GFR-CAM: Gram-Schmidt Feature Reduction for Hierarchical Class Activation Maps

Kaveh Safavigerdini[1] Bahram Yaghooti[2] Amir Erfan Zareei Shams Abadi[1]
Kannappan Palaniappan[1]

[1]University of Missouri-Columbia, USA [2]Washington University in St. Louis, USA

{ksgh2, azzgg, pal}@missouri.edu {byaghooti}@wustl.edu

Abstract

Deep learning models have achieved remarkable success in computer vision tasks, yet their decision-making processes remain largely opaque, limiting their adoption especially in safety-critical applications. While Class Activation Maps (CAMs) have emerged as a prominent solution for visual explanation, existing methods suffer from a fundamental limitation: they produce single, consolidated explanations leading to "explanatory tunnel vision." Current CAM methods fail to capture the rich, multi-faceted reasoning that underlies model predictions, particularly in complex scenes with multiple objects or intricate visual relationships. We introduce the Gram-Schmidt Feature Reduction Class Activation Map (GFR-CAM), a novel gradient-free framework that overcomes this limitation through hierarchical feature decomposition that provides a more holistic view of the architecture's explanatory power. Unlike existing feature reduction methods that rely on Principal Component Analysis (PCA) and generate a single dominant explanation, GFR-CAM leverages Gram-Schmidt orthogonalization to systematically extract a sequence of orthogonal, information rich components from model feature maps. The subsequent orthogonal components are shown to be meaningful explanations, not mere noise, that decomposes single objects into semantic parts and systematically disentangles multi-object scenes to identify co-existing entities. We show that GFR-CAM applied to ResNet-50 and Swin Transformer architectures across ImageNet and PASCAL VOC datasets achieves competitive performance with state-of-the-art methods.

1. Introduction

Deep architectures such as Convolutional Neural Networks (CNNs) [10, 19], and vision transformers [7, 31] have revolutionized computer vision by achieving remarkable performance on tasks ranging from visual object detection to video object tracking. However, these models [6, 14, 30] often remain opaque "black boxes," providing little insight into the decision-making processes behind their predictions [21, 24, 54]. In safety-critical domains such as medical diagnostics [32, 45], autonomous and control systems [2, 49, 52], and manufacturing [42, 55], this lack of interpretability remains a significant barrier, driving the need for methods that can reliably explain the inner workings of these complex models. We were motivated by our own applications that could benefit from improved deep architecture explainability, including malaria infection red blood cell segmentation [46], histopathology [3], aerial vehicle tracking [28, 39], vessel segmentation [48], and analysis of carbon nanotube properties [37, 38, 44] and growth rates [35, 40].

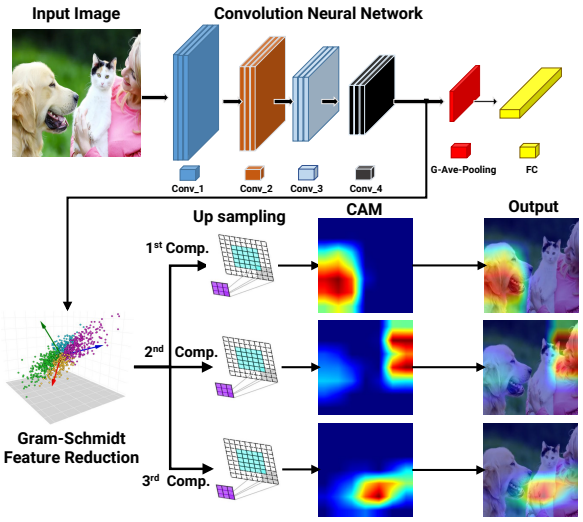


Figure 1. GFR-CAM integrated within a CNN. GFR-CAM uses Gram-Schmidt orthogonalization to generate a hierarchical explanation, sequentially identifying the primary object (Component 1) and distinct secondary objects (Components 2 and 3) to provide a disentangled hierarchical view of the model's explainability.

Class Activation Maps (CAM) [54] have emerged as a leading approach to this challenge, visually highlighting image regions that influence a model's output [15, 39]. While early methods relied on gradients (e.g., Grad-CAM[41]), a new class of more robust, gradient-free techniques has gained prominence. Methods like Eigen-CAM [27] and Kernel-PCA CAM (KPCA-CAM) [18] leverage powerful decomposition techniques such as Principal Component Analysis (PCA) [22] and its nonlinear extension, Kernel PCA [25]. By directly analyzing convolutional feature maps, these methods produce clean explanations without gradient-related noise and vanishing gradient issues.

However, despite their advancements, these state-of-the-art decomposition methods [1, 11, 26, 43] share a fundamental limitation: they are designed to produce a *single, consolidated explanation*. By projecting all feature map information onto a single principal component, they generate a monolithic heatmap that captures only the most dominant evidence for a prediction. This "winner-takes-all" approach creates a form of explanatory tunnel vision. In multi-object scenes, it highlights one object while ignoring others that may be relevant to the model's overall understanding. For single-object images, it obscures the rich "visual subtext"—crucial secondary features like texture, context, or distinct object parts—that are essential for a complete and robust interpretation of the model's reasoning.

This limitation is a direct consequence of the feature reduction techniques employed. PCA and its variants are designed to find the single direction of maximum variance; subsequent components, by definition, capture progressively less information and often represent noise rather than meaningful, distinct concepts. The core challenge, therefore, is not to simply find the best single explanation, but to *decompose the feature space into a hierarchy of interpretable parts*. To achieve this, we move away from PCA and turn to a different, theoretically-grounded decomposition method: *Gram-Schmidt orthogonalization* [50, 51, 53]. Unlike PCA, the Gram-Schmidt process is explicitly designed to construct a set of linearly independent basis vectors. This inherent property makes it an ideal candidate for systematically extracting a sequence of distinct, information-rich, and non-redundant components from the feature space, paving the way for a multi-faceted explanation.

Building on this foundation, we introduce the *Gram-Schmidt Feature Reduction Class Activation Map (GFR-CAM)*, a novel framework that leverages Gram-Schmidt orthogonalization to deconstruct a model's reasoning into a hierarchy of visual evidence. As illustrated in Figure 1, GFR-CAM analyzes the final convolutional feature maps to generate a sequence of activation maps. The first component reliably identifies the primary evidence, achieving performance competitive with state-of-the-art methods. However, the true innovation lies in the subsequent components. Because the Gram-Schmidt process isolates linearly independent and information-rich features, the second, third, and further activations are not noise; they are meaningful explanations that solve the tunnel vision problem. In single-object images, they isolate supporting evidence like texture and distinct parts. In multi-object scenes, they systematically disentangle the visual field, identifying co-existing objects one by one to reveal a comprehensive scene understanding that monolithic CAMs cannot provide.

