A Biologically Inspired Filter Significance Assessment Method for Model Explanation

Emirhan Böge[1(✉)], Yasemin Gunindi[2,3], Murat Bilgehan Ertan[4], Erchan Aptoula[3], Nihan Alp[2], and Huseyin Ozkan[3]

[1] School of Informatics, University of Edinburgh, Edinburgh, UK
e.boge@sms.ed.ac.uk
[2] Faculty of Arts and Social Sciences, AlViNlab, Sabanci University, Istanbul, Turkey
{yasemingunindi,nihan.alp}@sabanciuniv.edu
[3] Faculty of Engineering and Natural Sciences, VPAlab, Sabanci University, Istanbul, Turkey
{erchan.aptoula,huseyin.ozkan}@sabanciuniv.edu
[4] Centrum Wiskunde & Informatica (CWI), Amsterdam, The Netherlands
bilgehan.ertan@cwi.nl

Abstract. The interpretability of deep learning models remains a significant challenge, particularly in convolutional neural networks (CNNs) where understanding the contributions of individual filters is crucial for explainability. In this work, we propose a biologically inspired filter significance assessment method based on Steady-State Visually Evoked Potentials (SSVEPs), a well-established neuroscience principle. Our approach leverages frequency tagging techniques to quantify the importance of convolutional filters by analyzing their frequency-locked responses to periodic contrast modulations in input images. By blending SSVEP-based filter selection into Class Activation Mapping (CAM) frameworks such as Grad-CAM, Grad-CAM++, EigenCAM, and LayerCAM, we enhance model interpretability while reducing attribution noise. Experimental evaluations on ImageNet using VGG-16, ResNet-50, and ResNeXt-50 demonstrate that SSVEP-enhanced CAM methods improve spatial focus in visual explanations, yielding higher energy concentration while maintaining competitive localization accuracy. These findings suggest that our biologically inspired approach offers a robust mechanism for identifying key filters in CNNs, paving the way for more interpretable and transparent deep learning models.

Keywords: Explainable AI · Interpretability · Neuroscience-inspired AI · SSVEP · Filter Importance · CAM

1 Introduction

Convolutional neural networks (CNNs) have demonstrated remarkable performance across various computer vision tasks, but understanding their decision-making processes remains a challenge. eXplainable Artificial Intelligence (XAI)

R. Guidotti et al. (Eds.): xAI 2025, CCIS 2577, pp. 422–435, 2026.
https://doi.org/10.1007/978-3-032-08324-1_19

techniques aim to bridge this gap by offering insights into network activations [10]. One of the key challenges in XAI is assessing the importance of individual neurons and filters within deep learning models [20]. Identifying critical neurons aids in model pruning, robustness analysis, and fairness evaluation, while also contributing to a deeper understanding of network behavior [37].

In this work, we build upon the Steady-State Visually Evoked Potential (SSVEP)-based filter assessment framework [5], which draws inspiration from neuroscience [2,23,25]. SSVEPs describe neural responses to flickering visual stimuli at fixed frequencies, revealing frequency-specific tuning in biological neurons [25]. Previous research suggests that CNNs exhibit analogous responses, enabling a biologically inspired approach to filter selection [5]. This makes SSVEP-based filter analysis a promising direction for model interpretability. SSVEP analysis provides a stable measure of convolutional filter importance by leveraging frequency-domain representation of activations. This ability to capture fine-grained filter responses allows for precise identification of highly responsive filters [5]. Most importantly, by aligning with established neuroscience principles, this approach enhances the plausibility of AI explanations, offering a biologically grounded perspective on model behavior.

Our method incorporates SSVEP-based importance assessment with Class Activation Mapping (CAM) techniques, enhancing model interpretability by refining heatmap localization. Unlike traditional CAM methods [6,14,21,30], which consider all convolutional filters, our approach selectively focuses on the most responsive filters. This improves explanation quality by concentrating activation energy on the most relevant regions while reducing noise from less informative convolutional filters.

