SaliencyDecor: Enhancing Neural Network Interpretability through Feature Decorrelation

Ali Karkehabadi[1*], Jamshid Hassanpour[2*], Houman Homayoun[1], Avesta Sasan[1]
[1]University of California, Davis, USA
[2]Georgia Institute of Technology, USA
{akarkehabadi, hhomayoun, asasan}@ucdavis.edu, jamshid@gatech.edu
* These authors contributed equally to this work.

***Abstract*—Gradient-based saliency methods are widely used to interpret deep neural networks, yet they often produce noisy and unstable explanations that poorly align with semantically meaningful input features. We argue that a fundamental cause of this behavior lies in the geometry of learned representations: correlated feature dimensions diffuse attribution gradients across redundant directions, resulting in blurred and unreliable saliency maps. To address this issue, we identify feature correlation as a structural limitation of gradient-based interpretability and propose *SaliencyDecor*, a training framework that enforces feature decorrelation to improve attribution fidelity without modifying saliency methods or model architectures by reshaping the feature space toward orthogonality, our approach promotes more concentrated gradient flow and improves the fidelity of saliency-based explanations. SaliencyDecor jointly optimizes classification, prediction consistency under feature masking, and a decorrelation regularizer, requiring no architectural changes or inference-time overhead. Extensive experiments across multiple benchmarks and architectures demonstrate that our method produces substantially sharper and more object-focused saliency maps while simultaneously improving predictive performance, achieving accuracy gains across the datasets. These results establish our method as a principled mechanism for enhancing both interpretability and accuracy, challenging the conventional trade-off between explanation quality and model performance.**

***Index Terms*—Interpretability, Representation collapse**

I. Introduction

Deep neural networks (DNNs) have achieved state-of-the-art performance across a wide range of tasks, including computer vision, natural language processing, and time series analysis [1], [2]. Despite their success, the lack of interpretability in these models remains a major obstacle to their deployment in high-stakes applications. Understanding which input features drive model predictions is essential for validating model behavior, diagnosing failures, and establishing user trust [3], [4]. Saliency methods address this challenge by assigning importance scores to input features based on their influence on model outputs [5], [6]. However, many existing approaches produce explanations that are noisy, unstable, and poorly aligned with human intuition [7]. In practice, saliency maps often exhibit weak foreground-background separation and inconsistent attribution of semantically relevant regions, limiting their usefulness for both interpretation and debugging.

A key source of this behavior lies in the structure of the learned feature representations. Gradient-based explanations are particularly sensitive to correlations in the underlying feature space. When multiple features encode redundant information, gradients may diffuse across correlated dimensions, obscuring truly influential input elements and amplifying noise in irrelevant regions. Related phenomena have been studied in representation learning under the notion of dimensional collapse, where learned representations concentrate along a small number of correlated directions despite high nominal dimensionality [8]. Such collapse reduces representation efficiency and degrades gradient signal quality. Standard normalization techniques such as Batch Normalization [9] correct first-order statistics but do not explicitly remove inter-feature correlations, leaving this issue largely unaddressed. Several saliency methods attempt to mitigate noisy explanations through post-hoc smoothing or aggregation. SmoothGrad averages gradients over noisy input perturbations [10], while Integrated Gradients accumulates gradients along a path from a baseline input [6]. Although effective in some cases, these techniques do not modify the underlying representations and therefore cannot resolve the structural causes of saliency degradation.

Similarly, saliency-guided training approaches [11] incorporate gradient information into optimization but still operate on correlated feature spaces, resulting in unstable and misleading attributions [12]. Recent works have attempted to improve saliency-guided training through more adaptive masking strategies, such as SMOOT [13] and HLGM [14], which enhance both accuracy and saliency quality. However, these methods still rely on the underlying feature representations and do not explicitly address feature correlation, leaving structural sources of attribution instability unresolved.

The limitations of current saliency methods have been systematically documented. These findings suggest that improving explanation quality requires addressing structural properties of neural representations rather than modifying gradient computations alone.

