GLM-Based: CAM Based on Gradient Local Mean for Enhanced Explainability in Visual Transformer

1st Keyuan Wang
Northwest A&F University
Yangling, China
keyuanwang@nwafu.edu.cn

2nd Yixin Ren
Northwest A&F University
Yangling, China
renyixin@nwafu.edu.cn

3rd Lingyuan Kong
Harbin Institute of Technology
Harbin, China
2022056038@nwafu.edu.cn

4th Zhiyi Zhang*
Northwest A&F University
Yangling, China
zhangzhiyi@nwafu.edu.cn

***Abstract*—The rapid advancement of Visual Transformers (ViT) has created an urgent need for model interpretability, however, existing gradient-based explanation methods often suffer from significant visual noise caused by local gradient fluctuations. To address this challenge, this paper proposes the Gradient Local Mean-based (GLM) method, a novel framework designed to enhance the transparency of ViT models. The GLM method employs a stochastic sampling strategy by introducing Gaussian noise into the input image and aggregating the resulting Class Activation Maps (CAMs)—derived from gradient and attention matrices—to effectively mitigate noise instability. Extensive experiments on the ImageNet ILSVRC 2012 and CUB-200-2011 datasets demonstrate the method's superiority. On the fine-grained CUB-200-2011 dataset, the GLM method achieves a Top-1 Localization Intersection over Union (IoU) of 0.5958, significantly outperforming the state-of-the-art AGCAM method (0.4235) by 40.7%, while attaining a Pixel Accuracy of 0.8405. Furthermore, in pixel perturbation tests on ImageNet, the proposed method yields a superior Area Between Perturbation Curves (ABPC) score of 0.3739, statistically outperforming leading alternatives ($p < 0.05$). These results confirm that the GLM method provides robust and precise visual explanations, facilitating reliable decision-making in practical applications.**

***Index Terms*—class activation map, visual transformer, attention mechanism, gaussian noise, explainability.**

I. Introduction

In recent years, Visual Transformer (ViT) models [1] have made impressive achievements in the field of computer vision. ViT models apply the attention mechanism [2] to achieve remarkable performance in image classification, semantic segmentation, and target detection tasks [3]–[5]. For example, Touvron et al. [6] demonstrated the superiority of ViT performance by simplifying the model on the ResNet-50 training set, while Yin et al. [7] proposed A-ViT to enable adaptive tuning of the inference cost. Additionally, Tumanyan et al. [8] utilized deep ViT features to stitch together desired structures and appearances.

The rapid development of ViT models has resulted in a notable increase in their complexity, which is a significant challenge limiting their further application [9]–[12]. This increased complexity has created an urgent need for model explainability to help understand and explain the model's decision-making process [13]–[17]. Gaining an understanding of the internal mechanisms of these models is crucial for building reliable applications and solving real-world problems [18]–[20].

Existing approaches for ViT model explainability, such as Attention Rollout [21], Layer-wise Relevance Propagation (LRP)-based methods [22], and Attention Guided CAM (AGCAM) [23], have been proposed to visualize the decision-making process. However, as we will discuss in Section II, these methods often suffer from significant shortcomings, such as low localization accuracy or visual noise caused by gradient fluctuations.

To address these challenges, this paper proposes the Gradient Local Mean-based (GLM) method. As illustrated in the following sections, the GLM method employs a stochastic sampling strategy by introducing Gaussian noise into the input image and aggregating the resulting Class Activation Maps (CAMs) to effectively mitigate noise instability.

The main contributions of this paper are as follows:

- This paper proposes an explainability method for the ViT model that enhances the transparency and explainability of the model by generating the CAM that reveals the model's attention to different regions in the image.
- The GLM-based method generates multiple noise samples by introducing Gaussian noise into the input image, calculates the CAM of each sample based on the gradient matrix and the attention score matrix, and averages the CAM of these samples. Compared with the existing methods, the proposed method effectively reduces the influence of noise caused by gradient fluctuation.
- This paper provides a visualisation method and performs well in pixel perturbation tests. The generated heat maps are less noisy, thus providing a more reliable basis for model decisions.

