From Weight Perturbation to Feature Attribution for Explaining Fully Connected Neural Networks.

Thodoris Lymperopoulos
NCSR Demokritos
Athens, Greece
t.lymperopoulos@iit.demokritos.gr

Denia Kanellopoulou
NCSR Demokritos
Athens, Greece
denia@iit.demokritos.gr

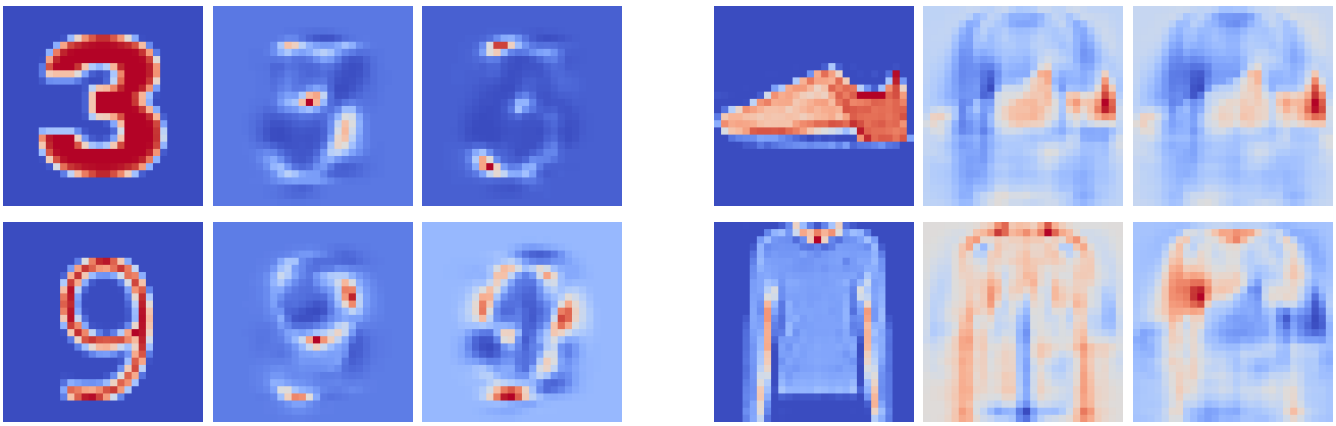

Figure 1: The produced attribution maps for our proposed methods, using weight perturbation for estimating feature attribution.

Abstract

Fully Connected Neural Networks (FCNNs) are often regarded as simple and intuitive architectures, yet they serve as the foundation for more complex models. Nonetheless, the lack of consensus on their interpretability continues to pose challenges, underscoring the enduring relevance of simpler, attribution-based approaches for understanding even the most advanced neural architectures. In this regard, we explore a novel idea for estimating feature attribution, by applying perturbation to the features' attached weights instead of their values. This method offers a fresh perspective aimed at mitigating common limitations in Occlusion techniques, such as Added Bias and Out-of-Distribution data. The application of this rule leads to the formation of a pair of novel attribution methods we call XWP and XWP^c. Founded on simple rules, our methods achieve competitive performance in identifying image signals for simple DNNs, competing with the most established attribution methods on standard baseline metrics. Our work thus contributes to the field of Explainability by introducing a robust framework that paves the way for addressing these long-standing vulnerabilities, and leads to more reliable and interpretable model explanations.

CCS Concepts

• **Computing methodologies** → *Feature selection*; **Machine learning**.

Keywords

Explanable AI, Interpretability, Attribution Methods, Occlusion Techniques

ACM Reference Format:
Thodoris Lymperopoulos and Denia Kanellopoulou. 2026. From Weight Perturbation to Feature Attribution for Explaining Fully Connected Neural Networks.. In *Proceedings of (Manuscript under review)*. ACM, New York, NY, USA, 9 pages.

