathbf{0.493 \pm 0.279}$	0.535 ± 0.277	0.515 ± 0.279	0.495 ± 0.280
	AD ↓	71.68 ± 30.22	39.21 ± 33.41	38.99 ± 31.08	$\mathbf{37.59 \pm 31.20}$	46.74 ± 34.12
	AG ↑	-0.611 ± 0.296	-0.313 ± 0.299	-0.318 ± 0.285	$\mathbf{-0.305 \pm 0.289}$	-0.382 ± 0.311
	AI ↑	0.035 ± 0.183	0.121 ± 0.327	0.125 ± 0.331	$\mathbf{0.128 \pm 0.334}$	0.102 ± 0.302
Stability (LIP) ↓	Blur	-	2.601 ± 1.619	$\mathbf{2.003 \pm 1.368}$	2.059 ± 1.452	2.381 ± 1.548
	Noise	-	0.895 ± 0.131	0.783 ± 0.123	0.779 ± 0.117	$\mathbf{0.778 \pm 0.109}$
	Perspective	-	0.733 ± 0.271	$\mathbf{0.650 \pm 0.244}$	0.653 ± 0.242	0.678 ± 0.253
	Brightness	-	0.737 ± 0.294	0.603 ± 0.220	$\mathbf{0.589 \pm 0.233}$	0.684 ± 0.254

Drop (AD), Average Increase (AI) [7], and Average Gain (AG) [23] to quantify the changes in the model's confidence when the input is masked using the generated explanation map.

Stability Assessment: Finally, we evaluate the robustness of our explanations using the Lipschitz criterion (LIP) [1]. This metric measures the local stability of an explanation to small input perturbations, where lower values indicate greater stability. To extensively evaluate the stability of our methods, we applied a comprehensive set of image distortions, including additive Gaussian noise, Gaussian blur, uniform brightness shifts, and perspective transformations, strictly following the evaluation protocol established by Zhukov et al. [25].

Center Bias Baseline: Human eye movements exhibit a strong central bias [6,22]. To account for this, we introduce a naive baseline modeled as a centered 2D Gaussian distribution ($\sigma = 0.25$). This baseline quantifies the extent to which plausibility metrics are driven by image centrality rather than semantic content.

4.3 Quantitative Results

In this section, we provide a detailed analysis of the performance of our proposed Class-Specific FEM compared to the baseline Grad-CAM and the original FEM. We organize the discussion into three key dimensions: plausibility, correctness, and stability. Statistical significance was determined using paired t-tests, with a significance threshold of $p < 0.05$.

Plausibility Analysis We evaluated plausibility across CAT2000, SALICON, and the ground-truth subset of the MexCulture datasets.

Regarding the CAT2000 dataset (Table 1), CS-FEM (Pos) demonstrates the strongest alignment with human attention, achieving a SIM score of 0.452. Statistical analysis confirms this is significantly better than Grad-CAM ($0.390, p < 0.001$) and the original FEM ($0.445, p < 0.001$). Interestingly, CS-FEM (Comb)

Table 2. Quantitative evaluation of XAI methods on the SALICON dataset across plausibility, correctness, and stability metrics. Best results are highlighted in **bold**. The arrows (↑ / ↓) indicate whether higher or lower values are desirable.