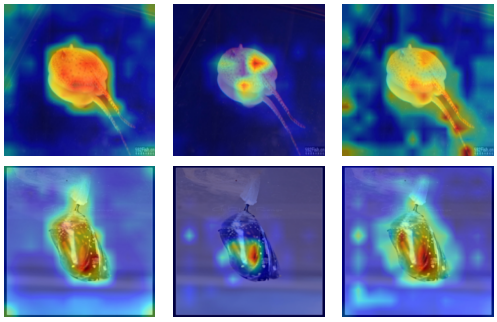
II. Related work

A. Visual Transformers and Explainability

Since the introduction of the ViT [1], self-attention mechanisms have fundamentally shifted computer vision paradigms. However, the lack of inductive bias compared to CNNs leads to increased model complexity [7], [8], necessitating robust interpretability methods [13]–[17].

B. Gradient-based & Attention-based Methods

To elucidate the decision-making of ViTs, various methods have been developed. The Attention Rollout method [21] evaluates the impact of input features by aggregating attention



(a) Rollout (b) LRP-based (c) AGCAM

Fig. 1. Visualisation of regions of interest for different explainability methods. (a) Attention Rollout. (b) Layer-wise Relevance Propagation-based method. (c) Attention Guided CAM.

matrices. However, as shown in Fig. 1(a), it ignores the complex dynamics between layers, often resulting in low localization accuracy .

The LRP-based method [22] attempts to propagate relevance scores backward. Although theoretically sound, the nonlinearities in deep networks cause approximation errors, leading to sparse visualizations that fail to highlight the complete object region, as observed in Fig. 1(b) .

The Attention Guided CAM method [23] improves resolution by combining gradient information with attention features [24]. However, CAMs relying on a single input are susceptible to gradient fluctuations. As demonstrated in Fig. 1(c), this leads to significant visual noise in the heatmap.

C. Noise Reduction in Interpretability

The issue of visual noise caused by local partial derivatives is a known challenge [25]. In traditional networks, regularization techniques like adding noise to samples have been shown to have a denoising effect [26]. Inspired by this, our GLM method extends this principle to ViTs. Unlike AGCAM, GLM introduces stochastic sampling to cancel out noise, providing more stable explanations.

III. Methodology

In this section, we present the Gradient Local Mean-based (GLM) method, a novel framework designed to enhance the explainability of Visual Transformer (ViT) models. By addressing the sensitivity of gradient-based methods to local perturbations, GLM generates more robust and visually coherent explanations. The remainder of this section is organized as follows: Section A provides a high-level overview of the proposed framework; Section B details the specific feature extraction process for individual samples; Section C introduces the core stochastic aggregation strategy for noise reduction; and Section D describes the visualization technique used to generate the final heatmaps.

A. Overview of the GLM Framework

The core motivation of the GLM-based method is to mitigate the instability of gradient-based explanations by introducing a stochastic smoothing mechanism. As illustrated in Fig. 2, unlike traditional methods that compute a single saliency map, the GLM framework operates in three distinct stages:

- Stochastic Sampling: The original input image is replicated and perturbed with Gaussian noise to create a sequence of noisy samples (see "Sample sequence" in Fig. 2).
- Independent Feature Extraction: Each sample is independently processed by the Visual Transformer to generate a specific Class Activation Map (CAM) based on its unique gradient and attention patterns.
- Local Mean Aggregation: The sequence of generated CAMs is aggregated and averaged to cancel out random gradient fluctuations, resulting in a stable "Smooth CAM".

B. CAM generation for a single sample

To implement the GLM framework, we first define the process for generating a CAM for a single, specific sample.

ViT model is a deep neural network model based on self-attention mechanism [1]. It accomplishes the image classification task by segmenting the input image into a number of patches and analyzing the relationship between these patches. Different from traditional convolutional neural networks, the ViT model relies on a multi-layer self-attention mechanism, and the explainability of its decision-making process can be achieved by analyzing the attention matrix and gradient matrix of the model. Based on this property, the explainability of ViT can be guided by using the attention weights matrix on the gradient matrix.