We evaluate our approach on VGG-16 [32], ResNet-50 [13], and ResNeXt-50 [35] using ImageNet [28]. The SSVEP-enhanced method is applied to Grad-CAM [30], GradCAM++ [6], EigenCAM [21], and LayerCAM [14], and its performance is compared against their respective baselines. Our findings demonstrate that SSVEP-enhanced CAM methods obtain more concentrated and focused explanations while maintaining competitive localization accuracy. *Source code is available at:* github.com/emirhanboge/SSVEP-CAM-Enhancement.

The remainder of this paper is structured as follows: Sect. 2 discusses related work and background, Sect. 3 details our proposed methodology, and Sect. 4 presents experimental findings along with an analysis of the results. Finally, Sect. 5 concludes with key insights and potential future directions.

2 Related Work

2.1 Explainable AI and Model Transparency

The rapid advancements in deep learning have led to significant breakthroughs in computer vision and decision-making systems. However, the black-box nature of these models presents a major challenge in critical applications, particularly

in domains requiring fairness, accountability, and interpretability [1,7,11,12,15, 19,36]. Explainable AI (XAI) methods aim to provide insights into these models' predictions, making their reasoning more transparent. In vision-based applications, saliency methods and Class Activation Mapping (CAM) techniques are widely used to visualize model decisions.

2.2 Saliency-Based Methods and Attribution Techniques

Saliency-based methods highlight the most relevant regions of an input image that influence a model's decision. These approaches can be broadly classified into gradient-based, perturbation-based, and decomposition methods. Gradient-based approaches, such as Grad-CAM [30] and Integrated Gradients [34], compute the gradient of a model's output with respect to input features to determine their importance. Perturbation-based methods, such as LIME [26] and SHAP [18], approximate model behavior by altering input features and observing prediction changes. Decomposition techniques, such as Layer-wise Relevance Propagation (LRP) [3], redistribute the model output back to input features, ensuring the conservation of relevance.

2.3 Class Activation Mapping (CAM) Techniques

CAM-based techniques provide class-specific visualizations by aggregating feature importance scores across a model's activation maps. The original CAM approach [16,38] computes feature importance using global average pooling but requires architectural modifications. Grad-CAM [30] extends CAM by incorporating gradient information, making it applicable to a wider range of architectures. Further improvements were introduced with Grad-CAM++ [6], which refines pixel-wise gradient weighting to improve localization accuracy. Eigen-CAM [21] introduces principal component analysis (PCA) to derive discriminative feature representations without requiring gradients. LayerCAM [14] refines this approach by aggregating activations from multiple layers to improve localization precision. These techniques enable visualization of model focus areas, providing valuable insights into learned representations. While these CAM-based approaches enhance interpretability, the challenge remains in rigorously assessing the quality and faithfulness of the produced explanations. Samek et al. [29] emphasized that visualization methods should be evaluated not only on their perceptual appeal but also on their alignment with model decision-making processes. Their findings highlight the necessity of grounding interpretability techniques in well-defined validation frameworks that can systematically determine whether a heatmap truly reflects the model's reasoning or merely produces plausible-looking attributions.

3 Methodology

This section describes our approach to enhancing explainability in CNNs by utilizing SSVEP analysis for neuron/filter importance assessment. We propose

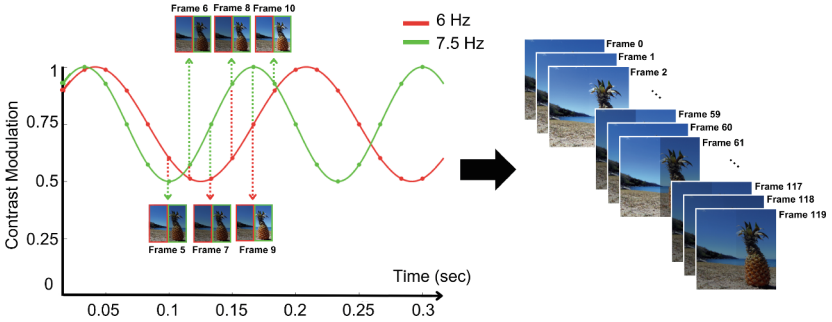

Fig. 1. Sinusoidal contrast modulation of an ImageNet image [28]. Left and right halves are modulated at 6 Hz and 7.5 Hz, respectively, generating 120 flickering frames.

a method that first identifies the most significant neuron filters using SSVEP-inspired frequency tagging and then applies CAM techniques using only the most critical filters.