Our contributions are threefold:

- We introduce GFR-CAM, a novel, gradient-free, and computationally efficient CAM framework built on Gram-Schmidt orthogonalization, which is theoretically suited for hierarchical feature decomposition.
- We demonstrate that GFR-CAM overcomes the single-explanation limitation of existing methods by generating a hierarchy of interpretable activation maps. We show that its subsequent components are not noise but instead reveal secondary features, contextual clues, and distinct objects in complex scenes.
- Through extensive quantitative and qualitative evaluation across multiple models and rigorous benchmarks, we prove that GFR-CAM provides a more complete and disentangled understanding of a model's decision-making, particularly in complex, multi-object scenes, thereby setting a new standard for comprehensive AI explanation.

2. Methodology

In this section, we present the Gram-Schmidt Class Activation Maps (GFR-CAM) framework. Our method enhances class activation maps for interpretability using Gram-Schmidt orthogonalization and feature reduction [50, 53]. This approach identifies discriminative image regions with higher precision and reduced redundancy, revealing the model's hierarchical visual evidence.

The overall GFR-CAM workflow is summarized as follows (see Figure 1). The key innovation is in Step 3, where hierarchical, orthogonal components are generated to produce a sequence of interpretable activation maps:

1. **Feature Extraction:** Extract the feature maps $\{A^k\}_{k=1}^{K}$ from the final convolutional layer of a pre-trained CNN or Vision Transformer (ViT) for a given input image.
2. **Data Preparation:** Reshape the feature maps into a matrix $F \in \mathbb{R}^{K \times L}$, where K is the number of channels and $L = H \times W$ is the product of the spatial dimensions.
3. **Hierarchical Component Generation:** Apply the Gram-Schmidt Functional Reduction (GFR) procedure (detailed in Section 3.2) to the feature matrix F. This iteratively computes a set of orthogonal direction vectors $\{\nu_1, \nu_2, \ldots, \nu_m\}$ and their corresponding activation maps $\{M_1, M_2, \ldots, M_m\}$.

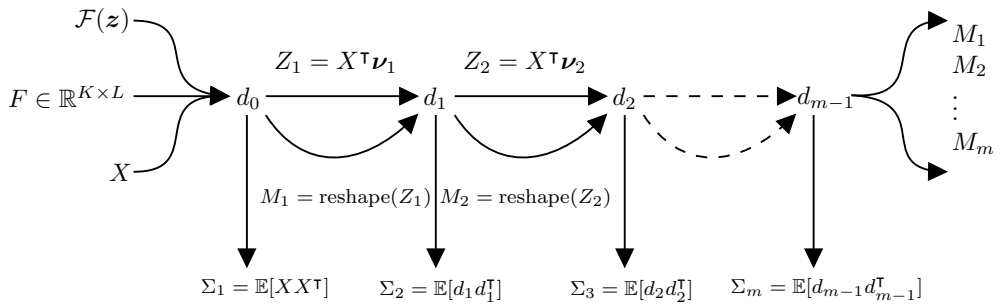


Figure 2. Schematic overview of the GFR-CAM framework. It begins with a family of functions and independent CNN feature samples, iteratively replacing fixed parameters with linear data projections, orthogonalizing the functions, and subtracting their contributions to obtain a new random variable $d_j = X - \sum_{f\in\hat{\mathcal{F}}(Z_1,\ldots,Z_j)} \mathbb{E}[Xf]f$ at iteration j. Each d_j yields a covariance matrix $\Sigma_{j+1} = \mathbb{E}[d_j(X)d_j(X)^\intercal]$, whose largest eigenvector ν_j indicates the direction of maximum variance for the next feature to be extracted [50, 53].

4. **Post-processing:** Normalize each M_j to [0,1], then upsample to original image size via bilinear interpolation.
5. **Visualization:** Overlay each processed map onto the original image to generate sequential explanations.

In what follows, we first describe the foundational orthogonalization algorithm before detailing how it is employed within the GFR procedure to generate our hierarchical GFR-CAMs.

2.1. Gram-Schmidt Orthogonalization: Core Mechanism

The foundation of our method is the Gram-Schmidt process, which we use to systematically decompose the CNN/Vit's feature space. This process is formalized in Algorithm 1. At each iteration j, the algorithm takes a new feature component Z_j and makes it orthogonal to all previously computed components. This is achieved by subtracting its projection onto the already-orthogonalized function set $\mathcal{T}$ (Line 4) and then normalizing the result (Line 5).

This process can be concisely represented by the classical Gram-Schmidt projection formula. For a new component Z_j and a set of existing orthonormal components $\{\hat{Z}_i\}_{i=1}^{j-1}$, the orthogonalization is:

$$\tilde{Z}_j = Z_j - \sum_{i=1}^{j-1} \langle Z_j, \hat{Z}_i\rangle \hat{Z}_i, \quad \text{and} \quad \hat{Z}_j = \frac{\tilde{Z}_j}{\|\tilde{Z}_j\|} \tag{1}$$

where $\tilde{z}_j$ is the orthogonalized component and $\hat{Z}_j$ is its normalized version. A critical output of Algorithm 1 is the residual covariance matrix Σ_{j+1} (Line 9). This matrix captures the variance structure of the data *after* the information from the first j components has been removed, which is essential for finding the next most important feature direction.

2.2. Hierarchical Activation Map Generation with GFR

We operationalize the orthogonalization process using Gram-Schmidt Functional Reduction (GFR), as detailed in Algorithm 2, to produce our sequence of activation maps.

Algorithm 1: Orthogonalize$(\hat{\mathcal{F}}(Z_1,\ldots,Z_{j-1}), \mathcal{F}(\mathbf{z}_{[j]})\setminus \mathcal{F}(\mathbf{z}_{[j-1]}), \{Z_i\}_{i=1}^{j})$

Input:
- $\hat{\mathcal{F}}(Z_1,\ldots,Z_{j-1})$: an orthogonalized function family in random variables $Z_1,\ldots,Z_{j-1}$.
- $\mathcal{F}(z_{[j]}) \setminus \mathcal{F}(z_{[j-1]})$: all the functions in $\mathcal{F}(\mathbf{z})$ which depend on $\mathbf{z}_{[j]}$ but not only on $\mathbf{z}_{[j-1]}$.
- $\{Z_i\}_{i=1}^{j}$: the random variables $Z_1,\ldots,Z_{j-1}$.
- Z_j: a new random variable.