In the forward propagation process of the ViT model, first, the input image x is divided into multiple non-overlapping patches, each of which is mapped into a high-dimensional vector by a linear transformation to generate the corresponding feature representation. In addition, the ViT model adds a learnable position embedding for each patch and introduces a [class] token at the beginning of the input sequence to capture information about the global image. The patch embedding vectors, comprising the position embedding and the [class] token, are then fed into the Transformer encoder.

In the Transformer encoder, the embedding vectors are fed into the multi-head attention layer for processing. Let N be the number of patches (including the classification token) and d_h be the dimension of each attention head. As shown in Fig. 2, the attention score matrix $S_h^k \in \mathbb{R}^{N\times N}$ is generated by computing the dot product between the query matrix $Q_h^k \in \mathbb{R}^{N\times d_h}$ and the key matrix $K_h^k \in \mathbb{R}^{N\times d_h}$ in the h-th attention head of the k-th layer:

$$S_h^k = Q_h^k (K_h^k)^T \tag{1}$$

Then, in order to normalize the attention score matrix S_h^k into an attention weight matrix in the interval [0,1], the softmax function can usually be used. However, the softmax function tends to amplify local maxima, resulting in some attention weights being too high while suppressing other important attention weights [23]. Therefore, the sigmoid function

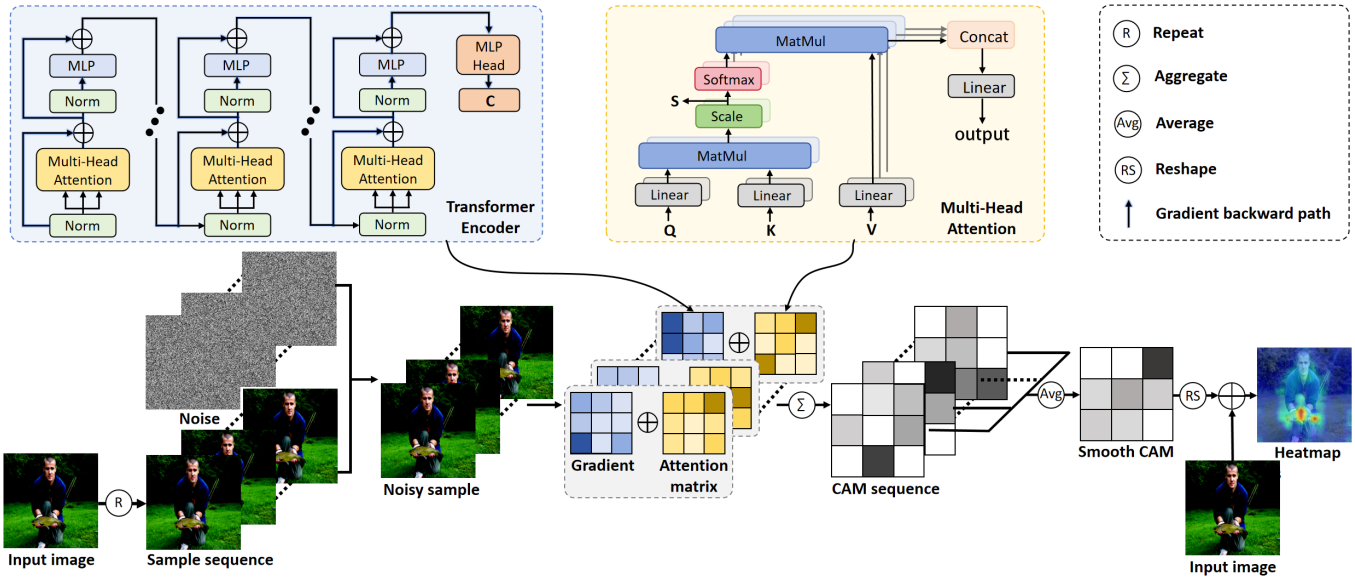


Fig. 2. Structure of the GLM-based method model. The input image is perturbed by adding Gaussian noise to generate multiple noise samples. These samples are passed through the Transformer encoder to obtain the attention score matrix S during forward and the gradient during backward. Combining the gradient and the attention weight matrix of each noise sample yields a sequence of CAMs. These CAMs are then aggregated and averaged to create smoother and more stable CAM. Finally, the resulting smoothed CAM is reshaped to the spatial resolution of the original image and superimposed on the input image to generate the final heatmap.