3.1 Preliminaries: SSVEP-Based Filter Selection for Explainability

The work presented in this paper builds on the SSVEP-based neuron importance framework [5]. This method quantifies filter significance with a biologically inspired approach, and leverages frequency tagging to analyze filter activations under periodic contrast modulations. Analyzing this allows to identify the neurons that exhibit strong phase-locked responses to specific input stimuli.

We follow the SSVEP-based neuron importance setup introduced in [5], which builds upon frequency-tagging principles from neuroscience [2,4,23,25]. To analyze neural responses over time, we apply periodic contrast variations to a static input image using sinusoidal contrast modulation, as visualized in Fig. 1. This process generates a sequence of frames from the same image, where the contrast is dynamically modulated over time. The contrast of each frame i is adjusted on the basis of its temporal position within the sequence. Specifically, for each frame i, the modulation angle $\omega_i \in [0, 2\pi]$ is defined as:

$$\omega_i = 2\pi f \frac{i}{\text{FPS}} + \phi \tag{1}$$

where $f \in \mathbb{R}^+$ is the modulation frequency, $\text{FPS} \in \mathbb{N}$ is the frame rate, $i \in \mathbb{N}$ is the sequential frame index, and $\phi \in \mathbb{R}$ is the phase shift, which is set to zero in our experiments. The method is generic and can be applied to various values of f and i. However, in our specific implementation, we use a sequence of 120 frames ($i \in [0, 119]$), with spatially distinct modulation frequencies. Specifically, following previous work [4], the left half of the image is modulated at $f_{\text{left}} = 6$ Hz and the right half at $f_{\text{right}} = 7.5$ Hz.

As shown in Fig. 1, the modulation is applied independently to each pixel, dynamically adjusting its intensity over time, thereby creating a periodic

flickering effect. To modulate contrast at each pixel location, we scale the original intensity values I according to the sinusoidal function:

$$I' = \left(\frac{\sin(\omega_i) + 1}{2} (s_{\max} - s_{\min}) + s_{\min} \right) \cdot I \tag{2}$$

where $I \in \mathbb{R}^{H \times W}$ represents the original image intensity values, with H and W denoting the image height and width, respectively. The parameters $s_{\max}, s_{\min} \in \mathbb{R}^+$ define the maximum and minimum modulation factors, set to $s_{\max} = 1.0$ and $s_{\min} = 0.5$ in our experiments. This periodic contrast modulation influences the frequency-specific activations in the network, and allows us to evaluate filter responses in the frequency domain, motivated by evidence that neural network exhibit frequency-tuning characteristics analogous to those of biological neurons [5]. As illustrated in Fig. 1, this process generates a sequence of 120 frames over a 2-second window at 60 FPS. Each frame is derived from the same static image, with pixel intensities oscillating according to their modulation frequency. No spatial transformations occur, which ensures that only the temporal flickering effect is introduced. This controlled contrast modulation is key to assessing how neurons respond to periodic visual changes. By analyzing how different convolutional filters in a pretrained network respond to these modulated inputs, we can quantify their frequency selectivity and importance.

These tagged images are forward-passed through a pretrained network, and the resulting activations of each filter are recorded over time. Later, the frequency response of each filter is analyzed using Fast Fourier Transform (FFT) to extract frequency components. To quantify the relevance of each filter, we compute Signal-to-Noise Ratio (SNR) in the frequency domain, using the FFT magnitudes of activations. The SNR at a given frequency f, denoted as $\mathrm{SNR}(f) \in \mathbb{R}^+$, and is computed as follows [8,17]:

$$\mathrm{SNR}(f) = \frac{F(f)}{\frac{1}{|\mathcal{N}(f)|} \sum_{k \in \mathcal{N}(f)} F(k)} \tag{3}$$