1 Introduction

Modern deep learning models have achieved remarkable performance across a wide range of tasks, from image recognition to natural language processing. Yet, this success comes at a cost: as models grow in scale and complexity, their internal decision-making process becomes increasingly opaque, rendering them, in effect, black boxes. This opacity poses significant challenges in high-stakes domains such as medicine and autonomous systems, where understanding and trusting a model's predictions is as important as the predictions themselves. The field of Explainable AI (XAI) [3] emerged precisely to address this gap, seeking to provide human-interpretable insights into the model's behavior [16].

Attribution methods represent the earliest and most widely adopted family of approaches within XAI, aiming to quantify the contribution of each input feature to the model's output [31]. The core intuition is to produce a saliency map — a score assigned to each input dimension — reflecting its relative importance for a given prediction. Over time, a range of methodologies has been developed to estimate these scores. Gradient-based methods, such as Vanilla Gradients [24] and Integrated Gradients [26], derive these

Manuscript under review,
2026.

scores directly from the model's gradients with respect to the input. Backpropagation-based methods, such as Layer-wise Relevance Propagation (LRP) [5] and DeepLIFT [23], instead propagate the output signal backwards through the network, redistributing relevance across layers according to defined propagation rules. Perturbation-based methods, such as LIME [21] and SHAP [17], approximate the model locally by observing how the output changes as portions of the input are systematically masked or altered. Collectively, these methods constitute a robust framework for interpreting a model's decisions.

In a separate body of work [19] authors introduced a complementary perspective, applying well-established techniques — weight visualization and neuron optimization — to probe the model's internal representations directly. Yet, the scale and complexity of modern architectures rendered manual inspection of individual components intractable, motivating the development of Mechanistic Interpretability (MI) [4] — a systematic framework for reverse-engineering a model's decision-making process by identifying the circuits and components responsible for specific computations. As the field matured, attention shifted to more complex architectures [9], and MI established a natural connection to the field of Causality [12, 15]: both disciplines share a commitment to interventional reasoning, seeking not merely to correlate inputs with outputs but to identify the underlying mechanisms that govern model behavior. This convergence gave rise to structured methods for uncovering causal relationships among model components, yielding foundational insights into how neural networks encode and process information [4, 12].

Despite the recent success of MI to interpret different aspects of the model's inner processes, no definite consensus on the model's interpretability has been reached to this day, even for the simplest among complex architectures — the Fully Connected Neural Networks (FCNNs) [9]. This comprises an essential challenge, since FCNNs serve as core components in Transformers [27]. In this work, we depart from conventional occlusion approaches by revisiting the fundamentals of model computation through the lens of weight perturbation. Adopting this perspective, we formulate a novel framework for attribution estimation, which circumvents common pitfalls of input perturbation in existing methodologies, such as the problem of Out-of-Distribution data. Within this framework, we introduce two variants—XWP and XWP^c— designed to assess attribution scores with improved accuracy. We validate their efficacy on FCNNs, where heatmaps generated for the Typeface MNIST and Fashion MNIST datasets (Figure 1) reveal clean, input-aligned patterns, devoid of the artifacts prevalent in prior work. Together, these variants provide a synergistic approach to explaining a model's decision-making process, establishing a novel paradigm for perturbation-based interpretability. Our code is available at: https://anonymous.4open.science/r/Explainable-Weight-Perturbation-0DDC

2 Related Work

Occlusion Methods. They comprise a prominent family of attribution methods relying on occlusion and perturbation of input data to quantify the contribution of individual features or regions to a model's prediction. One of the earliest approaches, Occlusion Sensitivity [31], slides a mask patch across an input image and observes the resulting change in output confidence, producing a saliency map that reflects each region's importance. Building on this principle, LIME (Local Interpretable Model-agnostic Explanations) [21] generates perturbed samples around a given input, weights them by their proximity to the original, and fits a sparse linear model to approximate the local decision boundary of any black-box classifier. Similarly, SHAP (SHapley Additive exPlanations) [17] grounds feature attribution in cooperative game theory, computing Shapley values that fairly distribute the prediction output among all input features by averaging their marginal contributions across all possible feature subsets — a process inherently based on systematic feature inclusion and exclusion.