In this work, we identify feature correlation in learned representations as a key bottleneck for saliency-based interpretability. Correlated features introduce redundancy, distort importance distributions, and reduce the contrast between relevant and irrelevant regions in saliency maps. We propose *SaliencyDecor*, a training framework that explicitly addresses this issue by integrating ZCA whitening with saliency-guided

optimization. By decorrelating intermediate representations during training, our method promotes more balanced gradient propagation and yields saliency maps that are sharper, more stable, and better aligned with semantically meaningful features. Extensive experiments on image classification benchmarks demonstrate consistent improvements in saliency quality, accompanied by competitive or improved classification performance.

II. Related Work

A. Saliency Methods

Gradient-based saliency attributes predictions using input derivatives, beginning with raw gradients [5] and later refinements [6], [10], [15], [16]. Despite their popularity, these methods often produce noisy and unstable explanations. Prior studies show that saliency maps can fail sanity checks [7] and are highly sensitive to small input perturbations [17].

B. Feature Decorrelation

Decorrelation has been explored to improve representation quality and optimization. DeCov [18] penalizes correlated activations, while Decorrelated Batch Normalization [8] incorporates whitening into normalization layers. In self-supervised learning, whitening and redundancy-reduction techniques have been shown to mitigate dimensional collapse and improve representation diversity [19], [20]. Recent studies further highlight the prevalence and impact of dimensional collapse in learned representations, demonstrating its effect on feature quality and downstream performance [21]. However, the connection between feature decorrelation and gradient-based interpretability remains largely unexplored.

C. Saliency-Guided Training and Interpretability

Several works integrate interpretability objectives into training, including gradient regularization [22], explanation supervision [23], and saliency-guided training [11]. While effective in some settings, these approaches operate on correlated representations and do not address structural sources of gradient noise [12]. Alternative strategies modify architectures or objectives through class activation mapping [24], distillation [25], or robustness-based alignment [26].

D. Normalization and Representation Structure

Normalization methods such as Batch Normalization [9] stabilize first-order statistics but leave feature correlations largely unchanged, while whitening variants mainly target optimization efficiency [8]. In contrast, we explicitly leverage feature decorrelation to improve saliency fidelity, linking representation geometry with attribution quality.

III. Problem Statement: The Dimensional Collapse Challenge in Interpretability

Neural network explanations based on input gradients suffer from geometric limitations induced by correlated feature spaces, leading to two forms of collapse. Complete collapse occurs when representations become constant, yielding $\nabla_X f_\theta(X) \approx 0$ and eliminating informative gradients. Dimensional collapse arises when representations lie on a low-dimensional manifold, i.e., the covariance $\mathbf{\Sigma}$ has low effective rank,

$$\mathrm{rank}_{\mathrm{eff}}(\mathbf{\Sigma}) \ll \dim(\mathbf{\Sigma}), \quad \mathrm{rank}_{\mathrm{eff}}(\mathbf{\Sigma}) = \exp\big(-\textstyle\sum_i \tilde{\lambda}_i \log \tilde{\lambda}_i\big),$$

where $\tilde{\lambda}_i = \lambda_i / \sum_j \lambda_j$ are normalized eigenvalues of $\mathbf{\Sigma}$.

In such settings, correlated features cause gradients to diffuse across redundant dimensions rather than concentrate on informative ones, limiting interpretability. Current methods fail to recognize that the geometric structure of the representation space directly impacts gradient flow and feature attribution. Without proper decorrelation, we show that saliency methods suffer from:

- Gradient Entanglement: Correlated features produce entangled gradients that distribute attribution across redundant features.
- Vanishing Saliency Gaps: The separation between salient and non-salient features diminishes as correlation increases.
- Stochastic Axis Swapping: During training, correlated representations cause unstable gradient directions.

These challenges require a principled approach that transforms the geometry of the feature space to enable faithful gradient-based explanations.

IV. Methodology: SaliencyDecor Framework

SaliencyDecor enhances gradient-based interpretability through feature decorrelation during training. Our framework addresses dimensional collapse by combining ZCA whitening with saliency-guided optimization, creating representations where gradients naturally align with human-interpretable features. Figure 1 shows the general architecture of our method.