where $F(f)$ denotes the Fourier magnitude at the discrete frequency bin f, representing the activation strength of a filter at that frequency. The set $\mathcal{N}(f) \subset \mathbb{R}^+$ consists of neighboring frequencies used to estimate the noise baseline, and $|\mathcal{N}(f)|$ is the total number of noise bins considered. To further improve numerical stability, we express SNR in decibels (dB), given by:

$$\mathrm{SNR}_{\mathrm{dB}}(f) = 10 \log_{10}(\mathrm{SNR}(f)) \tag{4}$$

A higher SNR value indicates that the filter exhibits a strong phase-locked response at the target frequency, which signifies its importance. Finally, as each convolutional filter produces SNR values across multiple discrete frequency bins, we summarize these values into a single scalar SNR score for each filter by computing the mean SNR across all frequencies. This metric is then used to rank all filters within the selected convolutional layers.

This ranking allows us to ensure that the explanation process is driven by filters with the strongest responses to specific stimuli. By selectively retraining only the top-K most responsive filters, we enforce a biologically inspired sparsity constraint on the explainability process.

3.2 Combining SSVEP-Based Filter Selection Into CAM

After identifying the most important neurons, we combine this information into the CAM generation process. Unlike traditional CAM methods, which compute class activation maps using a weighted sum of all convolutional feature maps based on either global average pooled classification weights [38] or gradient-based importance scores [30,31], our approach pre-selects only the most relevant filters based on their SSVEP responses. This ensures that only neurons with strong, frequency-locked activity contribute to the generated heatmaps, improving interpretability.

SSVEP-Guided CAM Generation: Traditional CAM techniques [30,31,38] compute interpretability maps by applying a weighted sum over all activation maps, where the weighting is determined either by classification-layer coefficients [38] or gradient-based saliency scores [30,31]. A broad formulation of CAM is given by:

$$M_{\text{CAM}}(x, y) = \sum_{k} w_k A_k(x, y), \tag{5}$$

where $A_k(x, y)$ represents the activation map for the k-th convolutional filter, and w_k denotes its corresponding importance weight [38].

However, this approach does not distinguish between relevant and irrelevant filters, potentially introducing noise and reducing interpretability and explainability. Many filters may activate in response to non-discriminative patterns, leading to scattered or misleading heatmaps [22,33]. To address this, we propose an SSVEP-guided filter selection strategy, which enhances interpretability and reduces noise by restricting heatmap computation to the top-K most relevant filters, ranked based on their SNR-derived importance from Eq. 3:

$$M_{\text{SSVEP-CAM}}(x, y) = \sum_{k \in \mathcal{K}} w_k A_k(x, y), \tag{6}$$

where $\mathcal{K}$ is the set containing the indices of the top-K filters identified through SSVEP analysis. By leveraging SSVEP-based filter selection, our approach ensures that heatmaps are more concentrated around salient regions, enhancing interpretability by improving localization and reducing irrelevant activations. Furthermore, by aligning with neuroscience-inspired principles, SSVEP-CAM provides a biologically plausible explainability method.

3.3 Evaluation and Metrics

To validate our method, we conducted various experiments. We compare traditional CAM methods that use all filters against our SSVEP-enhanced CAM approach. To assess the quality of the explanations, we use the following metrics.

Energy Concentration (EC): Quantifies how concentrated the CAM heatmap is in the most salient regions. Similar approaches have been explored in the evaluation of explainability methods, particularly in assessing the complexity and focus of neural network visualizations [29]. A higher EC score indicates that the explanation is more localized, avoiding unnecessary dispersion. It is defined as:

$$EC = \frac{\sum_{i \in S} H_i}{\sum_{j \in \Omega} H_j} \tag{7}$$

where $H_i \in \mathbb{R}^+$ represents the heatmap intensity at pixel i, $S \subset \Omega$ denotes the subset of pixels corresponding to the top 20% highest heatmap intensities, and Ω is the set of all pixels in the heatmap. Specifically, this metric computes the proportion of the total heatmap energy that is concentrated within the most activated 20% of pixels. We expect this to be higher with CAM methods only using top-K filters, as they generate more concentrated attributions.