Subsequent work has expanded and refined perturbation-based attribution in several directions. KernelSHAP [17] and SampleSHAP [2] provide computationally efficient approximations of Shapley values, since its estimation is computationally intractable. Meaningful Perturbations [11] formulates the search for an explanatory mask as a continuous optimization problem, penalizing perturbations that minimally alter the input while maximally reducing model confidence. More recently, RISE (Randomized Input Sampling for Explanation) [20] estimates importance maps by feeding thousands of randomly masked versions of an image through the network and computing a weighted average of the masks, scaled by the corresponding output scores. Collectively, however, these methods suffer from severe theoretical limitations, particularly in the process of selecting the baseline values that are meant to represent the absence of information. Specifically, the Added Bias problem [10, 14, 25] typically lead to information addition with respect to the model's inherent biases and image characteristics, while the Out-Of-Distribution problem [13] highlights that the resulting perturbed samples may lie outside the margins of the training data distribution, rendering the model's responses on such inputs unreliable as a basis for attribution.

Weight Contribution. A different line of approaches focuses on the critical role of weights to the model's decision making process. *Weight perturbation* is based on techniques for altering a model's parameters and is used for improving the model's learning process [18], robustness to adversarial examples [29] and data reconstruction [22]. In the context of interpretability, authors of [19] found matching pairs of weights and optimal neuronal activations for interpreting individual neurons. Authors of [1] replaced standard linear transformations in deep networks with a new B-cos operation that enforces alignment between model weights and task-relevant input patterns during training. This *ante-hoc* approach manages to effectively summarize model behaviour in a single, highly interpretable linear map. In another line of work, *Embedded Methods* [7] introduced an artificial layer between the input and the first hidden layer to track the evolution of training statistics—most commonly weights and gradients - which are then used to infer feature importance using different techniques to the collected gradient or weight profiles. However, the interpretation of these statistics remains elusive.