A. Geometric Whitening for Decorrelated Features

To prevent dimensional collapse, SaliencyDecor applies ZCA whitening. Given features $\mathbf{Z} \in \mathbb{R}^{d \times m}$,

$$\tilde{\mathbf{Z}} = \mathbf{\Sigma}^{-1/2}(\mathbf{Z} - \boldsymbol{\mu}\mathbf{1}^T),$$

where $\boldsymbol{\mu}$ is the batch mean and $\mathbf{\Sigma} = \frac{1}{m}(\mathbf{Z} - \boldsymbol{\mu}\mathbf{1}^T)(\mathbf{Z} - \boldsymbol{\mu}\mathbf{1}^T)^T = \mathbf{D}\mathbf{\Lambda}\mathbf{D}^T$ is the covariance eigendecomposition, with $\mathbf{D}$ the eigenvectors and $\mathbf{\Lambda}$ the eigenvalues, so that $\mathbf{\Sigma}^{-1/2} = \mathbf{D}\mathbf{\Lambda}^{-1/2}\mathbf{D}^T$. We choose ZCA because, unlike PCA whitening, it preserves feature orientation while enforcing orthogonality, ensuring gradients flow along consistent semantic directions.

B. Efficient Group-wise Feature Decorrelation

To address computational challenges in high-dimensional spaces, SaliencyDecor implements efficient group-wise whitening. Features are partitioned into groups of size G, and whitening is applied within each group:

$$\mathbf{Z}^{[h]} = \mathrm{ZCA}(\mathbf{Z}^{[h]}) \tag{1}$$

where $\mathbf{Z}^{[h]}$ represents the h-th group of features. This reduces computational complexity from $\mathcal{O}(d^2 m)$ to $\mathcal{O}(dmG)$, making our method scalable to modern architectures with millions of parameters.

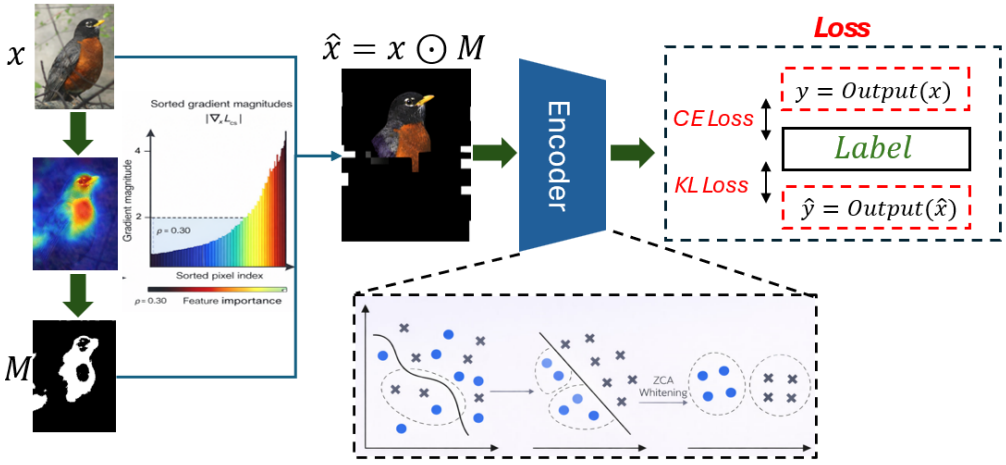

Fig. 1. **Overview of the proposed SaliencyDecor framework.** Given an input X, gradient-based importance scores are used to identify non-informative regions and generate a saliency mask M. Intermediate encoder features are decorrelated via group-wise ZCA whitening to reduce redundancy and stabilize gradient attribution. The network is trained with a multi-objective loss combining classification, consistency, and decorrelation terms, encouraging orthogonal feature representations that produce sharper and more faithful saliency maps without inference-time overhead.