is chosen here to normalize the matrix S_h^k and generate the normalized attention weight matrix $A_h^k \in \mathbb{R}^{N\times N}$:

$$A_h^k = \text{sigmoid}(S_h^k) \tag{2}$$

In order to generate a complete CAM, it is necessary to combine not only the attention weight matrix obtained through forward, but also the gradient information obtained through backward with respect to a specific class c. Let the gradient propagated from the output class c to each attention weight matrix along the gradient backward path in Fig. 2 be $G_h^{k,c} \in \mathbb{R}^{N\times N}$, which reflects the sensitivity of the class to the attention weight matrix. To ensure that only gradients that contribute positively to the class c are considered, $G_h^{k,c}$ is corrected to filter out the negative portion using the Rectified Linear Unit (ReLU) activation function:

$$G_h'^{k,c} = ReLU(G_h^{k,c}) \tag{3}$$

The attention weight matrix of each attention head is multiplied element by element with the modified gradient matrix $G_h'^{k,c}$ to obtain the weighted gradient matrix $g_h^{k,c} \in \mathbb{R}^{N\times N}$:

$$g_h^{k,c} = A_h^k \bigodot G_h'^{k,c} \tag{4}$$

where $\bigodot$ denotes the Hadamard product.

Subsequently, in order to generate the CAM for individual samples, the weighted gradient matrix is subjected to an aggregation operation that fuses all the heads and layers:

$$M^c = \sum_{k=1}^{K}\sum_{h=1}^{H} g_h^{k,c} \tag{5}$$

In the above equation, K denotes the number of layers of the Transformer encoder and H denotes the number of attention heads in each layer. Ultimately, $M^c \in \mathbb{R}^{N\times N}$ denotes the CAM generated by the model for the class c,Since the dimensionality of M^c corresponds to the number of patches N, it reflects how much attention the model pays to the input features when predicting the class c.

C. Gradient Local Mean Smoothing

The visual noise in the heatmap is related to the way the CAM of the model is computed, in the case where the gradient is computed only once for the original input image, the apparent noise seen in the heat map is caused by fluctuations in the local partial derivative of the CAM with respect to the input image [25]. The common regularisation technique of adding noise to the sample has a denoising effect [26]. By introducing Gaussian noise to generate multiple samples, multiple computations are performed on the input image under several different perturbation conditions. Although the CAM of each sample is affected by the respective noise, these effects are random and inconsistent. By averaging the results of multiple CAMs, the effects of the noise cancel each other out, thus reducing the effect of the overall noise.

As shown in Fig. 2, N samples identical to the original input image x are first generated. Subsequently, an independent Gaussian noise is added to each sample to form multiple noisy samples, allowing the model to process the input under different noise conditions. For the i-th sample x_i, the generation process can be expressed as follows:

$$x_i = x + \epsilon_i, \epsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{6}$$

where $\mathcal{N}(0, \sigma^2)$ is denoted as a Gaussian noise with mean 0 and standard deviation σ and ϵ_i is the random variable of this Gaussian noise. A sample with different noise is generated by adding a different noise ϵ_i to each sample $\{x_1, x_2, \ldots, x_N\}$. The essence of this is to perturb the model inputs so that the gradient response of the model reflects small changes in the inputs.