Output:
- Orthogonalized function family $\hat{\mathcal{F}}(Z_1,\ldots,Z_j)$
- A covariance matrix Σ_{j+1}.

1 **Initialize:** $\mathcal{T} \leftarrow \hat{\mathcal{F}}(Z_1,\ldots,Z_{j-1})$. Denote $\mathcal{F}(z_{[j]}) \setminus \mathcal{F}(z_{[j-1]})) \triangleq \{f_1(z),\ldots,f_\ell(z)\}$, where $f_1(z) = z_j$.
2 **for** $k \leftarrow 1$ **to** ℓ **do**
3 $\quad g(Z_1,\ldots,Z_j) \leftarrow f_k(Z_1,\ldots,Z_j)$.
4 $\quad g(Z_1,\ldots,Z_j) \leftarrow g(Z_1,\ldots,Z_j) - \sum_{f\in\mathcal{T}} \mathbb{E}[gf]f$;
5 $\quad g(Z_1,\ldots,Z_j) \leftarrow \frac{g(Z_1,\ldots,Z_j)}{\sqrt{\mathbb{E}[g(Z_1,\ldots,Z_j)^2]}}$.
6 $\quad \mathcal{T} \leftarrow \mathcal{T} \cup \{g(Z_1,\ldots,Z_j)\}$.
7 Define $\hat{\mathcal{F}}(Z_1,\ldots,Z_j) = \mathcal{T}$.
8 Define $d_j(X) = X - \sum_{f\in\mathcal{T}} \mathbb{E}[Xf]f$.
9 Define $\Sigma_{j+1} = \mathbb{E}[d_j(X)d_j^\intercal(X)]$.
10 **Return** Σ_{j+1}, $\mathcal{T}$.

The GFR algorithm iteratively identifies orthogonal components that maximize variance. At each step j, it finds the principal eigenvector ν_j of the current (residual) covariance matrix Σ_j. This eigenvector represents the direction of maximum variance in the information that has not yet been explained by previous components.

Algorithm 2: Gram-Schmidt Functional Reduction (GFR) for GFR-CAM

Input: Random variable X taken from feature maps $F \in \mathbb{R}^{K\times(H,W)}$, function family $\mathcal{F}(z)$, and threshold $\epsilon > 0$.

Output: Orthogonal activation maps $\{M_j\}_{j=1}^{m}$.

Initialize: $\Sigma_1 \leftarrow \mathbb{E}\big[XX^\intercal\big]$.

for $j \leftarrow 1$ **to** K **do**

  $\nu_j \leftarrow \arg\max_{\|\nu\|=1} \nu^\intercal \Sigma_j \nu$;

  if $\nu_j^\intercal \Sigma_j \nu_j \leq \epsilon^2$ **then**

    break.

  $Z_j \leftarrow X^\intercal \nu_j$.

  $M_j \leftarrow \text{reshape}(Z_j, H, W)$.

  $\Sigma_{j+1}, \hat{\mathcal{F}}(Z_{[j]}) \leftarrow$ Orthogonalize$\big(\Sigma_j, \hat{\mathcal{F}}(Z_{[j-1]}), \mathcal{F}(\mathbf{z}_{[j]}) \setminus \mathcal{F}(\mathbf{z}_{[j-1]}), \{Z_i\}_{i=1}^{j}\big)$.

Return $\{M_j\}_{j=1}^{m}$

Primary Activation Map (M_1): The first activation map, analogous to the output of Eigen-CAM, is generated by projecting the feature maps F onto the principal eigenvector ν_1 of their covariance matrix $\Sigma_1 = \mathbb{E}[XX^\intercal]$. This map, M_1, highlights the most dominant visual evidence for the model's prediction:

$$M_1 = F^\intercal \nu_1 \tag{2}$$

Subsequent Activation Maps (M_j): Beyond the primary map, subsequent maps are generated by iteratively applying the Gram Schmidt algorithm. For each new map M_j, the algorithm first projects out the information captured by previous components $(\nu_1, \ldots, \nu_{j-1})$. It then finds the principal eigenvector ν_j of the remaining data. This vector is guaranteed to be orthogonal to its predecessors and thus highlights a new, independent facet of the model's reasoning, visualized as:

$$M_j = F^\intercal \nu_j \tag{3}$$

This iterative decomposition continues until the residual variance is negligible (below ϵ), providing a richer, multi-faceted explanation that goes beyond the single focus of existing methods.

3. Experiment Results

Our comprehensive evaluation is twofold. First, we establish GFR-CAM's primary component as a competitive baseline by benchmarking it against state-of-the-art methods across diverse architectures and metrics, including ROAD. Second, and more critically, we demonstrate its unique explanatory power by analyzing its subsequent components. We show qualitatively and quantitatively how GFR-CAM surpasses other decompositional methods in providing richer, more disentangled explanations, particularly for complex scenes. The results confirm that GFR-CAM is not only a competitive general-purpose CAM but also offers a superior method for generating multi-faceted, interpretable insights into a model's reasoning.

3.1. Experimental Setup

Models and Datasets. We conduct a comprehensive evaluation on two representative architectures: ResNet-50 [13], a standard CNN, and the Swin Transformer [20]. For ResNet-50, feature maps are extracted from the final convolutional block (`layer4`), while for Swin-T, we use the output of the last normalization layer in the final stage. Our evaluation is performed on two standard benchmarks: a 5,000-image subset of the ImageNet (ILSVRC2012) validation set [36] (1,000 classes) and a 4,000-image subset from PASCAL VOC 2012 [9] (20 classes). For preprocessing, all images are resized and center-cropped to 224×224 pixels and normalized using the standard ImageNet mean and STD.

Evaluation Metrics. To assess the quality of the generated CAMs, we employ a standard suite of metrics that measure an explanation's faithfulness and interpretability [5, 17, 29]. Faithfulness is measured by quantifying the change in model confidence as regions identified by the CAM are manipulated. Specifically, we use the following six established metrics.

Average Drop (AD) [5] measures the average percentage drop in confidence when only the region highlighted by the CAM is provided as input. A lower AD is better, indicating the map preserves class-discriminative features:

$$\text{AD} = \frac{1}{N}\sum_{i=1}^{N} \frac{\max(0, y_i^c - o_i^c)}{y_i^c} \times 100 \tag{4}$$

Here, y_i^c and o_i^c are the model's confidence scores for class c on the original image and the explanation-preserved image, respectively.