C. Multi-objective Optimization Framework

Our learning objective combines three components:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \alpha \cdot \mathcal{L}_{\text{consistency}} + \lambda \cdot \mathcal{L}_{\text{decorr}} \tag{2}$$

where:

- $\mathcal{L}_{\text{cls}}$ is the standard cross-entropy loss for classification
- $\mathcal{L}_{\text{consistency}} = \mathcal{D}_{\text{KL}}(f_\theta(X) \| f_\theta(\tilde{X}))$ is the KL-divergence between predictions from original and masked inputs
- $\mathcal{L}_{\text{decorr}} = \|\tilde{\mathbf{Z}}^T \tilde{\mathbf{Z}} - \mathbf{I}\|_F$ measures deviation from orthogonality
- α and λ control the trade-off between accuracy, consistency, and decorrelation

For feature masking, we leverage decorrelated gradients to identify truly non-informative features. Specifically, we compute importance scores from gradients in the whitened space and project them back to the input domain:

$$\text{importance} = |\nabla_X f_\theta(X)| \tag{3}$$

This ensures that masking decisions reflect true information content rather than correlation artifacts.

Algorithm 1: SaliencyDecor: Decorrelated Saliency-Guided Training

Input: Data $\{X, y\}$, mask ratio ρ, weights α, λ, group size G

Initialize parameters θ ;

for *each minibatch* $\{X, y\}$ **do**

- $Z \leftarrow f_{\text{enc}}(X)$;
- $\tilde{Z} \leftarrow \text{GroupZCA}(Z, G)$;
- $\hat{y} \leftarrow f_{\text{cls}}(\tilde{Z})$;
- $\mathcal{L}_{\text{cls}} \leftarrow \text{CE}(\hat{y}, y)$;
- $\mathcal{L}_{\text{decorr}} \leftarrow \|\tilde{Z}^\top \tilde{Z} - I\|_F$;
- $S_{\tilde{Z}} \leftarrow |\nabla_{\tilde{Z}} \mathcal{L}_{\text{cls}}|$;
- $S_X \leftarrow \text{ProjectToInput}(S_{\tilde{Z}})$;
- $\tilde{X} \leftarrow \text{Mask}(X, \rho, S_X)$;
- $\mathcal{L}_{\text{cons}} \leftarrow \mathcal{D}_{\text{KL}}\big(f_\theta(X) \,\|\, f_\theta(\tilde{X})\big)$;
- $\mathcal{L} \leftarrow \mathcal{L}_{\text{cls}} + \alpha \mathcal{L}_{\text{cons}} + \lambda \mathcal{L}_{\text{decorr}}$;
- Update θ ;

end

D. SaliencyDecor Algorithm

In Algorithm 1, we present the complete SaliencyDecor procedure. For each minibatch, we first compute feature representations through the encoder. We then apply group-wise ZCA whitening to these features, decorrelating them while preserving their original orientation. After computing the classification loss, we calculate a decorrelation regularization term that encourages orthogonality in feature space. For creating input masks, we compute gradients of the classification loss with respect to the input and identify the least important features based on gradient magnitude. We then generate a masked input by replacing these features and compute a consistency loss to ensure that the model's predictions remain stable despite feature masking. Finally, we update the parameters using the combined loss that balances classification accuracy, mask consistency, and feature decorrelation. SaliencyDecor is theoretically grounded in geometric deep learning principles, where the structure of the representation space directly impacts gradient flow and feature attribution. By reshaping the geometry of the feature space, SaliencyDecor creates representations where gradient-based saliency naturally aligns with human-interpretable features without requiring post-hoc modifications to existing saliency methods. This foundational approach provides consistent improvements across various architectures and saliency methods, addressing the core geometric limitations that have hindered interpretability in deep learning.