For each noisy sample x_i, we compute its corresponding CAM according to the CAM generation method for a single sample, we obtain the CAM of class c by forward and backward propagation. After completing the CAM for each noisy sample, the CAM results of all samples are averaged to generate smoother CAM. The smoothed CAM M^c_{smooth} for a specific class c can be expressed by the following equation:

$$M^c_{smooth} = \frac{1}{N} \sum_{i=1}^{N} M^c_i \tag{7}$$

D. CAM Visualization

The dimensionality of M^c_{smooth} corresponds to the number of patches input to the ViT model. Since the input to the ViT model is a chunked image, the generated CAM has a low spatial resolution. In order to make the CAM visualize the regions of interest of the model at the spatial resolution of the original image, the generated CAM is spatially interpolated using a bilinear interpolation method [13]. After completing the up-sampling, the generated high-resolution CAM is superimposed with the original image to generate visualized heatmap that intuitively demonstrates the model's attention to different regions in the input image.

IV. Experiments

This section evaluates the GLM-based method against three leading explainability approaches [21-23] on the ImageNet ILSVRC 2012 [27] and CUB-200-2011 [28] datasets. We utilize pre-trained models from the timm library [29], which are directly evaluated on ImageNet and fine-tuned specifically for the CUB-200-2011 tasks. Performance is assessed via pixel accuracy, Intersection over Union (IoU), Dice coefficient, Precision, and Recall. Additionally, we employ the Area Between Perturbation Curves (ABPC) score [30] for pixel perturbation tests. Adopting the formal definition from the Appendix (Eq. 13) of [30], this metric quantifies the overall heatmap quality by computing the area between the Least Relevant First (LeRF) and Most Relevant First (MoRF) curves, where a larger value indicates better quality.

A. Comparative tests

The quantitative results of the weakly supervised localization test on ImageNet are presented in Table I. The GLM-based method shows consistent improvements across most metrics (statistically significant, $p < 0.05$). This suggests that the noise caused by gradient fluctuations is effectively reduced by CAM smoothing, thus generating smoother and more accurate CAMs. Furthermore, the high Recall and IoU scores demonstrate that our method comprehensively captures the target area, generating heatmaps that align well with the ground truth.

TABLE I
LOCALIZATION PERFORMANCE ON ILSVRC 2012

Model	Attention Rollout	LRP	AGCAM	Ours (GLM)
Pixel acc	0.6209	0.5863	0.7341	**0.7700**
IoU	0.3597	0.2029	0.52212	**0.5827**
Dice	0.4893	0.3055	0.6515	**0.7057**
Precision	0.7326	**0.9110**	0.8299	0.8382
Recall	0.4657	0.2176	0.6276	**0.6970**

Table II presents the pixel perturbation results on ImageNet. The GLM-based method demonstrates consistent robustness, achieving the highest LeRF (0.5350), which indicates effective accuracy retention when removing irrelevant pixels. With a comparable MoRF of 0.1611 ensuring precise feature localization, our method reaches an ABPC of 0.3739, statistically outperforming comparison methods ($p < 0.05$).

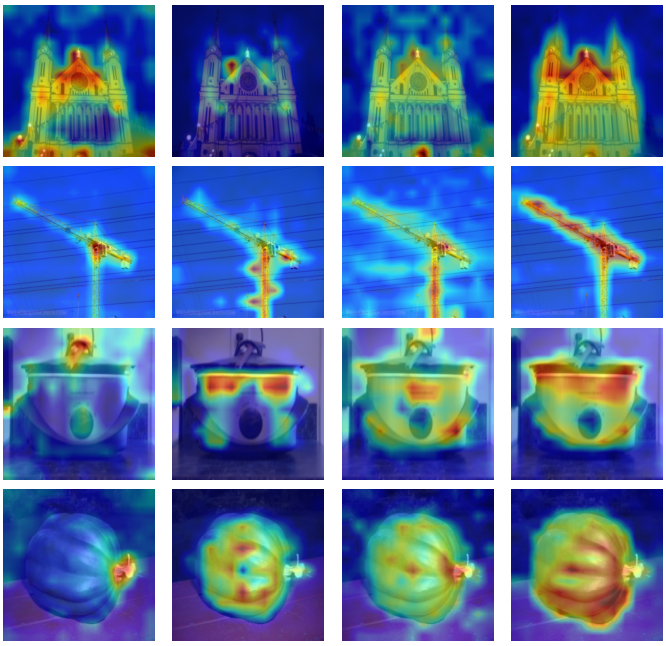
TABLE II
THE RESULT OF PIXEL PERTURBATION TEST ON ILSVRC 2012

Model	Attention Rollout	LRP	AGCAM	Ours (GLM)
LeRF	0.4739	0.5140	0.5298	**0.5350**
MoRF	0.2053	0.1736	**0.1607**	0.1611
ABPC	0.2685	0.3404	0.3691	**0.3739**

