One Wave To Explain Them All: A Unifying Perspective On Feature Attribution

Gabriel Kasmi [1,2], Amandine Brunetto [1], Thomas Fel [3], Jayneel Parekh [4]
[1] Mines Paris - PSL University [2] RTE France [3] Kempner Institute, Harvard University [4] ISIR, Sorbonne Université, France
Contact: gabriel.kasmi@minesparis.psl.eu

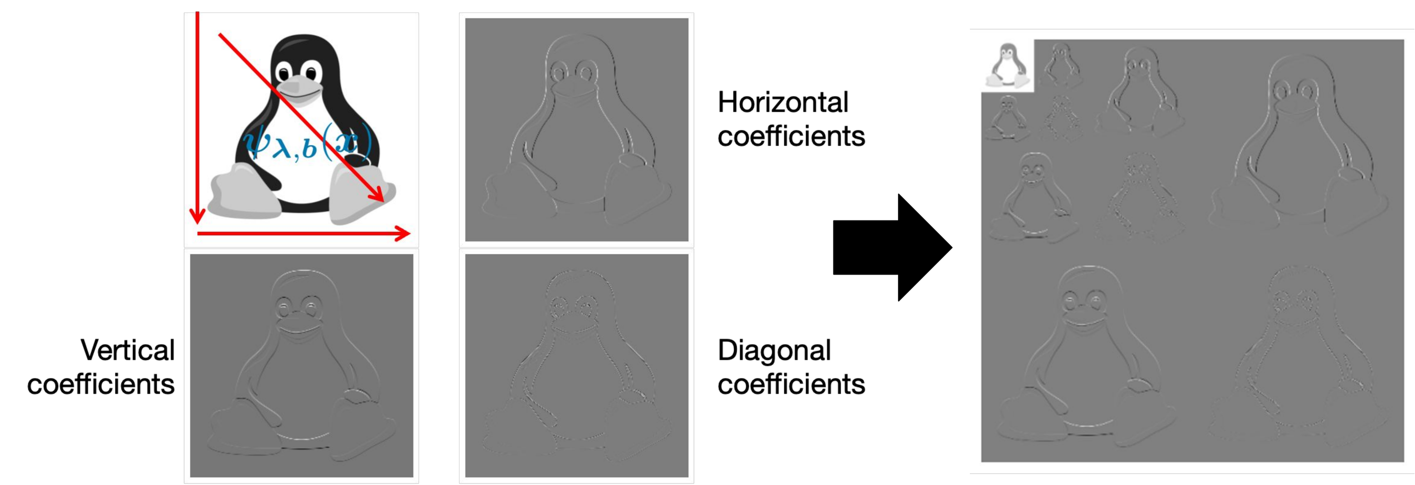
Background and Motivation

Feature attribution projects an explanation as a **pixel-based heatmap**.

Two main limitations:
- Ill-suited for **non-image modalities**
- **Overlooks structural information** of the input