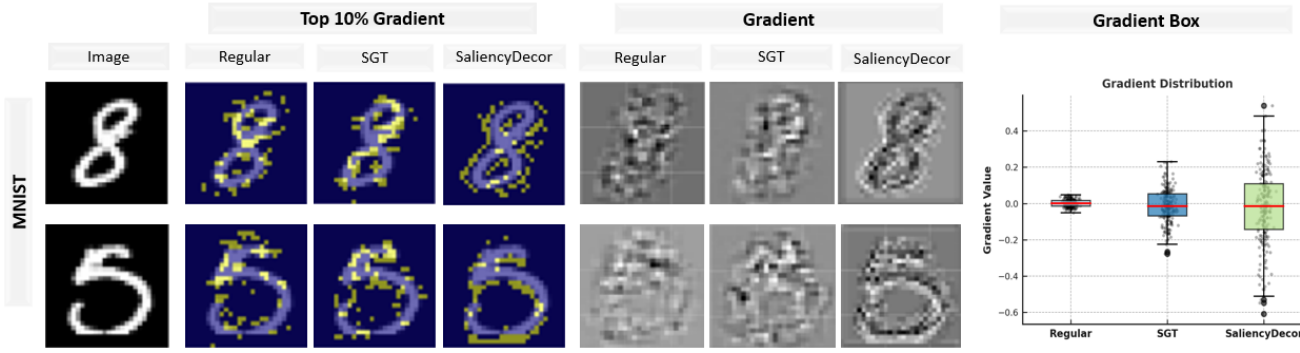

Fig. 2. Gradient visualization and distribution analysis for MNIST. Left: Original digit images. Middle columns: Top 10% gradients and full gradient maps for Regular, SGT, and SaliencyDecor methods. Right: Statistical distribution of gradient values showing box plots. SaliencyDecor produces focused attributions with significantly wider separation between important and non-important features.

V. Empirical Validation and Analysis

We benchmark our proposed method across different architectures and datasets. We further report the qualitative and quantitative results to show the performance of our model.

A. Implementation Details

We evaluate SaliencyDecor on standard benchmarks, including MNIST [27], CIFAR-10 [28], and Caltech-101 [29], Birds dataset [26] and Tiny Imagenet [30]. For architectures, we used a 2-layer CNN with 392 feature maps for MNIST, a 2-layer CNN with 512 feature maps for CIFAR-10, ResNet-18 with 512 feature maps for Caltech-101, and ResNet-50 for the Birds dataset. The models were implemented in PyTorch with SGT (momentum = 0.9), an initial learning rate of 0.01 with cosine annealing, and batch size of 128. Hyperparameters were consistent across experiments: decorrelation strength $\lambda = 0.01$, consistency weight $\alpha = 0.1$, masking ratio $\rho = 25\%$ and group size $G = 64$ for ZCA whitening.

B. Results

1) Visual Attribution Analysis: Figure 2 compares gradient maps on MNIST for standard training (Regular), SGT [11], and SaliencyDecor. Our method produces cleaner, more concentrated attributions aligned with digit structure, while baselines exhibit scattered and noisy gradients. Similar trends are observed on CIFAR-10, Caltech-101, and Birds (Fig. 4), where SaliencyDecor yields sharper, object-focused saliency and suppresses background noise.

2) Statistical Gradient Distribution: The box plots in Figure 2 (right) show clear differences in gradient behavior across methods. Standard training yields a collapsed distribution with little separation between important and non-important features, whereas SaliencyDecor produces a wider spread and stronger separation of attribution values, particularly on MNIST.

3) Robustness and Faithful Attribution Evaluation: We evaluate attribution fidelity using a perturbation-based masking test, where input features are progressively removed according to their saliency scores and the resulting accuracy degradation is measured. Figure 3 shows the results on MNIST. Models

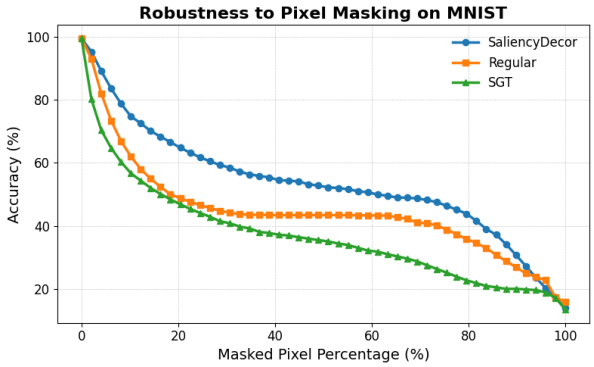

Fig. 3. Accuracy degradation under progressive feature masking on MNIST: the baseline model (AUC = 5353.33), the conventional saliency method (AUC = 4458.52), and SaliencyDecor (AUC = 3707.19) show that models trained with SaliencyDecor exhibit significantly steeper accuracy decline, confirming more precise identification of truly important features compared to conventional methods.