To qualitatively validate these metrics, we selected representative images for visualization display in Fig. 3. As observed, the Rollout method primarily focuses on the most discriminative feature regions, often neglecting the rest of the object structure. Although the LRP-based method captures broader areas, it still fails to achieve complete coverage of the entire object. The AGCAM method succeeds in covering the target object but suffers from noticeable noise interference in the background. In contrast, our GLM-based method demonstrates superior performance; it effectively covers the integral object structure with high activation intensity, while maintaining a clean background with significantly less noise. These visual results confirm that our method generates higher-quality heatmaps that stably and accurately reflect the model's region of interest.

To further verify generalization on fine-grained tasks, we conducted comparative experiments on the CUB-200-2011 dataset (Table III). The GLM-based method demonstrates a substantial performance advantage in this challenging scenario. Specifically, our method achieves a Top-1 Localization IoU of 0.5958, significantly outperforming the second-best method, AGCAM (0.4235), by 40.7%. Consistent improvements are also observed in Pixel Accuracy (0.8405) and Dice Score (0.7276). Notably, the high Recall (0.6996) indicates that our method provides more comprehensive target coverage compared to other approaches. Furthermore, the ABPC score reaches 0.6147, which is statistically superior to the comparison methods (paired T-test, $p < 0.05$). These results indicate that our method effectively captures subtle discriminative

features required for fine-grained categorization, proving its strong adaptability across different domains.



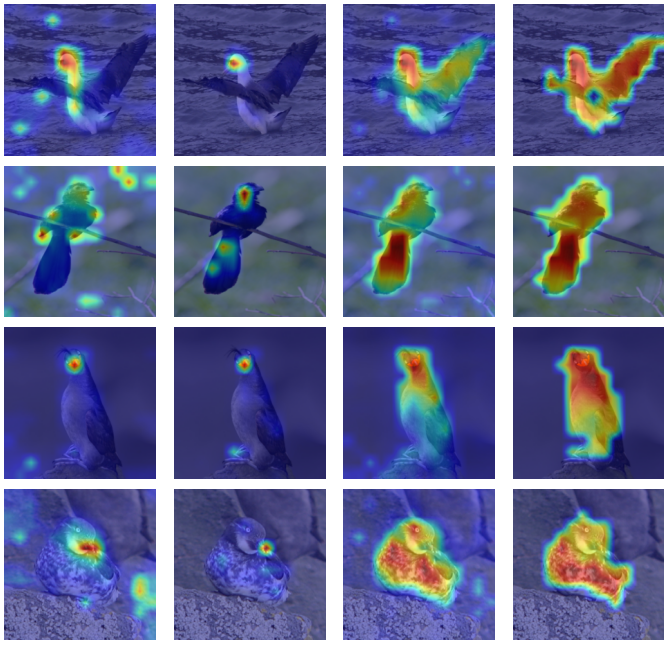
(a) Rollout (b) LRP-based (c) AGCAM (d) Ours

Fig. 3. Heatmaps of ImageNet ILSVRC 2012 dataset generated by different methods.

TABLE III
QUANTITATIVE COMPARISON ON CUB-200-2011 DATASET

Metric	Rollout	LRP	AGCAM	Ours (GLM)
Pixel acc	0.6883	0.6690	0.7919	**0.8405**
IoU	0.1516	0.0649	0.4235	**0.5958**
Dice	0.2388	0.1161	0.5646	**0.7276**
Precision	0.9109	**0.9866**	0.9765	0.8583
Recall	0.1539	0.0650	0.4315	**0.6996**
ABPC	0.5727	0.6112	0.6108	**0.6147**
MoRF	0.1126	**0.0978**	0.1001	0.1002
LeRF	0.6853	0.7090	0.7108	**0.7150**