While we often observe that higher EC scores visually correlate with more spatially localized explanations, it is important to note that the EC metric itself is agnostic to spatial structure. It does not explicitly account for whether high-intensity pixels are spatially contiguous or scattered. Thus, while we observe that higher EC values tend to coincide with more spatially compact heatmaps in our qualitative results, this spatial localization is not guaranteed by the metric formulation alone. To assess spatial coherence more directly, qualitative visualizations are used in conjunction with EC in our evaluation.

Loc-1 and Loc-5 Localization Accuracy: We evaluate the localization accuracy of model explanations using the loc1 (Top-1) and loc5 (Top-5) metrics [6,30,38]. These metrics assess whether the most activated region correctly identifies the target class (loc1) or falls within the top five predicted bounding boxes (loc5).

Given a predicted bounding box $\hat{B}$ and a ground-truth bounding box B, localization accuracy is measured using the Intersection over Union (IoU), defined as [9]:

$$\text{IoU}(\hat{B}, B) = \frac{|B \cap \hat{B}|}{|B \cup \hat{B}|} \tag{8}$$

where IoU $\in [0, 1]$, with higher values indicating greater spatial overlap. Localization is considered successful in loc1 if the highest-scoring predicted bounding box satisfies IoU $\geq \tau$, where τ is a predefined threshold (typically $\tau = 0.5$).

For loc5, at least one of the top five predicted bounding boxes must meet this criterion:

$$\max_{i=1,\ldots,5} \text{IoU}(\hat{B}_i, B) \geq \tau \tag{9}$$

Unlike conventional IoU-based evaluation, which measures spatial overlap between heatmaps and ground-truth annotations [9], loc1 and loc5 emphasize the interpretability of model explanations in classification tasks. These metrics ensure that the highlighted regions correspond to class-relevant areas rather than merely overlapping with object boundaries. For a detailed theoretical foundation of IoU, we refer to [9].

4 Experiments

This section presents a comprehensive evaluation of the SSVEP-enhanced CAM framework. The proposed methodology is compared against well-documented CAM techniques across three widely used architectures: VGG-16 [32], ResNet-50 [13], and ResNeXt-50 [35].

4.1 Experimental Setup

We identify the most relevant convolutional filters using the SSVEP-based frequency tagging method and select the top 100 filters for each model's explanation process. This selective approach ensures that only the most functionally significant filters contribute to the class activation maps. The final convolutional block before the fully connected layers is selected for all three models to generate the explanations as they capture high-level semantic features.

We conduct our experiments on the ImageNet (ILSVRC2012) validation dataset [28]. A subset of 5000 randomly sampled images is used to ensure diverse object classes while maintaining computational feasibility. All models used are pretrained on ImageNet, and are acquired using PyTorch's torchvision library [24]. The inputs are normalized using standard ImageNet mean and standard deviation values.

4.2 Experiment Results

Figure 2 presents visual comparisons between the methods. Across different methods, SSVEP-enhanced CAMs generate more spatially focused explanations by concentrating activation energy in regions most relevant to classification. Notably, EigenCAM shows significant improvements in spatial compactness, which reinforces our hypothesis that selecting functionally significant convolutional filters leads to more concentrated explanations.

Table 1 reports the Energy Concentration scores for each model and CAM method. This metric quantifies how concentrated the activation heatmap is in specific regions. Previous CAM methods assess their methods using conventional

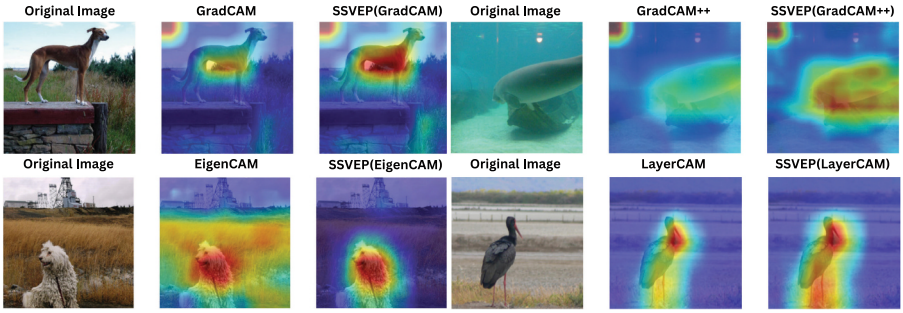

Fig. 2. Comparing baseline CAM methods and their SSVEP-enhanced counterparts.