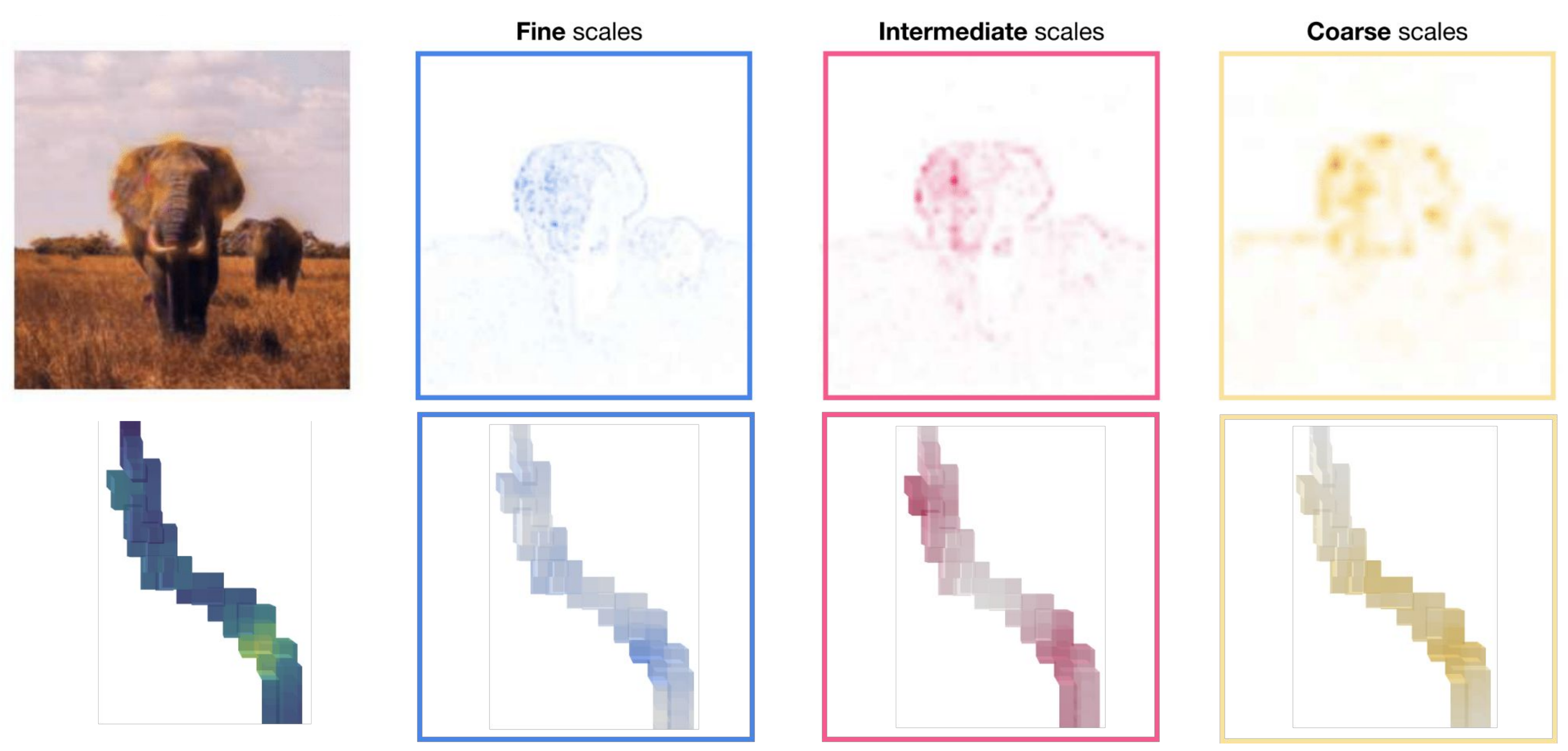
Surprisingly, only a handful of works explored feature attribution in alternative domains than the input domain.

Multiscale decompositions

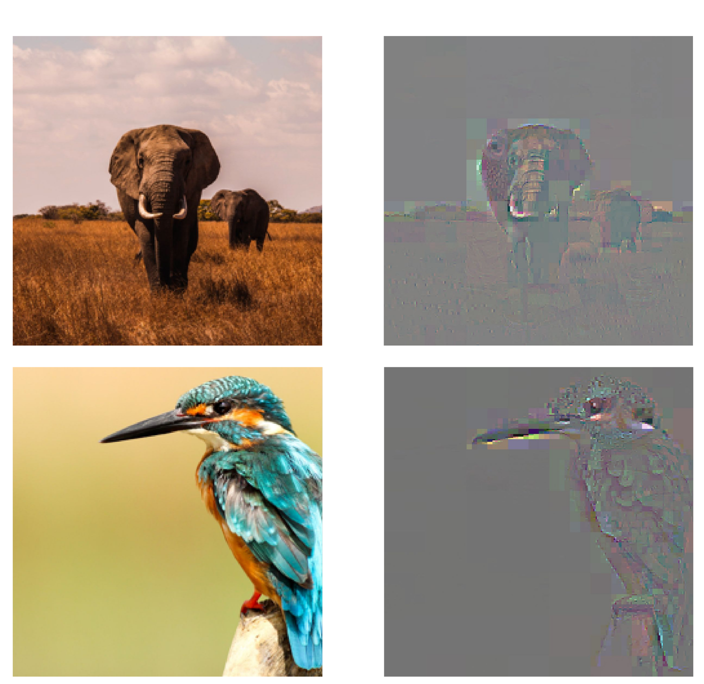


Wavelets [1] decompose an input into structural components which are **localized in space and scale**. From an explainability perspective, one can view wavelet coefficients as **low-level features.**