We selected representative samples for visualization in Fig. 4 to corroborate these quantitative findings. Distinct from general object detection, fine-grained tasks require capturing the object's full extent. As shown, baselines like Rollout and LRP-based tend to fixate solely on local discriminative parts (e.g., the bird's head), resulting in discrete, point-like activations. While AGCAM succeeds in covering the overall object structure, it exhibits lower activation intensity at the boundaries and introduces noticeable background noise. In contrast, our GLM-based method generates high-confidence heatmaps that maintain relatively high intensity throughout the object area, including the boundaries. By effectively suppressing background clutter and ensuring better boundary adherence, our method achieves a more precise alignment with the ground truth, which explains the significant boost in Localization IoU.



(a) Rollout (b) LRP-based (c) AGCAM (d) Ours

Fig. 4. Heatmaps of CUB-200-2011 dataset generated by different methods.

B. Ablation experiments

Through ablation experiments, we investigated the effects of the number of samples and Gaussian noise intensity on the performance of the CAM. As shown in Table IV, the model performs best when the number of samples is 10. For Gaussian noise, moderate noise, e.g., 0.10, helps to improve the model's performance in metrics such as pixel precision, IoU, and dice coefficients, whereas overly strong noise significantly reduces the performance, and especially performs poorly in the capture of target regions, i.e., recall. To summarize, choosing the appropriate number of samples and Gaussian noise intensity are the key factors in generating high-quality CAMs, and too many samples or too strong noise can negatively affect the model localization ability.

TABLE IV
RESULTS OF ABLATION EXPERIMENTS

Parameter	N=5 S=0.10	N=10 S=0.10	N=20 S=0.10	N=10 S=0.05	N=10 S=0.20
Pixel acc	0.7703	**0.7700**	0.7684	0.7666	0.7597
IoU	0.5821	**0.5827**	0.5796	0.5792	0.5670
Dice	0.7055	**0.7057**	0.7028	0.7030	0.6902
Precision	0.8407	**0.8382**	0.8401	0.8315	0.8316
Recall	0.6951	**0.6970**	0.6909	0.6988	0.6801

V. CONCLUSION

In this paper, we presented the GLM-based method, a novel framework designed to enhance the explainability of ViT models by addressing the significant visual noise caused by local gradient fluctuations. By introducing a stochastic sampling strategy that incorporates Gaussian noise into the

input image, our method aggregates CAMs derived from both gradient and attention matrices, effectively neutralizing noise instability through a smoothing process.

Extensive comparative experiments on the ImageNet ILSVRC 2012 and CUB-200-2011 datasets have validated the effectiveness of our approach compared to representative baseline methods. On the fine-grained CUB-200-2011 dataset, the GLM-based method achieved a Localization IoU of 0.5958, significantly outperforming the AGCAM method (0.4235) by 40.7%. Furthermore, in pixel perturbation tests conducted on both datasets, the proposed method yielded superior ABPC scores (0.3739 on ImageNet and 0.6147 on CUB-200-2011), demonstrating a statistically significant improvement ($p < 0.05$) over the comparison methods. Qualitative visualizations further confirm that our approach produces smoother heatmaps with reduced background clutter, successfully capturing the full extent of object structures rather than limiting focus to discrete discriminative regions.Additionally, our ablation studies identified that a sample size of 10 combined with a noise intensity of 0.10 strikes the optimal balance for generating precise CAMs.

In summary, the GLM-based method significantly improves the transparency and trustworthiness of ViT models, providing a reliable basis for decision-making in complex visual tasks such as image classification and fine-grained target detection.

REFERENCES

[1] Alexey, Dosovitskiy. "An image is worth 16x16 words: Transformers for image recognition at scale." arxiv preprint arxiv: 2010.11929 (2020).

[2] Vaswani, A. "Attention is all you need." Advances in Neural Information Processing Systems (2017).

[3] Tanzi, Leonardo, et al. "Vision transformer for femur fracture classification." Injury 53.7 (2022): 2625-2634.