Table 1. Energy Concentration (↑) (%) for Baseline and SSVEP CAM Methods, reported as **mean ± standard deviation** across 5000 validation samples.

Algorithm	VGG-16		ResNet-50		ResNeXt-50	
	Baseline	SSVEP	Baseline	SSVEP	Baseline	SSVEP
GradCAM	47.50 ± 8.26	**53.91 ± 11.05**	**45.24 ± 6.49**	44.31 ± 7.12	**42.02 ± 5.63**	39.39 ± 5.34
GradCAM++	39.18 ± 4.92	**44.89 ± 10.19**	39.31 ± 4.87	**42.68 ± 7.17**	**39.21 ± 4.06**	38.56 ± 5.04
EigenCAM	62.79 ± 17.33	**69.50 ± 13.52**	49.90 ± 11.48	**54.37 ± 11.68**	41.81 ± 7.28	**44.72 ± 7.60**
LayerCAM	46.63 ± 7.95	**47.09 ± 8.00**	45.66 ± 8.55	**46.48 ± 9.01**	44.27 ± 7.46	**44.75 ± 7.65**

metrics such as general alignment (e.g., IoU [9]) or impact on classification (e.g., ROAD [27]). Energy Concentration provides a direct measure of how focused an explanation is.

SSVEP-enhanced CAMs achieves higher Energy Concentration scores, which indicates that the activation energy is more focused. This confirms that selecting most responsive neurons enhances localization while reducing irrelevant activations. While localization accuracy measure overlap with ground-truth

Table 2. Localization Accuracy (↑) for Baseline and SSVEP CAM Methods.

Algorithm	Prediction Level	VGG-16		ResNet-50		ResNeXt-50	
		Baseline	SSVEP	Baseline	SSVEP	Baseline	SSVEP
GradCAM	loc1	**19.76%**	17.04%	18.52%	**18.92%**	**18.28%**	18.12%
	loc5	**24.50%**	21.16%	22.64%	**23.14%**	**22.18%**	21.86%
GradCAM++	loc1	**17.44%**	16.72%	18.72%	**19.90%**	18.12%	**18.52%**
	loc5	**21.54%**	20.66%	22.80%	**24.22%**	21.62%	**22.32%**
EigenCAM	loc1	**21.28%**	19.00%	25.06%	**26.12%**	21.72%	**22.78%**
	loc5	**25.78%**	22.78%	29.92%	**31.14%**	25.88%	**27.24%**
LayerCAM	loc1	21.84%	**22.00%**	**23.14%**	23.08%	**22.52%**	22.26%
	loc5	26.52%	**26.94%**	**28.16%**	28.06%	**27.04%**	26.82%

annotations, it does not account for the scattering of attributions. A heatmap may achieve a high IoU but it still can distribute energy across irrelevant areas. Therefore, it is important to find a balance between Energy Concentration and localization accuracy.

Table 2 presents localization accuracy results for each model and CAM method. Higher localization accuracy indicates a better match with human-annotated regions, and it does not necessarily reflect how precisely focused the explanation is. This distinction explains why SSVEP-enhanced CAMs do not always achieve higher localization accuracy in every case. However, they remain comparable to their baseline counterparts, and in some cases they even surpass the standard CAM methods despite utilizing only a subset of the all convolutional filters. This reinforces the idea that many activations may be redundant or even introduce noise for visual explanations.

4.3 Discussion

The results show that SSVEP-enhanced CAMs produce more focused and spatially compact explanations across multiple architectures. Higher Energy Concentration scores confirm that these methods reduce scattered attributions. This provides evidence that selecting only the most functionally significant filters improves interpretability by reducing noise from irrelevant activations.