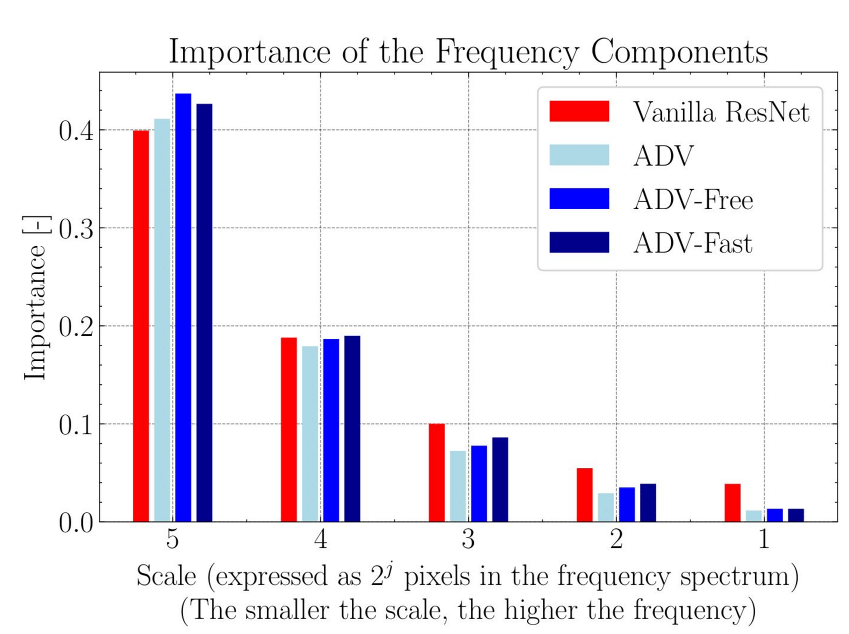
Applications



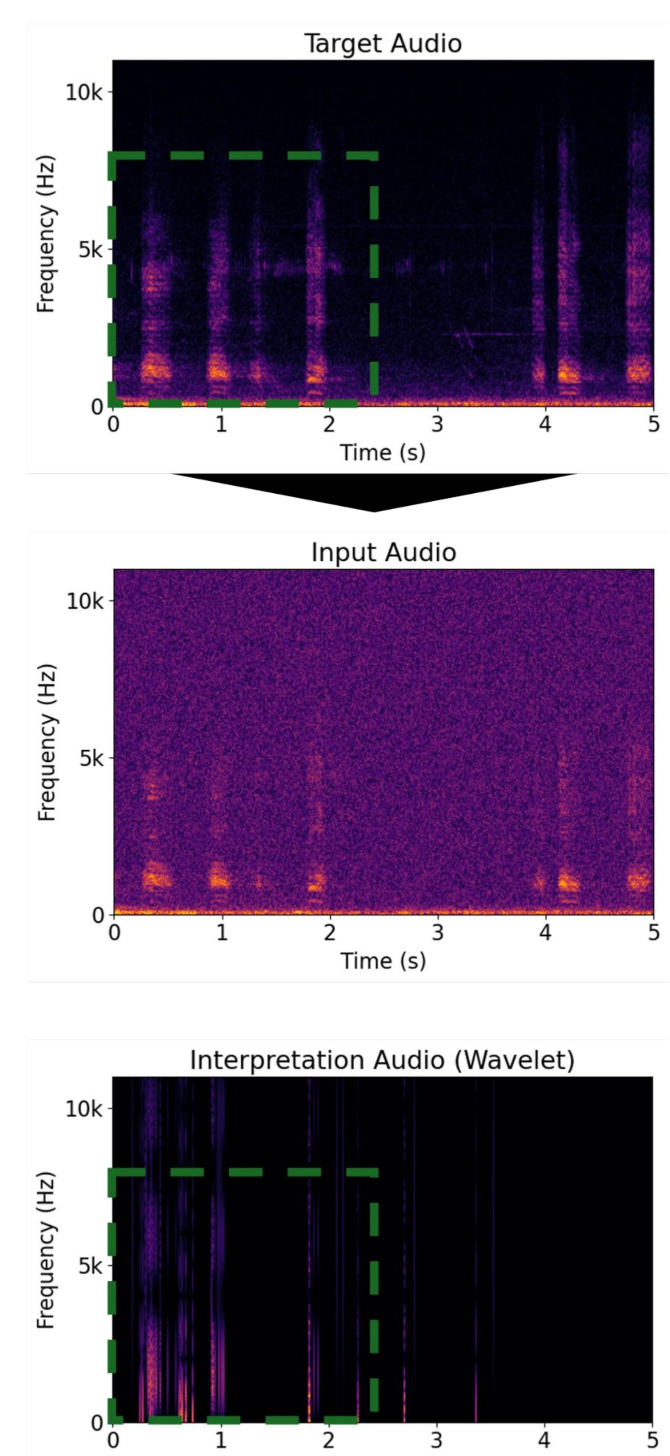
Explanations beyond important locations: decomposition of the **importance across scales**, which correspond to **different interpretable components**.