Coherency (Coh) [29] measures the stability of an explanation by calculating the Pearson correlation between the original heatmap and a heatmap generated from the explanation itself. Higher values indicate more consistent explanations:

$$\text{Coh}(x) = \frac{1}{2}\frac{\text{Cov}\Big(H_c\big(x \otimes H_c(x)\big),\, H_c(x)\Big)}{\sigma_{H_c\left(x\otimes H_c(x)\right)}\,\sigma_{H_c(x)}} + \frac{1}{2} \tag{5}$$

where $H_c(\cdot)$ is the CAM generation function, $\otimes$ is element-wise multiplication, and $\text{Cov}(\cdot,\cdot)$ denotes the covariance between two random variables.

Complexity (Com) [29] measures the sparsity of an explanation via its L_1 norm. Lower complexity signifies a more focused and simpler explanation:

$$\text{Com}(x) = ||H_c(x)||_1$$

Average Drop, Coherency, and Complexity (ADCC) [29] is a unified metric combining the above via their harmonic mean. A higher ADCC score indicates a better overall balance of explanation qualities:

$$\text{ADCC} = 3\left(\frac{1}{1-\text{AD}} + \frac{1}{\text{Coh}} + \frac{1}{1-\text{Com}}\right)^{-1} \quad (6)$$

Increase in Confidence (IC) [5] measures the percentage of images for which the model's confidence increases when shown only the CAM-highlighted region:

$$\text{IC} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(y_i^c < o_i^c) \times 100 \quad (7)$$

Here, $\mathbb{I}(\cdot)$ is the indicator function, and y_i^c and o_i^c are the confidences for the original and explanation-preserving images, respectively.

Average Drop in Deletion (ADD) [17] measures the drop in model confidence when the CAM-highlighted region is removed from the image. A higher ADD is better, as it indicates the removed region was critical to the prediction:

$$\text{ADD} = \frac{1}{N}\sum_{i=1}^{N} \frac{\max\left(0,\, y_i^c - d_i^c\right)}{y_i^c} \times 100 \quad (8)$$

Here, y_i^c is the original confidence and d_i^c is the confidence after deleting the highlighted pixels.

3.2. Evaluation of the Primary GFR-CAM Component against General CAM Baselines

In this section, we validate the effectiveness of our GFR-CAM by evaluating its *primary component*, which captures the most salient features identified by the model. We benchmark its performance against a comprehensive set of state-of-the-art CAM methods. These include gradient-based approaches like Grad-CAM [41], Grad-CAM++ [5], and XGradCAM [12]; and other notable techniques including LayerCAM [16], HiResCAM [8], KPCA CAM [18], and ShapleyCAM [4]. The comparison is conducted on two distinct architectures: ResNet-50 on a subset of the ImageNet validation set and the Swin Transformer on a subset of PASCAL VOC. The quantitative results demonstrate that GFR-CAM's *primary component* achieves performance that is highly competitive with these leading baselines.

Qualitative Evaluation of Visual Explanations: In our qualitative analysis, we employ ResNet-50 as the backbone network and generate visual explanations using its final convolutional layer. As depicted in Figure 3, GFR-CAM produces the most comprehensive activation maps in comparison to other methods. Grad-CAM struggles to identify the main target objects, often highlighting irrelevant regions, while both KPCA-CAM and Grad-CAM++ correctly localize the objects. However, our method exceeds these approaches in accurately pinpointing the relevant regions.

Performance on ResNet-50: Table 1 presents the results on the ResNet-50 architecture with ImageNet data. Our GFR-CAM demonstrates excellent faithfulness to the model's decision-making, achieving the best scores in Average Drop (11.17), Increase in Confidence (43.31), and ADD (45.57). This indicates that the primary GFR-CAM component is highly effective at identifying the most critical regions for the model's prediction. While Shapley-based methods excel in Coherency and ADCC, our method provides a competitive performance on the core faithfulness metrics, establishing it as a highly competitive CAM.

Table 1. Quantitative comparison of CAM methods on ResNet-50: leveraging the final convolutional layer

Method	AD ↓	Coh ↑	Com ↑	ADCC ↑	IC ↑	ADD ↑
GradCAM	12.73	88.12	40.13	67.22	41.97	44.19
GradCAM++[5]	13.79	96.49	**48.22**	71.23	42.85	37.02
HiResCAM	14.35	61.74	20.49	5.07	40.52	43.71
ShapleyCAM-E	12.28	**98.43**	41.27	82.59	39.33	36.23
KPCA CAM	13.98	90.21	35.67	16.23	40.01	41.11
ShapleyCAM	11.75	89.36	39.58	**83.74**	38.29	43.57
GFRCAM (Ours)	**11.17**	91.74	30.91	20.27	**43.31**	**45.57**

Performance on Swin Transformer. In Table 2, the evaluation on the Swin Transformer with PASCAL VOC data further confirms the robustness of GFR-CAM on non-CNN architectures. Our method achieves top performance in Complexity (32.81) and ADD (27.43), producing maps that are both informative and sparse. It also remains highly competitive in Coherency (62.17), closely following the leading method, ShapleyCAM. GFR-CAM demonstrates robust performance, achieving state-of-the-art results in Complexity and ADD while remaining competitive in other key metrics like Coherency. This solidifies its efficacy as a general-purpose explanation method, applicable to diverse model architectures.

Table 2. Quantitative comparison of CAM methods on the Swin Transformer: targeting the initial normalization layer in the final transformer block

Method	AD ↓	Coh ↑	Com ↑	ADCC ↑	IC ↑	ADD ↑
GradCAM	57.76	41.38	23.77	6.16	5.45	23.17
GradCAM++	68.15	49.67	25.03	3.65	12.42	24.85
XGradCAM[12]	83.73	50.48	31.63	4.06	21.71	19.58
LayerCAM[16]	71.28	58.94	28.26	5.17	–	–
HiResCAM	82.54	51.48	17.25	6.19	**25.12**	21.45
ShapleyCAM	67.22	**64.93**	30.20	**7.49**	17.42	27.18
GFRCAM (Ours)	**72.39**	62.17	**32.81**	6.53	13.82	**27.43**