Category	Metric	Baseline	Grad-CAM	FEM	CS-FEM (Pos)	CS-FEM (Comb)
Plausibility	SIM ↑	0.475 ± 0.110	0.478 ± 0.119	$\mathbf{0.517 \pm 0.107}$	0.504 ± 0.100	0.449 ± 0.112
	PCC ↑	0.458 ± 0.198	0.403 ± 0.211	$\mathbf{0.462 \pm 0.173}$	0.444 ± 0.175	0.378 ± 0.204
	AUC-J ↑	0.815 ± 0.102	0.747 ± 0.156	0.788 ± 0.112	$\mathbf{0.797 \pm 0.108}$	0.748 ± 0.152
	NSS ↑	1.087 ± 0.856	0.917 ± 0.720	$\mathbf{1.029 \pm 0.658}$	1.028 ± 0.673	0.891 ± 0.750
Correctness	IAUC ↑	0.730 ± 0.266	$\mathbf{0.803 \pm 0.222}$	0.741 ± 0.254	0.778 ± 0.237	0.799 ± 0.225
	DAUC ↓	0.632 ± 0.335	$\mathbf{0.544 \pm 0.358}$	0.625 ± 0.342	0.582 ± 0.353	0.551 ± 0.358
	AD ↓	29.37 ± 35.01	$\mathbf{4.11 \pm 13.53}$	8.58 ± 18.67	7.63 ± 18.32	9.29 ± 21.29
	AG ↑	-0.276 ± 0.337	$\mathbf{-0.027 \pm 0.127}$	-0.075 ± 0.172	-0.065 ± 0.173	-0.079 ± 0.204
	AI ↑	0.150 ± 0.357	$\mathbf{0.485 \pm 0.500}$	0.205 ± 0.404	0.325 ± 0.468	0.380 ± 0.485
Stability (LIP) ↓	Blur	-	1.570 ± 0.872	$\mathbf{1.116 \pm 0.451}$	1.264 ± 0.590	1.430 ± 0.721
	Noise	-	0.671 ± 0.202	$\mathbf{0.489 \pm 0.145}$	0.521 ± 0.157	0.584 ± 0.155
	Geometry	-	0.546 ± 0.182	$\mathbf{0.468 \pm 0.152}$	0.487 ± 0.153	0.503 ± 0.154
	Brightness	-	0.356 ± 0.263	$\mathbf{0.262 \pm 0.119}$	0.305 ± 0.173	0.347 ± 0.216

Table 3. Quantitative evaluation of XAI methods on the MexCulture dataset across plausibility, correctness, and stability metrics. Best results are highlighted in **bold**. The arrows (↑ / ↓) indicate whether higher or lower values are desirable.

Category	Metric	Baseline	Grad-CAM	FEM	CS-FEM (Pos)	CS-FEM (Comb)
Plausibility	SIM ↑	0.591 ± 0.081	0.633 ± 0.123	$\mathbf{0.651 \pm 0.081}$	0.648 ± 0.081	0.600 ± 0.113
	PCC ↑	0.737 ± 0.151	0.555 ± 0.270	0.571 ± 0.203	$\mathbf{0.582 \pm 0.197}$	0.532 ± 0.254
	AUC-J ↑	0.928 ± 0.095	0.795 ± 0.205	0.835 ± 0.122	$\mathbf{0.840 \pm 0.120}$	0.797 ± 0.188
	NSS ↑	2.318 ± 0.928	1.215 ± 0.955	1.267 ± 0.835	$\mathbf{1.310 \pm 0.829}$	1.223 ± 0.981
Correctness	IAUC ↑	0.689 ± 0.138	$\mathbf{0.713 \pm 0.126}$	0.693 ± 0.133	0.699 ± 0.131	0.712 ± 0.127
	DAUC ↓	0.651 ± 0.145	$\mathbf{0.634 \pm 0.154}$	0.661 ± 0.146	0.654 ± 0.148	0.640 ± 0.156
	AD ↓	26.94 ± 21.93	$\mathbf{8.20 \pm 11.15}$	13.55 ± 17.29	13.33 ± 16.31	12.43 ± 14.85
	AG ↑	-0.188 ± 0.184	$\mathbf{-0.022 \pm 0.127}$	-0.070 ± 0.151	-0.068 ± 0.149	-0.062 ± 0.144
	AI ↑	0.144 ± 0.352	$\mathbf{0.415 \pm 0.494}$	0.352 ± 0.478	0.335 ± 0.473	0.370 ± 0.484
Stability (LIP) ↓	Blur	-	1.396 ± 0.472	1.252 ± 0.417	$\mathbf{1.236 \pm 0.422}$	1.358 ± 0.450
	Noise	-	0.736 ± 0.120	$\mathbf{0.640 \pm 0.114}$	0.642 ± 0.111	0.688 ± 0.100
	Geometry	-	0.508 ± 0.127	$\mathbf{0.474 \pm 0.114}$	0.475 ± 0.111	0.498 ± 0.119
	Brightness	-	0.454 ± 0.185	0.432 ± 0.135	$\mathbf{0.430 \pm 0.137}$	0.459 ± 0.165