Meaningful perturbations meets the wavelet domain
Extract the **most important components** (and not only location) to highlight **texture biases**



Assessment of a model's robustness
WAM highlights the **reliance on frequency ranges** and therefore bridges the gap with **frequency-centric perspectives on model robustness** [2]



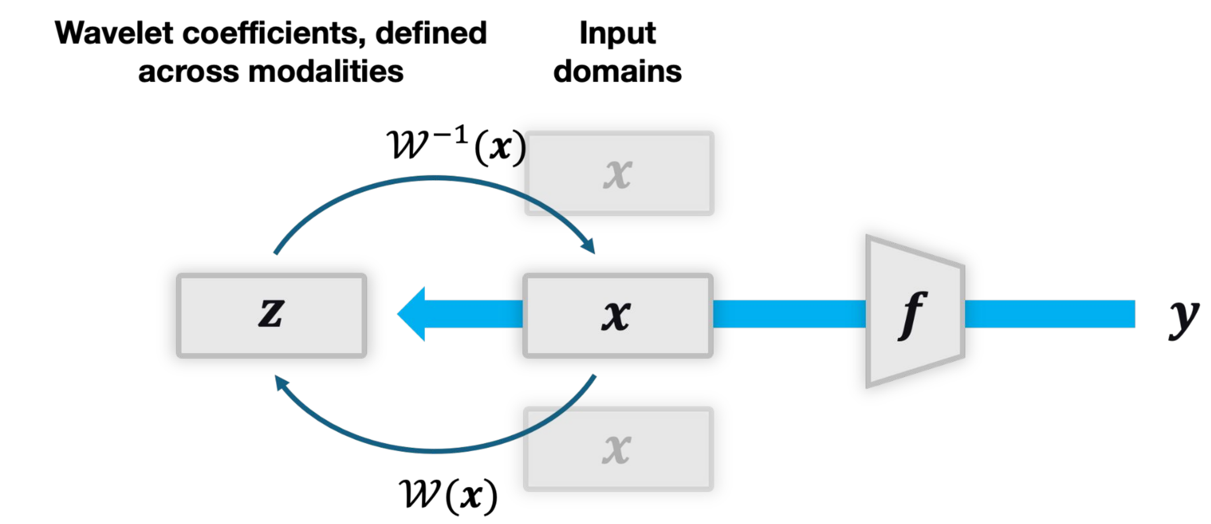
Reconstruction from **noisy audio** (also works for **audio overlaps**)

Git

Project page

[1] Mallat, S. (1999). *A wavelet tour of signal processing*. Elsevier.
[2] Chen, Y., Ren, Q., & Yan, J. (2022). Rethinking and Improving Robustness of Convolutional Neural Networks: a Shapley Value-Based Approach in Frequency Domain. Advances in neural information processing systems, 35, 324-337.

Proposed approach

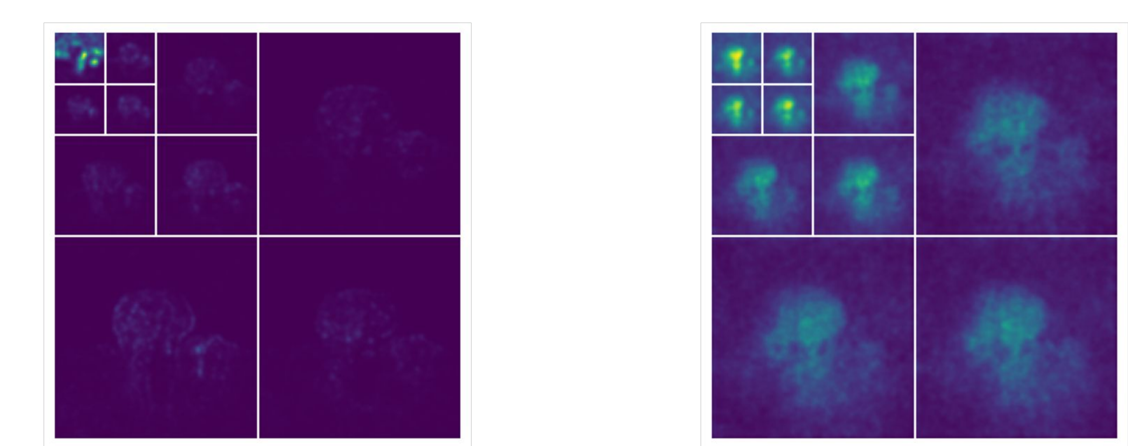


1. Compute gradients

Compute the gradients with respect to the **wavelet coefficients** of the input modality