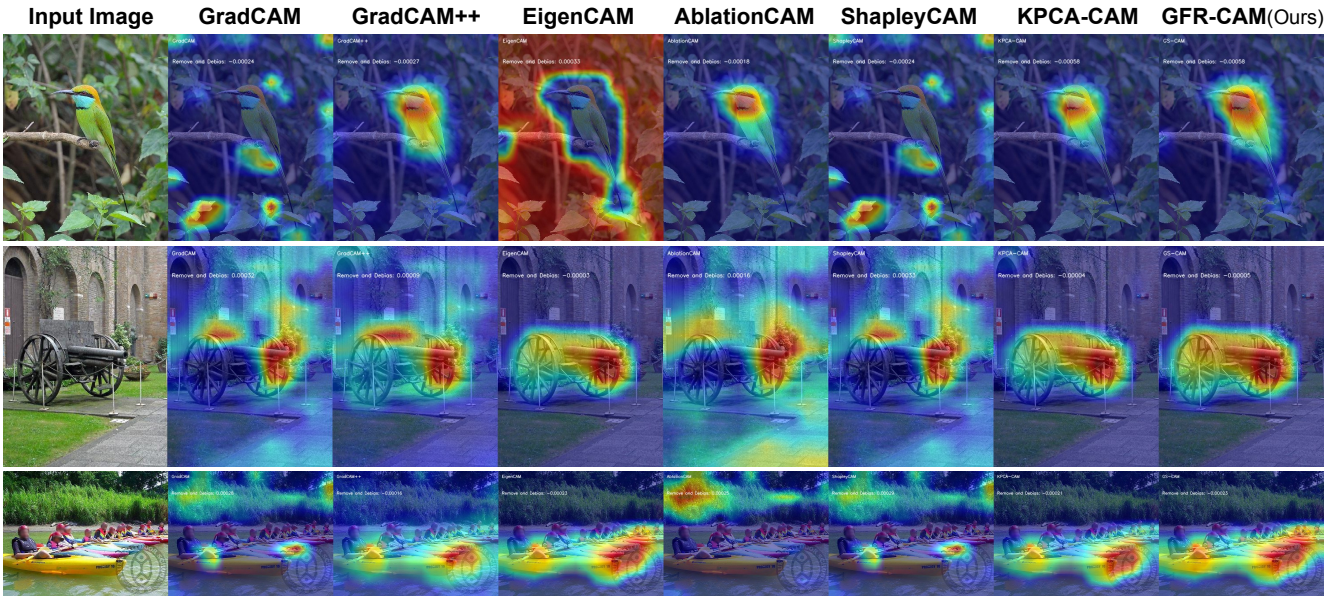


Figure 3. Qualitative comparison of CAM methods on a ResNet-50 model. The figure demonstrates that our proposed GFR-CAM generates more precise and comprehensive activation maps, accurately localizing target objects compared to other leading methods.

RemOve And Debias (ROAD) Benchmark [34]: The ROAD framework provides an efficient evaluation strategy for attribution methods by measuring how well saliency maps identify relevant features. The metric is defined as:

$$\text{ROAD} = \frac{1}{2}\left(\text{LeRF} - \text{MoRF}\right), \tag{9}$$

where LeRF and MoRF represent model confidence scores after perturbing least and most relevant features respectively. ROAD employs a perturbation scheme that removes features in order of importance (MoRF: Most Relevant First; LeRF: Least Relevant First) and uses linear imputation with neighbor-weighted interpolation to fill perturbed regions. A more effective attribution map will identify features that cause a rapid degradation in performance, resulting in a lower Area Under the Curve (AUC) of the performance curve. Therefore, for this benchmark, a lower score is better, signifying a more accurate identification of crucial features.

Our quantitative evaluation using the ROAD benchmark is presented in Table 3. The results highlight the exceptional performance of GFR-CAM, particularly on CNN architectures. Our method achieves state-of-the-art (lowest) scores on VGG-16 (59.38), ResNet-50 (-58.7), and DenseNet-161 (13.11), demonstrating its superior ability to pinpoint the most influential features in convolutional models. While AblationCAM [33] is specialized for and performs best on the transformer-based DeiT-Base, GFR-CAM's strong and consistent results across multiple backbones confirm its robustness. The qualitative visualization in Figure 4 illustrates why GFR-CAM excels: it identifies a compact and semantically critical region, whose removal causes significant information loss and thus explains the rapid performance drop measured by the ROAD metric.

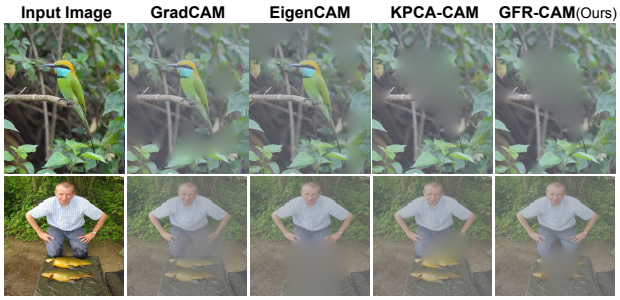
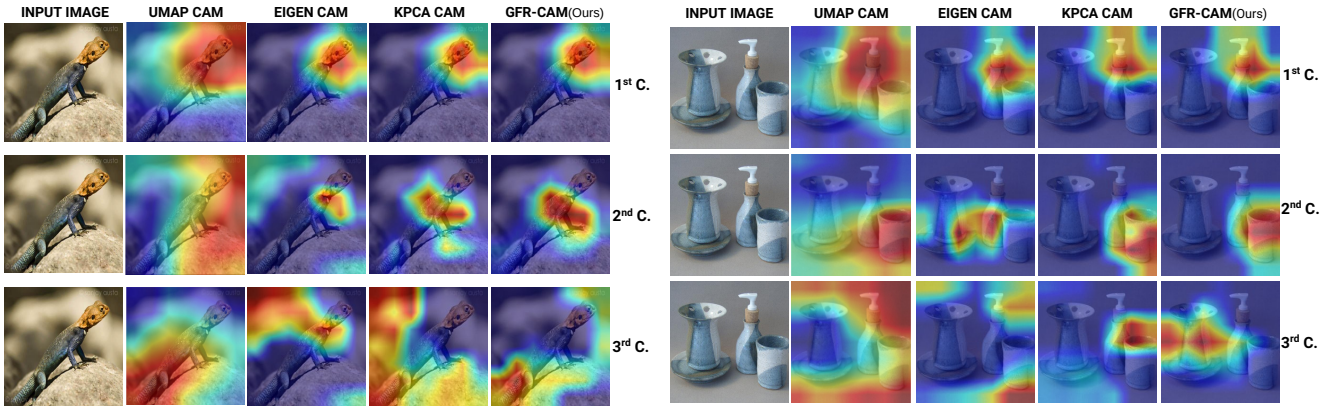


Figure 4. Qualitative results on ROAD. GFR-CAM identifies compact critical regions, yielding greater performance drops under blurring and superior quantitative scores.