performs worse (0.398 SIM). This highlights an inherent trade-off: dropping inhibitory weights improves alignment with human attention by removing background context, but conflates what humans look at with what the model actually uses.

On SALICON (Table 2), a similar trend is observed. CS-FEM (Pos) significantly outperforms Grad-CAM across all metrics ($p < 0.001$). While it is statistically tied with FEM on NSS ($p > 0.05$), it significantly outperforms FEM on the Similarity (SIM) metric ($p < 0.001$). CS-FEM (Comb) variant again underperforms, yielding significantly lower scores than both Pos and FEM ($p < 0.001$). This reinforces our hypothesis that negative class weights, while useful for discrimination, may aggressively remove context that is salient to human observers.

Regarding the MexCulture subdataset (Table 3), CS-FEM (Pos) achieves the highest NSS score of 1.310, significantly outperforming Grad-CAM ($p < 0.05$) and FEM ($p < 0.01$). CS-FEM (Comb) remains competitive here but does not surpass the Positive-only variant, indicating that focusing strictly on supportive

evidence is the most effective strategy for explaining classification decisions to humans.

Correctness Analysis The Correctness metrics (DAUC, IAUC, AD, AG, AI) assess how accurately the explanation reflects the model's decision-making process.

In terms of Average Drop, CS-FEM (Pos) demonstrates superior faithfulness on CAT2000, achieving the lowest drop of 37.59%, significantly better than Grad-CAM ($p < 0.005$) and FEM ($p < 0.001$). CS-FEM (Comb), however, performs worse (46.74% AD). This is likely because the ReLU applied after summation effectively zeros out regions where negative contributions outweigh positive ones. While mathematically correct for classification, this creates holes in the explanation map that disrupt the semantic coherence required for metrics like Deletion and Average Drop.

On MexCulture, Grad-CAM achieves the best DAUC/IAUC scores, as retaining unbinarized, continuous gradients more faithfully tracks the model's exact computational graph. Still, CS-FEM (Pos) significantly outperforms the original class-agnostic FEM ($p < 0.001$), partially restoring faithfulness.

Stability Analysis A major contribution of our method is its robustness, as evaluated using the LIP metric (where lower values indicate greater stability).

As shown in the Stability sections of all three tables, CS-FEM (Pos) yields significantly lower LIP scores compared to Grad-CAM across all datasets and distortions (Blur, Noise, Geometry, Brightness), with paired t-tests confirming significance at $p < 0.005$. This is an interesting behavior that demonstrates that CS-FEM provides greater stability than the widely used Grad-CAM.

CS-FEM (Comb) is also significantly more stable than Grad-CAM in most cases (e.g., Gaussian Noise on CAT2000, $p < 0.001$). However, it is generally less stable than the Positive variant (e.g., LIP of 2.381 vs 2.059 for Blur on CAT2000). This slight instability arises because the Combined variant involves a subtraction operation (Positive minus Negative weights), which can amplify variance when the input is perturbed, whereas the Positive-only variant purely aggregates constructive interference.

In summary, while both proposed variants improve on the baseline, CS-FEM (Pos) consistently offers the best balance among human plausibility, model faithfulness, and stability, establishing it as a superior and more robust alternative to Grad-CAM.

4.4 Qualitative Results

To complement our quantitative analysis, we present visual comparisons of the saliency maps generated by Grad-CAM, the original FEM, and our proposed CS-FEM variants against the GFDM in Figure 4 on the SALICON dataset. The behavior on other datasets is very much the same.