Furthermore, localization accuracy results show that SSVEP-based filtering does not always achieve higher alignment with human-annotated ground truths. While some methods maintain or even improve loc1 and loc5 scores, others experience slight drops, particularly on VGG-16. Importantly, the most improvement in Energy Concentration score is seen on VGG-16 as well. This suggests that optimizing for focus and compactness sometimes comes at the cost of completeness in object localization. The trade-off between Energy Concentration and localization accuracy highlights the need for a balanced approach in evaluating explainability methods.

Variations in results across architectures indicate that model-specific characteristics can influence the impact of SSVEP filtering. For instance, ResNeXt-50 shows a more moderate gain in Energy Concentration compared to VGG-16 and ResNet-50, which suggests that the benefits of filter selection depend on how distributed a model's feature representation is.

Another limitation of our study is the scope of the evaluation dataset. All experiments were conducted on a subset of the ImageNet validation set consisting of 5000 randomly selected images. While this subset ensures class diversity, it does not capture the full variability of the dataset or test cross-dataset generalizability. A more comprehensive evaluation across multiple datasets is needed to robustly assess the general applicability of SSVEP-based filter selection.

It is also important to acknowledge that the three networks differ substantially in capacity: VGG-16 contains significantly more parameters than ResNet-50 and ResNeXt-50. Despite using the same number of top filters across models, this design choice may not normalize for architectural differences, potentially leading to performance disparities. A percentage-based selection strategy—e.g.,

retaining the top $K\%$ of filters per model could offer a fairer baseline across architectures. While this remains an open avenue for future work, we acknowledge that such a strategy may yield more consistent results and help disentangle architecture-specific effects.

These results underscore the importance of Energy Concentration as a complementary metric for evaluating explainability. While existing benchmarks primarily assess interpretability through localization accuracy or classification impact, our findings suggest that the spatial precision of attributions is equally critical. A more focused heatmap reduces ambiguity in model decision-making and offers better insights.

In summary, incorporating additional constraints to balance localization accuracy with Energy Concentration could further improve CAM-based explanations. Exploring different filter selection strategies across architectures may help optimize the trade-off between focus and completeness. These insights provide a strong foundation for improving model interpretability while maintaining practical usability across different architectures and datasets.

5 Conclusion

Our results demonstrate that SSVEP-based filter selection enhances the focus and interpretability of CAM-based explanations without significantly compromising localization accuracy. By concentrating activation energy in the most relevant regions, this method reduces noise and improves the clarity of attributions. While performance gains vary across architectures, the overall improvements in Energy Concentration suggest that selective filter activation is a promising direction for refining explainability methods. Future work could explore dynamic selection mechanisms that adapt to different network architectures and tasks, further optimizing the trade-off between focus and completeness in visual explanations.

References

1. Akhtar, N.: A survey of explainable AI in deep visual modeling: Methods and metrics. arXiv preprint arXiv:2301.13445 (2023)
2. Alp, N., Ozkan, H.: Neural correlates of integration processes during dynamic face perception. Sci. Rep. **12**(1), 118 (2022)
3. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE **10**(7), e0130140 (2015)
4. Boremanse, A., Norcia, A.M., Rossion, B.: An objective signature for visual binding of face parts in the human brain. J. Vision **13**(11), 6–6 (09 2013). https://doi.org/10.1167/13.11.6, https://doi.org/10.1167/13.11.6
5. Böge, E., Gunindi, Y., Aptoula, E., Alp, N., Ozkan, H.: Adapting the biological ssvep response to artificial neural networks (2024). https://arxiv.org/abs/2411.10084

