*Proceedings of the 11th World Congress on Electrical Engineering and Computer Systems and Sciences (EECSS'25)*
*Paris, France - August, 2025*
*Paper No. MVML 138*
*DOI: 10.11159/mvml25.138*

# G3D-Vit, A Small Step Towards 3D Interpretability Of Vision Transformers

**Thomas Gillet[1], Marta Oliveira[1], Ben Bausch[1], Diego Andrés Blanco-Mora[1], Jorge Gonçalves[1], Esmeralda Blaney Davidson[2], Marian Van Der Meulen[1]**
[1]University of Luxembourg
2 Av. de l'Université, 4365 Esch-Belval Esch-sur-Alzette
thomas.gillet@uni.lu
[2]Radboud University Medical Center
Geert Grooteplein Zuid 10, 6525 GA Nijmegen, Netherlands

## Extended Abstract

Vision Transformers [1] (ViTs) have demonstrated strong performance in medical image analysis [2], yet their interpretability on complex 3D data such as MRI remains limited. We introduce G3D-ViT, a novel 3D GradCAM approach for ViTs, that generates gradient-based class activation maps, identifying relevant spatial regions in 3D volumes. Originating from the fMRI2Vec project [3], this work provides insights into brain regions driving predictions in MRI-based tasks such as gender or age classification.

To benchmark and motivate our method, we first evaluated existing interpretability techniques for both Convolutional Neural Networks (CNN) [4, 5, 6] and ViT [4, 7, 8, 9, 10, 11] with our 3D data, which consisted of in-house resting-state fMRI volumes. In experiments using a 3D ResNet, some CNN-based methods (e.g., LayerCAM and GradCAMElementWise) [3] yielded consistent results, while others were unreliable or incompatible. Regarding ViTs, the tested interpretability tools often lacked 3D support or depended on specific architectures. These evaluations helped establish a reliable ground truth to benchmark ViT interpretability in 3D models, against which ViT-based approaches could be assessed.

To address these gaps, we developed G3D-ViT a 3D GradCAM for Vision Transformers. Our approach registers hooks on the final normalization layer of a 3D ViT architecture (specifically, a direct implementation of Phil Wang's 3D ViT [12]). These hooks capture activations during the forward pass and gradients during the backward pass. We generate 3D GradCAM maps by weighting activations with their corresponding gradient importance scores, highlighting the most influential regions for classification. We explored gradient averaging methods, finding that Mean Average Pooling provided effective results, and an optional thresholding step further refines the attention maps.

To validate G3D-ViT, we used a synthetic 3D dataset consisting of a large cube with an embedded target cube. A 3D ViT was trained to classify the target cube's spatial position and its content. G3D-ViT accurately localized the target cube across aligned (at exact grid multiples of its size), unaligned (at random) cube positions, and noisy scenarios. Interpretability was influenced by the patch size relative to the target cube, with patches smaller than the region of interest offering better localization. Finally, among the explored gradient averaging methods, Mean Average Pooling (our default configuration) consistently proved to be the most effective in most cases.

G3D-ViT successfully extends GradCAM to 3D ViTs, offering a valuable tool for interpreting model decisions on 3D data. Our validation on synthetic data demonstrates its ability to accurately highlight relevant spatial regions. Preliminary tests on our resting-state fMRI volumes also showed encouraging results. However, further evaluation on more 3D datasets is necessary to assess the 3D GradCAM reliability in clinical scenarios.

The full code implementation and results of G3D-ViT can be found at https://github.com/gillet-thomas/G3D-ViT.

## References


[1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, & N. Houlsby. (2021). "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale", in *Proceedings of the International Conference on Learning Representations*.

[2] R. Azad, A. Kazerouni, M. Heidari, E. K. Aghdam, A. Molaei, Y. Jia, A. Jose, R. Roy, & D. Merhof. (2023). "Advances in Medical Image Analysis with Vision Transformers: A Comprehensive Review", *arXiv preprint arXiv:2301.03505*.

[3] T. Gillet. (2025). fMRI2vec [Online]. Available: https://github.com/gillet-thomas/fMRI2vec

[4] J. Goldblatt. (2022). pytorch-grad-cam [Online]. Available: https://github.com/jacobgil/pytorch-grad-cam

[5] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, & O. Reblitz-Richardson. (2020). "Captum: A unified and generic model interpretability library for PyTorch", *arXiv preprint arXiv:2009.07896*.

[6] S. Lundberg and S.-I. Lee. (2017). "A Unified Approach to Interpreting Model Predictions", in *Proceedings of the Conference on Neural Information Processing Systems*.

[7] H. Chefer, S. Gur, and L. Wolf. (2021). "Transformer Interpretability Beyond Attention Visualization", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[8] E. Voita, D. Talbot, F. Moiseev, R. Sennrich, & I. Titov. (2019). "Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned", *arXiv preprint arXiv:1905.09418*.

[9] J. Gildenblat. (2020). vit-explain [Online]. Available: https://github.com/jacobgil/vit-explain

[10] W. Bousselham. (2024). LeGrad [Online]. Available: https://github.com/WalBouss/LeGrad

[11] C. D. Pierse. (2021). transformers-interpret [Online]. Available: https://github.com/cdpierse/transformers-interpret

[12] P. Wang. (2020). vit-pytorch [Online]. Available: https://github.com/lucidrains/vit-pytorch