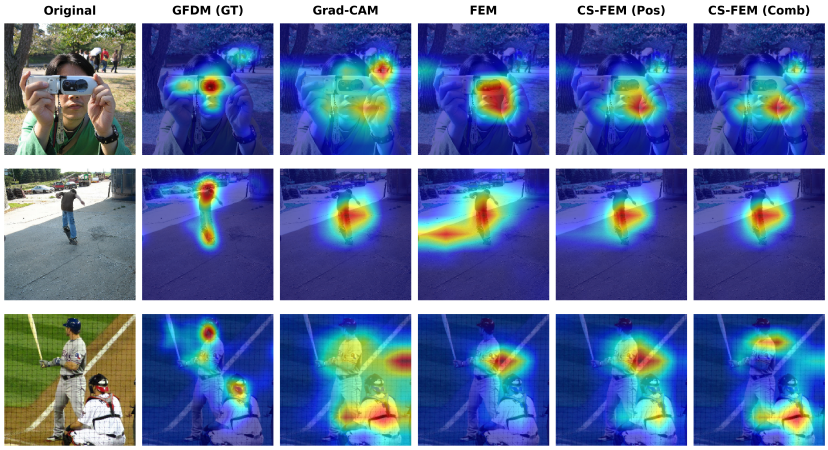

Fig. 4. Qualitative Results on SALICON.

A general observation across all datasets is the structural similarity between specific pairs of methods. The explanation maps produced by Grad-CAM often resemble those of CS-FEM (Comb), as both methods incorporate inhibitory signals (negative gradients or weights) that suppress certain regions. However, Grad-CAM frequently suffers from high-frequency noise and scattered activations, highlighting irrelevant background textures. In contrast, CS-FEM (Comb) produces more localized maps, though the subtraction of negative weights can sometimes result in fragmented explanations.

Similarly, the original FEM shares visual characteristics with CS-FEM (Pos), as both rely on the aggregation of feature maps without negative suppression. However, a distinct improvement in plausibility metrics is observable in our method. While FEM tends to indiscriminately highlight all dominant features in an image, often resulting in diffuse or oversaturated maps, CS-FEM (Pos) produces tighter, more semantically focused focal points. By filtering feature maps based on positive-class contribution, our method effectively removes additional noise that exists in FEM, yielding explanations that are less scattered and more aligned with human attention patterns.

5 Conclusion

In this work, we presented CS-FEM, a novel approach designed to bridge the trade-off between the inherent stability of statistical explanation methods and the semantic specificity required for classification tasks. By integrating global class-specific weights with the robust feature aggregation of the original FEM, our method effectively functions as a "semantic noise gate," filtering out irrelevant activations while preserving the structural coherence of the explanation. We have mathematically shown its equivalence to Grad-CAM except for its statistically filtered features.

Its evaluation across the CAT2000, SALICON, and MexCulture datasets demonstrates that CS-FEM significantly outperforms the popular Grad-CAM in terms of human plausibility and stability ($p < 0.001$), while simultaneously restoring the model faithfulness that the original class-agnostic FEM lacked. Furthermore, our analysis reveals a critical distinction in explanation strategies: the Positive-only variant consistently yields more plausible and stable explanations than the Combined variant, suggesting that, for human interpretability, focusing on supportive evidence is superior to aggressively subtracting inhibitory signals.

This contribution serves as the final puzzle piece in the evolution of the FEM family. The lineage of FEM began with the original single-layer approach for CNNs, later expanding to **MLFEM** for multi-layer aggregation, and more recently evolving into **RFEM** and **RFEM-Class** for Vision Transformers. By introducing a class-specific formulation specifically for Convolutional Neural Networks, CS-FEM fills the remaining gap in this landscape. It ensures that the stability and robustness benefits of the FEM framework are now available for class-discriminative CNN tasks, offering a comprehensive suite of explanation tools across architectures. Important directions for future research include evaluating CS-FEM across diverse architectures and benchmarking against a broader range of XAI methods.