trained with SaliencyDecor exhibit a substantially steeper accuracy decline than both standard training and SGT, indicating that the identified features are more critical for prediction. This effect is most evident between 12% and 36% masking, where accuracy drops from approximately 75–80% to around 60%, while competing methods maintain higher accuracy over the same range. Table I further quantifies attribution quality using the Area Under the Masking Curve (AUC), where lower values indicate better fidelity. SaliencyDecor achieves the lowest AUC while maintaining competitive classification accuracy, demonstrating improved identification of discriminative features.

TABLE I
MNIST Performance Comparison (AUC)

Baseline (w/o Saliency)	SGT	SaliencyDecor
5353.33	4458.52	**3707.19**

TABLE II
ACCURACY COMPARISON ACROSS COMMON BENCHMARKS FOR SALIENCY-/GRADIENT-BASED WORKS. "–" INDICATES THAT THE CORRESPONDING LITERATURE DOES NOT REPORT RESULTS FOR THAT DATASET.

Dataset **Model**	**MNIST** Simple CNN	**CIFAR-10** CNN	**CIFAR-10** ResNet-18	**Caltech-101** ResNet-18	**Birds** CNN / ResNet	**Tiny ImageNet** DeiT-Tiny
Baseline (Standard training without Saliency)	**99.4**	73.5	94.6	94.5	93.6	72.4
Gradient Regularization (SpectReg) [31]	97.7	–	–	–	–	50.8
Integrated Gradient Correlation [32]	99.3	–	–	–	–	–
SCAAT [33]	–	–	90.5	–	–	–
SGT [11]	99.3	73	92.9	94.7	94.3	72.9
SaliencyDecor (Ours)	**99.4**	**76.4**	**95.1**	**96.2**	**95.2**	**73.1**

4) Extending SaliencyDecor to Vision Transformers: We extend SaliencyDecor to Vision Transformers by applying feature decorrelation directly to the CLS token representation produced by the transformer encoder. After the final transformer block, the CLS embedding is extracted and partitioned into fixed-size groups, on which a group-wise ZCA whitening operation is applied to reduce feature correlation while preserving feature orientation. The decorrelated CLS representation is used solely to compute an auxiliary feature decorrelation loss, whereas classification logits are obtained from the original (non-whitened) CLS embedding. This design decouples predictive performance from the whitening operation, allowing decorrelation to act as a regularizer without altering the classifier decision space. The overall training objective combines standard cross-entropy loss with the decorrelation loss, encouraging the model to learn less redundant yet discriminative representations. This formulation enables a lightweight integration of SaliencyDecor into transformer architectures without modifying the attention mechanism or introducing inference-time overhead.

C. Ablation Study

1) Impact of feature decorrelation without saliency guidance: Table III compares SaliencyDecor against a decorrelation-only baseline using Decorrelated Batch Normalization (DBN) [8]. While DBN significantly degrades performance relative to the standard baseline, SaliencyDecor consistently improves accuracy across both seeds. These results indicate that whitening alone is insufficient, and that the benefits arise from the joint integration of decorrelation with saliency-guided masking and consistency objectives.

TABLE III
ABLATION RESULTS ON CIFAR-10 USING RESNET-18 UNDER IDENTICAL TRAINING SETTINGS. TEST ACCURACY (%) IS REPORTED.

Method	**Test Acc. (%)**
Baseline (Regular)	94.59
SGT	92.77
DBN	81.40
Ours	**95.12**

2) Effect of masking ratio k.: We study the impact of the masking ratio k used to remove low-saliency features during training. As shown in Table IV (a), masking 25% achieves the best test accuracy (95.12%). Increasing the masking ratio to 50% degrades accuracy (94.75%), suggesting that overly aggressive masking removes informative regions and makes the consistency objective harder to satisfy. Masking 75% partially recovers performance (94.96%), but remains below 25%. Moderate masking provides the best trade-off between enforcing robustness and preserving discriminative evidence.