3.3. Beyond the First Component: The Explanatory Power of GFR-CAM

A key limitation of existing decompositional CAMs like EigenCAM, KPCA-CAM, and UMAP-CAM [23] is their reliance on dimensionality reduction techniques that prioritize variance preservation, where subsequent components capture progressively less variance and often degrade into noise. Our GFR-CAM, however, uses Gram-Schmidt orthogonalization, which we hypothesize does not merely rank features by importance but instead isolates distinct, semantically meaningful concepts. This section validates this core advan-



(a) (b)

Figure 5. GFR-CAM component analysis. (a) Single object decomposition: Our method's 2nd and 3rd components isolate distinct semantic parts (e.g., lizard's head vs. hand/tail), providing finer-grained explanations than UMAP CAM, EigenCAM and KPCA-CAM. (b) Multiple Objects: It disentangles complex scenes by separating distinct object instances, while other methods merge or fail to separate them.

Table 3. ROAD benchmark scores (lower is better, ↓) across various architectures. GFR-CAM achieves state-of-the-art performance on CNN-based models. All values are $\times 10^2$.

Model	VGG-16	ResNet50	DenseNet-161	DeiT-Base
GradCAM	80.23	-24.3	67.69	5.85
GradCAM++	82.34	-27.8	82.63	9.84
HiResCAM	65.73	3.14	85.40	16.63
ScoreCAM [47]	—	12.1	15.23	10.10
AblationCAM[33]	68.51	-18.3	32.45	**3.89**
EigenCAM	75.23	-16.2	23.39	11.76
ShapleyCAM	87.57	-24.2	86.42	—
KPCA-CAM	82.03	-58.3	85.32	12.51
GFR-CAM (Ours)	**59.38**	**-58.7**	**13.11**	10.17

tage, demonstrating that GFR-CAM's subsequent components provide richer, more structured explanations than their dimensionality reduction-based counterparts, effectively decomposing a model's reasoning into a set of interpretable parts.

The qualitative results visually confirm this hypothesis. As shown in Figure 5a, when applied to a single object, GFR-CAM's components successfully partition the object into its constituent parts, such as a lizard's head and limbs, while competing methods produce diffuse or redundant maps. This ability extends to more complex scenarios, as seen in Figure 5b, where GFR-CAM effectively disentangles a multi-object scene by assigning each of its primary components to a distinct object instance, a task where other methods struggle and often merge the explanations.

This visual superiority is substantiated by the quantitative analysis in Table 4. For both the 2nd and 3rd components, GFR-CAM consistently outperforms EigenCAM and KPCA-CAM, leading in the majority of metrics. Notably, it achieves the best faithfulness scores (lowest Average Drop) while maintaining high coherency. While the performance of PCA-based methods degrades sharply for the 3rd component, GFR-CAM remains robust, proving its unique capability to generate a stable and high-fidelity set of orthogonal explanations beyond the primary one.

Table 4. Quantitative comparison of CAM methods using the Swin-T model for the 2nd and 3rd extracted components. We evaluate our proposed GFR-CAM against EigenCAM and KPCA-CAM across six different metrics.

2nd Component

Method	AD ↓	Coh ↑	Com ↑	ADCC ↑	IC ↑	ADD ↑
EigenCAM	91.31	44.61	21.61	4.74	10.95	21.73
KPCA-CAM	84.69	**59.03**	**22.45**	5.92	11.85	24.93
GFRCAM (Ours)	**80.43**	58.46	22.11	**6.14**	**12.57**	**26.03**

3rd Component

Method	AD ↓	Coh ↑	Com ↑	ADCC ↑	IC ↑	ADD ↑
EigenCAM	94.23	39.27	12.45	3.56	7.16	10.30
KPCA-CAM	89.76	43.79	15.70	4.29	**10.44**	17.44
GFRCAM (Ours)	**87.21**	**46.51**	**16.40**	**4.41**	9.91	**18.63**

3.4. Ablation Study

To validate the design choices of GFR-CAM, we conducted an ablation study on two key hyperparameters: the function family used for orthogonalization and the number of components to generate.

Choice of Orthogonalization Function. The computational cost of GFR-CAM is influenced by the complexity of the function family, $f(z)$, used in the Gram-Schmidt process. Following the methodology in [50], we evaluated linear functions (e.g., $f_1(z) = z_i$ for $i \in [d]$, polynomial functions up to degree 2 ($f_2(z) = z_i z_j$, for $i, j \in [d]$), and polynomial

functions up to degree 3 ($f_3(z) = z_i z_j z_k$, for $i, j, k \in [d]$), where d is the number of features in the feature map. polynomials. As summarized in Table 5, the linear option achieves the best trade-off between runtime and attribution quality, so we adopt $f_1(z)$.

Table 5. Computational cost comparison (seconds) for different orthogonalization functions in GFR-CAM.

Family Function	$f_1(z) = z_i$	$f_2(z) = z_i z_j$	$f_3(z) = z_i z_j z_k$
GFR-CAM($f(z)$)	0.45	0.91	2.21

Number of Components. We also analyzed the trade-off between the number of generated components (m) and their explanatory value. We found that the 2nd and 3rd components consistently provided valuable, distinct visual evidence, as detailed in Section 3.3. However, components beyond the third ($m \geq 4$) introduced significant computational latency while failing to provide a commensurate improvement in interpretable information; these higher-order maps often highlighted redundant features. Based on this analysis, we determined that focusing on the first three components offers the most effective compromise between detailed multi-faceted explanations and practical efficiency.

4. Conclusions and Future Work

We introduced GFR-CAM, a novel gradient-free framework that addresses the fundamental limitation of existing Class Activation Maps: their inability to provide comprehensive, multi-faceted explanations of CNN decision-making. By leveraging Gram-Schmidt orthogonalization instead of PCA-based decomposition, GFR-CAM generates hierarchical activation maps where each component captures distinct, meaningful visual evidence rather than progressively noisier approximations.