References

1. Alvarez Melis, D., Jaakkola, T.: Towards robust interpretability with self-explaining neural networks. Advances in neural information processing systems **31** (2018)
2. Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al.: Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information fusion **58**, 82–115 (2020)
3. Ayyar, M.P., Benois-Pineau, J., Zemmari, A.: There is more to attention: Statistical filtering enhances explanations in vision transformers. arXiv preprint arXiv:2510.06070 (2025)
4. Borji, A., Itti, L.: Cat2000: A large scale fixation dataset for boosting saliency research. arXiv preprint arXiv:1505.03581 (2015)
5. Bourroux, L., Benois-Pineau, J., Bourqui, R., Giot, R.: Multi layered feature explanation method for convolutional neural networks. In: International Conference on Pattern Recognition and Artificial Intelligence. pp. 603–614. Springer (2022)
6. Buswell, G.T.: How People Look at Pictures: A Study of the Psychology and Perception in Art. University of Chicago Press, Chicago, IL (1935)
7. Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE winter conference on applications of computer vision (WACV). pp. 839–847. IEEE (2018)
8. Fuad, K.A.A., Martin, P.E., Giot, R., Bourqui, R., Benois-Pineau, J., Zemmari, A.: Features understanding in 3d cnns for actions recognition in video. In: 2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA). pp. 1–6. IEEE (2020)

9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
10. Jiang, M., Huang, S., Duan, J., Zhao, Q.: Salicon: Saliency in context. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1072–1080 (2015)
11. Obeso, A.M., Benois-Pineau, J., Guissous, K., Gouet-Brunet, V., Vázquez, M.S.G., Acosta, A.A.R.: Comparative study of visual saliency maps in the problem of classification of architectural images with deep cnns. In: 2018 Eighth International Conference on Image Processing Theory, Tools and Applications (IPTA). pp. 1–6. IEEE (2018)
12. Petsiuk, V., Das, A., Saenko, K.: Rise: Randomized input sampling for explanation of black-box models. In: Proceedings of the British Machine Vision Conference (BMVC) (2018)
13. Ramaswamy, H.G., et al.: Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In: proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 983–991 (2020)
14. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
15. Šimić, I., Veas, E., Sabol, V.: A comprehensive analysis of perturbation methods in explainable ai feature attribution validation for neural time series classifiers. Scientific Reports **15**(1), 26607 (2025)
16. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: visualising image classification models and saliency maps. In: 2nd International Conference on Learning Representations, ICLR 2014 - Workshop Track Proceedings. International Conference on Learning Representations (2014)
17. Smilkov, D., Thorat, N., Kim, B., Viégas, F., Wattenberg, M.: Smoothgrad: removing noise by adding noise. arXiv preprint arXiv:1706.03825 (2017)
18. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: International conference on machine learning. pp. 3319–3328. PMLR (2017)
19. Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., Hu, X.: Score-cam: Score-weighted visual explanations for convolutional neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 24–25 (2020)
20. Wang, W., Shen, J.: Deep visual attention prediction. IEEE Transactions on Image Processing **27**(5), 2368–2378 (2017)
21. Wooding, D.S.: Eye movements of large populations: Ii. deriving regions of interest, coverage, and similarity using fixation maps. Behavior Research Methods, Instruments, & Computers **34**(4), 518–528 (2002)
22. Yarbus, A.L.: Eye Movements and Vision. Plenum Press, New York (1967)
23. Zhang, H., Torres, F., Sicre, R., Avrithis, Y., Ayache, S.: Opti-cam: Optimizing saliency maps for interpretability. Computer Vision and Image Understanding **248**, 104101 (2024)
24. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2921–2929 (2016)
25. Zhukov, A., Benois-Pineau, J., Giot, R.: Evaluation of explanation methods of ai-cnns in image classification tasks with reference-based and no-reference metrics. Advances in Artificial Intelligence and Machine Learning **3**(01), 620–646 (2023)