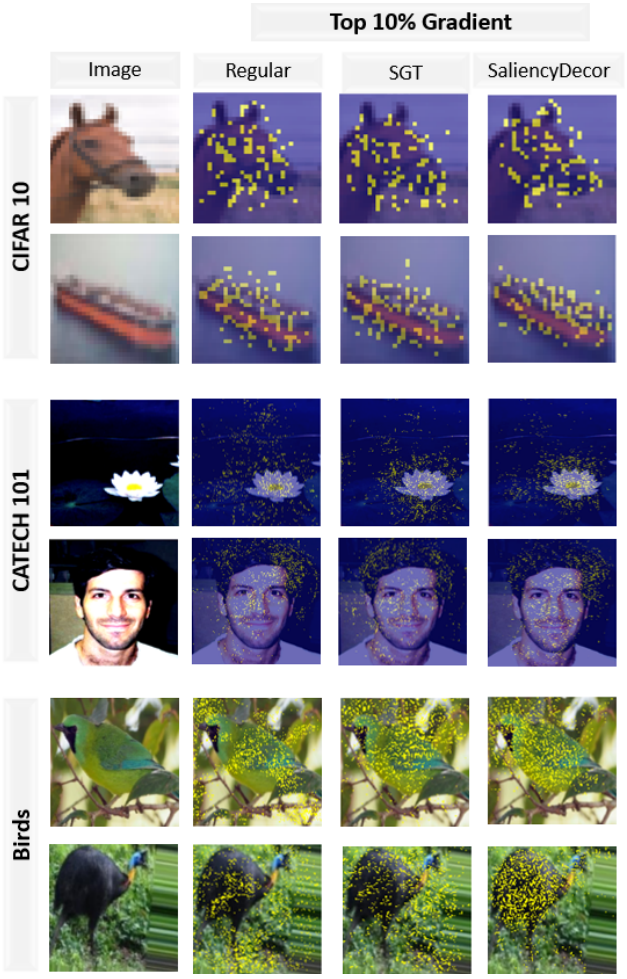

Fig. 4. Gradient visualization analysis for complex datasets. Shows original images from CIFAR-10, Caltech-101, and Birds datasets with corresponding gradient patterns from Regular, SGT, and SaliencyDecor methods. SaliencyDecor produces coherent attributions that align with discriminative features while suppressing noise across all complex visual recognition tasks.

TABLE IV
ABLATION STUDY ON MASKING RATIO k AND DECORRELATION WEIGHT λ.

Masking Ratio (k)	Acc. (%)
25%	**95.12**
50%	94.75
75%	94.96

(a) Masking ratio

λ	Acc. (%)
0.001	**95.17**
0.01	95.12
0.1	95.02

(b) Regularization weight

D. Effect of decorrelation weight λ

We vary the decorrelation weight λ (Table IV(b)). A small value ($\lambda = 0.001$) yields the best accuracy (95.17%), while larger values slightly degrade performance, indicating that decorrelation is most effective as a mild regularizer.

ACKNOWLEDGMENTS

This research was supported by the National Science Foundation under Award #2233893.

VI. CONCLUSION

This work links representation geometry to interpretability, identifying feature correlation as a key cause of degraded saliency. We introduce SaliencyDecor, which integrates ZCA whitening with saliency-guided training to learn decorrelated features and produce cleaner, more reliable attributions without inference overhead. Across benchmarks, our method consistently improves both attribution fidelity and classification performance, demonstrating that feature decorrelation provides a simple and principled approach to a reliable interpretability.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, pp. 1097–1105, 2012.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[3] Z. C. Lipton, "The mythos of model interpretability," *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[4] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1721–1730, 2015.

[5] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. Müller, "How to explain individual classification decisions," *Journal of Machine Learning Research*, vol. 11, pp. 1803–1831, 2010.

[6] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *International Conference on Machine Learning*, pp. 3319–3328, PMLR, 2017.