$$\gamma_{\mathrm{IG}} = (z - z_0) \cdot \int_0^1 \frac{\partial f_c\left(\mathcal{W}^{-1}\left(z_0 + \alpha(z - z_0)\right)\right)}{\partial z} \mathrm{d}\alpha$$

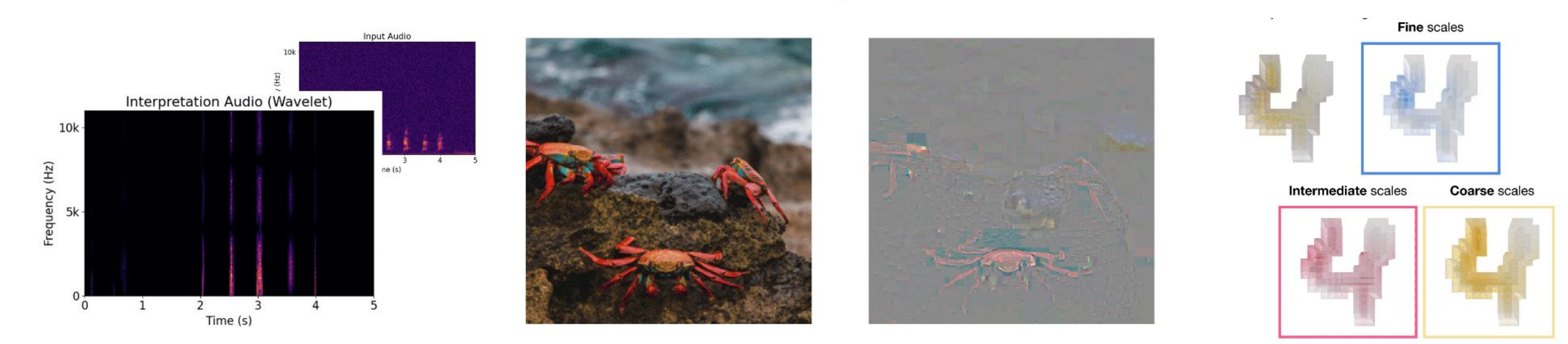
$$\gamma_{\mathrm{SG}}(z) = \frac{1}{n}\sum_{i=1}^{n} \nabla_{\tilde{z}} f(\mathcal{W}^{-1}(\tilde{z}))$$



2. Aggregate gradients

Path integration: highlights the inter-scale dependencies
Smoothing: focuses on intra-scale importance

Interpretation audio spectrograms — **Multiscale** heatmap **decomposition** — Attribution **on voxels**



Emphasis on **key audio parts** — Identification of **minimal images** — **Decomposition** across scales

Evaluation

WAM performs consistently across a wide range of **metrics, model topologies** and in **additional evaluation settings**.
It **passes the randomization check.**

	Audio			Images			Volumes		
Model	*ResNet*			*EfficientNet*			*3D Former*		
Dataset	*ESC-50*			*ImageNet*			*AdrenalMNIST3D*		
	Ins (↑)	Del (↓)	Faith (↑)	Ins(↑)	Del (↓)	Faith (↑)	Ins (↑)	Del (↓)	Faith (↑)
Integrated Gradients	0.267	**0.047**	**0.264**	0.113	0.113	0.000	0.666	0.743	-0.077
SmoothGrad	0.251	0.067	0.184	0.129	0.119	0.010	0.680	0.731	-0.051
GradCAM	0.274	0.201	0.072	0.364	0.303	0.061	0.689	0.744	-0.055
Saliency	0.220	0.154	0.066	0.148	0.140	0.008	0.751	0.742	0.009
WAM_{IG} (ours)	0.436	0.260	0.176	**0.447**	**0.049**	**0.370**	**0.719**	**0.621**	**0.098**
WAM_{SG} (ours)	**0.449**	0.252	0.197	0.419	0.097	0.350	0.718	0.648	0.070