[4] Wang, Wei, et al. "A ViT-based multiscale feature fusion approach for remote sensing image segmentation." IEEE Geoscience and Remote Sensing Letters 19 (2022): 1-5.

[5] Strudel, Robin, et al. "Segmenter: Transformer for semantic segmentation." Proceedings of the IEEE/CVF international conference on computer vision. 2021.

[6] Touvron, Hugo, Matthieu Cord, and Hervé Jégou. "Deit iii: Revenge of the vit." European conference on computer vision. Cham: Springer Nature Switzerland, 2022.

[7] Yin, Hongxu, et al. "A-vit: Adaptive tokens for efficient vision transformer." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

[8] Tumanyan, Narek, et al. "Splicing vit features for semantic appearance transfer." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[9] Han, Kai, et al. "A survey on vision transformer." IEEE transactions on pattern analysis and machine intelligence 45.1 (2022): 87-110.

[10] Wu, Bichen, et al. "Visual transformers: Where do transformers really belong in vision models?." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[11] Yang, Michael. "Visual transformer for object detection." arxiv preprint arxiv:2206.06323 (2022).

[12] Chen, Haoyuan, et al. "GasHis-Transformer: A multi-scale visual transformer approach for gastric histopathological image detection." Pattern Recognition 130 (2022): 108827.

[13] Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." Proceedings of the IEEE international conference on computer vision. 2017.

[14] Chattopadhay, Aditya, et al. "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks." 2018 IEEE winter conference on applications of computer vision (WACV). IEEE, 2018.

[15] Yang, Seunghan, et al. "Combinational class activation maps for weakly supervised object localization." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2020.

[16] Qin, Zhenyue, Dongwoo Kim, and Tom Gedeon. "Rethinking softmax with cross-entropy: Neural network classifier as mutual information estimator." arxiv preprint arxiv:1911.10688 (2019).

[17] Draelos, Rachel Lea, and Lawrence Carin. "Use HiResCAM instead of Grad-CAM for faithful explanations of convolutional neural networks." arxiv preprint arxiv:2011.08891 (2020).

[18] Zhao, Chenyang, et al. "Gradient-based Visual Explanation for Transformer-based CLIP." International Conference on Machine Learning. PMLR, 2024.

[19] Ma, Jie, et al. "Visualizing and understanding patch interactions in vision transformer." IEEE Transactions on Neural Networks and Learning Systems (2023).

[20] Mallick, Rupayan, Jenny Benois-Pineau, and Akka Zemmari. "I saw: a self-attention weighted method for explanation of visual transformers." 2022 IEEE international conference on image processing (ICIP). IEEE, 2022.

[21] Abnar, Samira, and Willem Zuidema. "Quantifying attention flow in transformers." arxiv preprint arxiv:2005.00928 (2020).

[22] Chefer, Hila, Shir Gur, and Lior Wolf. "Transformer interpretability beyond attention visualization." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.

[23] Leem, Saebom, and Hyunseok Seo. "Attention Guided CAM: Visual Explanations of Vision Transformer Guided by Self-Attention." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 38. No. 4. 2024.

[24] Zhou, Bolei, et al. "Learning deep features for discriminative localization." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[25] Smilkov, Daniel, et al. "Smoothgrad: removing noise by adding noise." arxiv preprint arxiv:1706.03825 (2017).

[26] Bishop, Chris M. "Training with noise is equivalent to Tikhonov regularization." Neural computation 7.1 (1995): 108-116.

[27] Russakovsky, Olga, et al. "Imagenet large scale visual recognition challenge." International journal of computer vision 115 (2015): 211-252.

[28] Wah, Catherine, et al. "The Caltech-UCSD Birds-200-2011 Dataset." California Institute of Technology (2011).

[29] R. Wightman, "PyTorch Image Models," GitHub repository, 2019. [Online]. Available: https://github.com/rwightman/pytorch-image-models

[30] Samek, Wojciech, et al. "Evaluating the visualization of what a deep neural network has learned." IEEE transactions on neural networks and learning systems 28.11 (2016): 2660-2673.