Our extensive evaluation demonstrates that GFR-CAM's primary component achieves state-of-the-art performance across multiple architectures and benchmarks, while additional subsequent components provide additional interesting explanatory power. We showed that these additional maps are not artifacts but genuine insights that decompose single objects into semantic parts and systematically disentangle complex multi-object scenes. This capability represents a significant advancement over existing methods that suffer from "explanatory tunnel vision."

The theoretical foundation of Gram-Schmidt orthogonalization ensures that each component is linearly independent and information-rich, making GFR-CAM particularly valuable for safety-critical applications where comprehensive understanding of model reasoning is essential. Several promising directions emerge from this work. First, extending GFR-CAM to other vision architectures, including newer transformer variants and hybrid CNN-transformer models, could further validate its generalizability. Second, investigating adaptive component selection mechanisms could automatically determine the optimal number of meaningful components for different image complexities, potentially improving efficiency.

The integration of GFR-CAM with interactive visualization tools presents opportunities for enhanced human-AI collaboration, particularly in medical imaging and autonomous systems where detailed explanatory feedback is crucial. Additionally, exploring temporal extensions for video understanding and applying the hierarchical decomposition principle to other modalities (e.g., NLP) could broaden the impact of our orthogonalization-based approach.

Finally, investigating the theoretical connections between Gram-Schmidt decomposition and other interpretability methods, as well as developing quantitative metrics specifically designed to evaluate multi-component explanations, would strengthen the foundation for next-generation explainable AI systems that move beyond single-focus explanations toward comprehensive explainable model understanding.

Acknowledgments

This work was partially supported by awards from the U.S. National Science Foundation (CMMI-2026847) and NIH National Institute of Neurological Disorders and Stroke (R01NS110915). Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the U. S. Government or agency thereof.

References

[1] Leila Arras, Bruno Puri, Patrick Kahardipraja, Sebastian Lapuschkin, and Wojciech Samek. A close look at decomposition-based XAI-methods for transformer language models. *arXiv preprint arXiv:2502.15886*, 2025. 2

[2] Shahin Atakishiyev, Mohammad Salameh, Hengshuai Yao, and Randy Goebel. Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions. *IEEE Access*, 2024. 1

[3] Rina Bao, Yunxin Zhao, Akhil Srivastava, Shellaine Frazier, and Kannappan Palaniappan. Lightweight sdenet fusing model-based and learned features for computational histopathology. *IEEE Journal of Selected Topics in Signal Processing*, 18(6):1123–1137, 2024. 1

[4] Huaiguang Cai. CAMs as Shapley value-based explainers. *arXiv preprint arXiv:2501.06261*, 2025. 5

[5] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847, 2018. 4, 5

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding, 2018. 1

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 1

[8] Rachel Lea Draelos and Lawrence Carin. Use HiResCAM instead of Grad-CAM for faithful explanations of convolutional neural networks. *arXiv preprint arXiv:2011.08891*, 2020. 5

[9] Mark Everingham, Luc Van Gool, Christopher K.I. Williams, John Winn, and Andrew Zisserman. The PASCAL visual object classes (VOC) challenge. *IJCV*, 2010. 4

[10] Ramy Farag, Parth Upadhay, Jacket Dembys, Yixiang Gao, Katherin Garces Montoya, Seyed Mohamad Ali Tousi, Gbenga Omotara, and Guilherme Desouza. Efficientnet-sam: A novel effecientnet with spatial attention mechanism for covid-19 detection in pulmonary ct scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5200–5206, 2024. 1

[11] Qizhang Feng, Ninghao Liu, Fan Yang, Ruixiang Tang, Mengnan Du, and Xia Hu. DEGREE: Decomposition based explanation for graph neural networks. In *10th Int. Conf. Learning Representations (ICLR)*, 2022. 2

[12] Ruigang Fu, Qingyong Hu, Xiaohu Dong, Yulan Guo, Yinghui Gao, and Biao Li. Axiom-based Grad-CAM: Towards accurate visualization and explanation of CNNs. In *British Machine Vision Conference (BMVC)*, 2020. 5

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 4

[14] Hamidreza Montazeri Hedesh, Moh Wafi, Bahram Shafai, Milad Siami, et al. Robust stability analysis of positive lure system with neural network feedback. *arXiv preprint arXiv:2505.18912*, 2025. 1

[15] Habib Irani and Vangelis Metsis. Enhancing time-series prediction with temporal context modeling: A Bayesian and deep learning synergy. In *The International FLAIRS Conference Proceedings*, 2024. 2

[16] Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. LayerCAM: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 30:5875–5888, 2021. 5

[17] Hyungsik Jung and Youngrock Oh. Towards better explanations of class activation mapping. In *ICCV*, 2021. 4, 5

[18] Sachin Karmani, Thanushon Sivakaran, Gaurav Prasad, Mehmet Ali, Wenbo Yang, and Sheyang Tang. KPCA-CAM: Visual explainability of deep computer vision models using Kernel PCA. In *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–5, 2024. 2, 5

[19] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. 1

[20] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10012–10022, 2021. 4

[21] Luca Longo, Mario Brcic, Federico Cabitza, Jaesik Choi, Roberto Confalonieri, Javier Del Ser, Riccardo Guidotti, Yoichi Hayashi, Francisco Herrera, Andreas Holzinger, et al. Explainable artificial intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion*, 106:102301, 2024. 1

[22] Andrzej Maćkiewicz and Waldemar Ratajczak. Principal components analysis (PCA). *Computers & Geosciences*, 19 (3):303–342, 1993. 2

[23] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3:861, 2018. 6

[24] Melkamu Mersha, Khang Lam, Joseph Wood, Ali K Alshami, and Jugal Kalita. Explainable artificial intelligence: A survey of needs, techniques, applications, and future direction. *Neurocomputing*, 599:128111, 2024. 1

[25] Sebastian Mika, Bernhard Schölkopf, Alex Smola, Klaus-Robert Müller, Matthias Scholz, and Gunnar Rätsch. Kernel PCA and de-noising in feature spaces. *Advances in Neural Information Processing Systems (NeurIPS)*, 11, 1998. 2

[26] Juan C Moreno, VB Prasath, Dmitry Vorotnikov, Hugo Proença, and Kannappan Palaniappan. Adaptive diffusion constrained total variation scheme with application tocartoon+ texture+ edge'image decomposition. *arXiv preprint arXiv:1505.00866*, 2015. 2

[27] Mohammed Bany Muhammad and Mohammed Yeasin. Eigen-CAM: Class activation map using principal components. In *IEEE International Joint Conference on Neural Networks (IJCNN)*, pages 1–7, 2020. 2