[7] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Advances in Neural Information Processing Systems*, pp. 9505–9515, 2018.

[8] L. Huang, D. Yang, B. Lang, and J. Deng, "Decorrelated batch normalization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 791–800, 2018.

[9] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*, pp. 448–456, PMLR, 2015.

[10] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "Smoothgrad: removing noise by adding noise," in *Workshop on Visualization for Deep Learning, ICML*, 2017.

[11] A. A. Ismail, S. Feizi, and H. C. Bravo, "Improving deep learning interpretability by saliency guided training," *Advances in Neural Information Processing Systems*, vol. 34, pp. 13890–13903, 2021.

[12] A. Kapishnikov, T. Bolukbasi, F. Viégas, and M. Terry, "Guided integrated gradients: An adaptive path method for removing noise," *arXiv preprint arXiv:2106.09650*, 2021.

[13] A. Karkehabadi, H. Homayoun, and A. Sasan, "Smoot: Saliency guided mask optimized online training," in *2024 IEEE 17th Dallas circuits and systems conference (DCAS)*, pp. 1–6, IEEE, 2024.

[14] A. Karkehabadi, B. S. Latibari, H. Homayoun, and A. Sasan, "Hlgm: A novel methodology for improving model accuracy using saliency-guided high and low gradient masking," in *2024 14th International Conference on Information Science and Technology (ICIST)*, pp. 909–917, IEEE, 2024.

[15] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *International Conference on Machine Learning*, pp. 3145–3153, PMLR, 2017.

[16] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, pp. 4765–4774, 2017.

[17] A. Ghorbani, A. Abid, and J. Zou, "Interpretation of neural networks is fragile," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 3681–3688, 2019.

[18] M. Cogswell, F. Ahmed, R. Girshick, L. Zitnick, and D. Batra, "Reducing overfitting in deep networks by decorrelating representations," in *International Conference on Learning Representations*, 2016.

[19] T. Hua, W. Wang, Z. Xue, S. Ren, Y. Wang, and H. Zhao, "On feature decorrelation in self-supervised learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9598–9608, 2021.

[20] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *International Conference on Machine Learning*, pp. 12310–12320, PMLR, 2021.

[21] J. Hassanpour, V. Srivastav, D. Mutter, and N. Padoy, "Overcoming dimensional collapse in self-supervised contrastive learning for medical image segmentation," pp. 1–5, 05 2024.

[22] A. S. Ross, M. C. Hughes, and F. Doshi-Velez, "Right for the right reasons: Training differentiable models by constraining their explanations," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pp. 2662–2670, 2017.

[23] R. Ghaeini, X. Z. Fern, and P. Tadepalli, "Saliency learning: Teaching the model where to pay attention," *arXiv preprint arXiv:1902.08649*, 2019.

[24] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016.

[25] N. Frosst and G. Hinton, "Distilling a neural network into a soft decision tree," *arXiv preprint arXiv:1711.09784*, 2017.

[26] C. Etmann, S. Lunz, P. Maass, and C. Schönlieb, "A connection between adversarial robustness and saliency map interpretability," in *International Conference on Machine Learning*, pp. 1823–1832, PMLR, 2019.

[27] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 2002.

[28] A. Krizhevsky, G. Hinton, *et al.*, "Learning multiple layers of features from tiny images," 2009.

[29] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," in *2004 conference on computer vision and pattern recognition workshop*, pp. 178–178, IEEE, 2004.

[30] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[31] D. Varga, A. Csiszárik, and Z. Zombori, "Gradient regularization improves accuracy of discriminative models," *arXiv preprint arXiv:1712.09936*, 2017.

[32] P. Lelièvre and C.-C. Chen, "Integrated gradient correlation: a dataset-wise attribution method," *arXiv preprint arXiv:2404.13910*, 2024.

[33] R. Xu, W. Qin, P. Huang, H. Wang, and L. Luo, "Scaat: Improving neural network interpretability via saliency constrained adaptive adversarial training," *arXiv preprint arXiv:2311.05143*, 2023.