6. 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) (2018)
7. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning (2017). https://arxiv.org/abs/1702.08608
8. Dzhelyova, M., Rossion, B.: Supra-additive contribution of shape and surface information to individual face discrimination as revealed by fast periodic visual stimulation. J. Vision **14**(14), 15–15 (12 2014). https://doi.org/10.1167/14.14.15
9. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. Int. J. Comput. Vision **88**, 303–338 (2010)
10. Ghorbani, A., Zou, J.Y.: Neuron shapley: Discovering the responsible neurons. In: Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020 (2020)
11. Gunning, D., Aha, D.: Darpa's explainable artificial intelligence (xai) program. AI Mag. **40**(2), 44–58 (2019)
12. Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., Yang, G.Z.: Xai—explainable artificial intelligence. Sci. Robot. **4**(37), eaay7120 (2019)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 770–778 (2016)
14. Jiang, P.T., Zhang, C.B., Hou, Q., Cheng, M.M., Wei, Y.: Layercam: exploring hierarchical class activation maps for localization. IEEE Trans. Image Process. **30**, 5875–5888 (2021)
15. Kazmierczak, R., Berthier, E., Frehse, G., Franchi, G.: Explainability for vision foundation models: A survey. arXiv preprint arXiv:2501.12203 (2025)
16. Li, K., Wu, Z., Peng, K.C., Ernst, J., Fu, Y.: Tell me where to look: Guided attention inference network (2018). https://arxiv.org/abs/1802.10171
17. Liu-Shuang, J., Norcia, A.M., Rossion, B.: An objective index of individual face discrimination in the right occipito-temporal cortex by means of fast periodic oddball stimulation. Neuropsychologia **52**, 57–72 (2014). https://doi.org/10.1016/j.neuropsychologia.2013.10.022
18. Lundberg, S., Lee, S.I.: A unified approach to interpreting model predictions (2017). https://arxiv.org/abs/1705.07874
19. Mersha, M., Lam, K., Wood, J., AlShami, A., Kalita, J.: Explainable artificial intelligence: a survey of needs, techniques, applications, and future direction. Neurocomputing p. 128111 (2024)
20. Molchanov, P., Mallya, A., Tyree, S., Frosio, I., Kautz, J.: Importance estimation for neural network pruning. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 11264–11272 (2019)
21. Muhammad, M.B., Yeasin, M.: Eigen-cam: Class activation map using principal components. In: 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1—-7 (Jul 2020)
22. Nielsen, I.E., Dera, D., Rasool, G., Ramachandran, R.P., Bouaynaya, N.C.: Robust explainability: a tutorial on gradient-based attribution methods for deep neural networks. IEEE Signal Process. Mag. **39**(4), 73–84 (2022)
23. Norcia, A.M., Appelbaum, L.G., Ales, J.M., Cottereau, B.R., Rossion, B.: The steady-state visual evoked potential in vision research: A review. J. Vision **15**(6) (2015)

24. Paszke, A., et al.: Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems, pp. 8024–8035 (2019)
25. Regan, D.: An effect of stimulus colour on average steady-state potentials evoked in man. Nature **210**(5040), 1056–1057 (1966)
26. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should i trust you?": Explaining the predictions of any classifier (2016). https://arxiv.org/abs/1602.04938
27. Rong, Y., Leemann, T., Borisov, V., Kasneci, G., Kasneci, E.: A consistent and efficient evaluation strategy for attribution methods (2022). https://arxiv.org/abs/2202.00449
28. Russakovsky, O., et al.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vision **115**, 211–252 (2015)
29. Samek, W., Binder, A., Montavon, G., Lapuschkin, S., Müller, K.R.: Evaluating the visualization of what a deep neural network has learned. IEEE Trans. Neural Netw. Learn. Syst. **28**(11), 2660–2673 (2016)
30. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: visual explanations from deep networks via gradient-based localization. Int. J. Comput. Vision **128**(2), 336–359 (2019)
31. Selvaraju, R.R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., Batra, D.: Grad-cam: Why did you say that? (2017). https://arxiv.org/abs/1611.07450
32. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: 3rd International Conference on Learning Representations, ICLR (2015)
33. Smilkov, D., Thorat, N., Kim, B., Viégas, F., Wattenberg, M.: Smoothgrad: removing noise by adding noise. arXiv preprint arXiv:1706.03825 (2017)
34. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks (2017). https://arxiv.org/abs/1703.01365
35. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1492–1500 (2017)
36. Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., Zhu, J.: Explainable AI: a brief survey on history, research areas, approaches and challenges. In: Tang, J., Kan, M.-Y., Zhao, D., Li, S., Zan, H. (eds.) NLPCC 2019. LNCS (LNAI), vol. 11839, pp. 563–574. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32236-6_51
37. Yu, R., et al.: NISP: pruning networks using neuron importance score propagation. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 9194–9203 (2018)
38. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2921–2929 (2016)