[28] Mahdieh Poostchi, Kannappan Palaniappan, and Guna Seetharaman. Spatial pyramid context-aware moving vehicle detection and tracking in urban aerial imagery. In *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, 2017. 1

[29] Samuele Poppi, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Revisiting the evaluation of class activation mapping for explainability: A novel metric and experimental analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2299–2304, 2021. 4, 5

[30] Mahdi Pourmirzaei, Farzaneh Esmaili, Mohammadreza Pourmirzaei, Mohsen Rezaei, and Dong Xu. Predicting kinase-substrate phosphorylation site using autoregressive transformer. *bioRxiv*, pages 2025–03, 2025. 1

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 1

[32] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. CheXNet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017. 1

[33] Harish Guruprasad Ramaswamy et al. Ablation-CAM: Visual explanations for deep convolutional network via gradient-free localization. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 983–991, 2020. 6, 7

[34] Yao Rong, Tobias Leemann, Vadim Borisov, Gjergji Kasneci, and Enkelejda Kasneci. A consistent and efficient evaluation strategy for attribution methods. *arXiv preprint arXiv:2202.00449*, 2022. 6

[35] Upasana Roy, Vani Seth, Mahady Hasan Rayhan, Ashish Pandey, Mauro Lemus Alarcon, Matthew Maschmann, and Prasad Calyam. Closed-loop cnt growth using retrieval augmented generation and human feedback. In *ICHMS*, pages 128–133, 2025. 1

[36] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115:211–252, 2015. 4

[37] Kaveh Safavigerdini, Koundinya Nouduri, Ramakrishna Surya, Andrew Reinhard, Zach Quinlan, Filiz Bunyak, Matthew R Maschmann, and Kannappan Palaniappan. Predicting mechanical properties of carbon nanotube (CNT) images using multi-layer synthetic finite element model simulations. In *IEEE International Conference on Image Processing (ICIP)*, pages 3264–3268, 2023. 1

[38] Kaveh Safavigerdini, Ramakrishna Surya, Andrew Reinhard, Zach Quinlan, Filiz Bunyak, Matthew R Maschmann, and Kannappan Palaniappan. Creating semi-quanta multi-layer synthetic CNT images using CycleGAN. In *IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–6, 2023. 1

[39] K Safavigerdini, J Collins, R Huynh, J Fraser, and K Palaniappan. EpiX 2.0: an enhanced 3D measurement tool with integrated triangulation for cultural heritage and aerial photogrammetry applications. In *Geospatial Informatics XIV*, page PC1303706, 2024. 1, 2

[40] Kaveh Safavigerdini, Ramakrishna Surya, Jaired Collins, Prasad Calyam, Filiz Bunyak, Matthew R Maschmann, and Kannappan Palaniappan. Automated feature tracking for real-time kinematic analysis and shape estimation of carbon nanotube growth. *arXiv preprint arXiv:2508.19232*, 2025. 1

[41] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. 2, 5

[42] Saleh Valizadeh Sotubadi, Shyam Sasi Pallissery, and Vinh Nguyen. Multi-modal explainable artificial intelligence for neural network-based tool wear detection in machining. *Engineering Applications of Artificial Intelligence*, 2025. 1

[43] Mingzhai Sun, Jiaqing Huang, Filiz Bunyak, Kristyn Gumpper, Gejing De, Matthew Sermersheim, George Liu, Pei-Hui Lin, Kannappan Palaniappan, and Jianjie Ma. Superresolution microscope image reconstruction by spatiotemporal object decomposition and association: Application in resolving t-tubule structure in skeletal muscle. *Optics Express*, 22(10): 12160–12176, 2014. 2

[44] Ramakrishna Surya, Gordon L Koerner, Taher Hajilounezhad, Kaveh Safavigerdini, Martin Spies, Prasad Calyam, Filiz Bunyak, Kannappan Palaniappan, and Matthew R Maschmann. CNT forest self-assembly insights from in-situ ESEM synthesis. *Carbon*, 229:119439, 2024. 1

[45] Dan Tang, Jinjing Chen, Lijuan Ren, Xie Wang, Daiwei Li, and Haiqing Zhang. Reviewing CAM-based deep explainable methods in healthcare. *Applied Sciences*, 14:4124, 2024. 1

[46] Deniz Kavzak Ufuktepe, Feng Yang, Yasmin M Kassim, Hang Yu, Richard J Maude, Kannappan Palaniappan, and Stefan Jaeger. Deep learning-based cell detection and extraction in thin blood smears for malaria diagnosis. In *IEEE AIPR*, pages 1–6, 2021. 1

[47] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-CAM: Score-weighted visual explanations for convolutional neural networks, 2020. 7

[48] Yang Yang Wang, O. V. Glinskii, Filiz Bunyak, and Kannappan Palaniappan. Ensemble of deep learning cascades for segmentation of blood vessels in confocal microscopy images. In *IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–7, 2021. 1

[49] Bahram Yaghooti, Ali Siahi Shadbad, Kaveh Safavi, and Hassan Salarieh. Adaptive synchronization of uncertain fractional-order chaotic systems using sliding mode control techniques. *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering*, 234(1):3–9, 2020. 1

[50] Bahram Yaghooti, Netanel Raviv, and Bruno Sinopoli. Beyond PCA: A Gram-Schmidt approach to feature extraction. In *Allerton Conference on Communication, Control, and Computing*, pages 1–8, 2024. 2, 3, 7

[51] Bahram Yaghooti, Netanel Raviv, and Bruno Sinopoli. Gram-Schmidt methods for unsupervised feature selection. In *IEEE Conference on Decision and Control (CDC)*, pages 7700–7707, 2024. 2

[52] Bahram Yaghooti, Kaveh Safavigerdini, Reza Hajiloo, and Hassan Salarieh. Stabilizing unstable periodic orbit of unknown fractional-order systems via adaptive delayed feedback control. *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering*, 238(4):693–703, 2024. 1

[53] Bahram Yaghooti, Netanel Raviv, and Bruno Sinopoli. Gram–schmidt methods for unsupervised feature extraction and selection. *IEEE Transactions on Information Theory*, 71 (10):7856–7885, 2025. 2, 3

[54] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2016. 1, 2

[55] Erfan Ziad, Zhuo Yang, Yan Lu, and Feng Ju. Knowledge constrained deep clustering for melt pool anomaly detection in laser powder bed fusion. In *IEEE Int. Conf. on Automation Science and Engineering (CASE)*, pages 670–675, 2024. 1