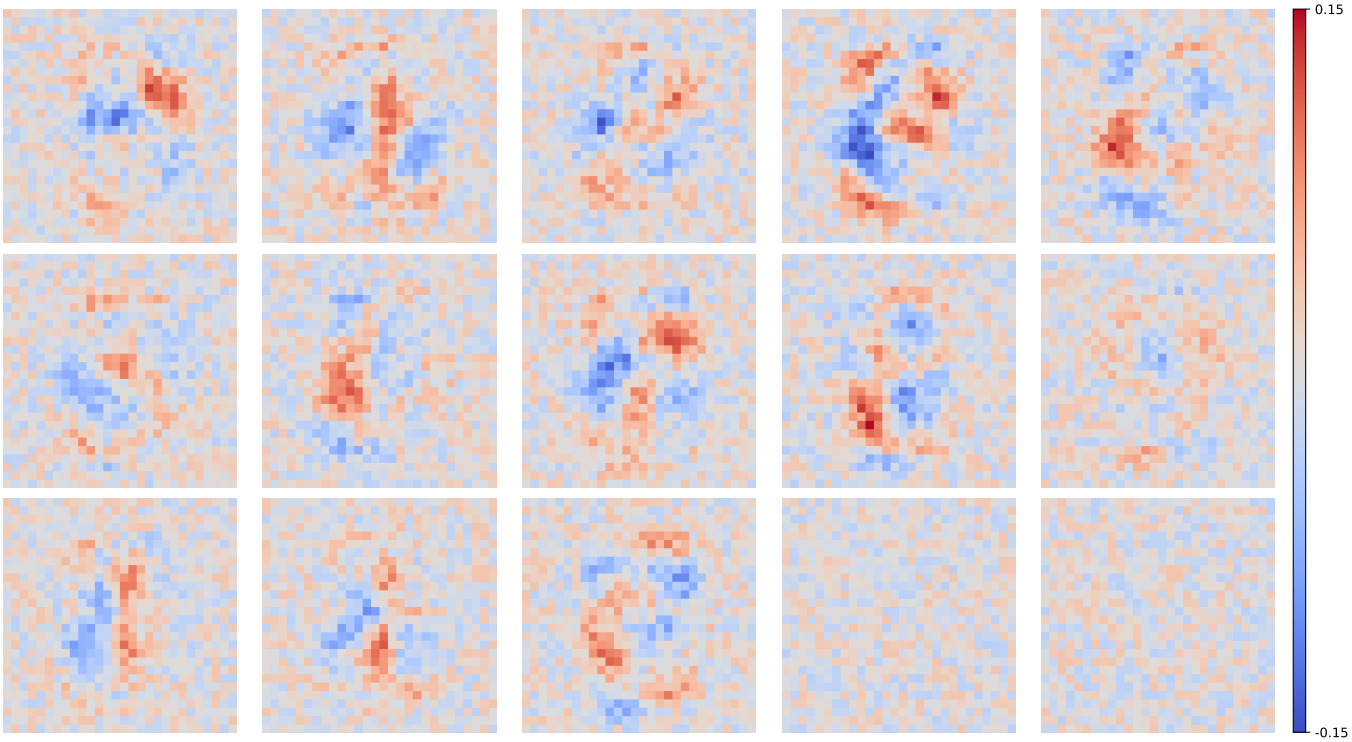

Figure 2: A visualization of first layer weights of a trained FCNN on TMNIST, for specific neurons of the first layer. Distinct patterns corresponding to the digits 0-9 can be identified (first two rows), among other, more abstract patterns (third row). The case for pattern absence is also present (last two images), signifying neutral neurons with unclear functionality.

3 Methodology

3.1 Mathematical Annotation

Consider an arbitrary fully connected neural network (FCNN) $f = f^L \circ f^{L-1} \circ ... \circ f^0$ receiving input $x \in \mathcal{X}$, from the input space. Let l denote an arbitrary inner layer of the network, consisting of n_l neurons with weights $W^l \in \mathbb{R}^{n_{l-1} \times n_l}$ and bias $b \in \mathbb{R}^l$. The layer's operation is described by the output z^l obtained by applying a non-linear activation function to the input x^l_j. Thus, for neurons i and j in layers $l-1$ and l, respectively, the following holds:

$$x^l_j = \sum_{i \in [n_{l-1}]} z^{l-1}_i * w^l_{ij} + b^l_j, \tag{1}$$

$$z^l_j = \sigma(x^l_j). \tag{2}$$

σ is the ReLU activation function, applied to all intermediate layers (except the final layer, where Softmax is used). For a neuron i in the input layer, we adopt the annotation $W_i = \{w_{ij} | j \in l^1\}$ denoting the set of its attached weights of the first layer, which will be our primary focus in our research.

3.2 Model Parameters and Pattern Formation

Our methodology treats a deep model as a computational graph—a structured sequence of operations involving neuron activations and model parameters. In this framework, parameters act as control mechanisms, shaping the formation of meaningful representations to minimize loss and enable accurate decision-making. Their role is thus pivotal for aligning machine and human objectives, comprising the starting point of our research towards interpreting the resulting patterns and representations.

In the context of FCNNs, the visualization of weights connecting input features to specific neurons of the first layer $W^T_j = \{w_{ij} | iin[n^0]\}$ reveals patterns that are interpretable (as demonstrated in Figure 2)) and can be well understood: they illustrate the model's motifs that either enhance or suppress a neuron's activation. Such dual behavior reflects the polysemantic nature of neurons [19], where a single unit integrates multiple feature detectors. In the first layer of FCNNs, this functionality manifests as strong activation in response to specific patterns and suppression in response to others.

The primary contribution of visualizing weights W^T_j is the revelation of spatial structure among input features. This is a key observation which will be exploited in our work. However, the individual interpretation of a weight does not directly translate to estimating the contribution of input neurons, which collectively influence all patterns encoded in the weights. In this regard, authors in [19] advanced our understanding by constructing optimal input signals: maximizing activation for specific inner neurons and matching their patterns with those of the neuron's weights. Although this approach enables interpretation of inner neurons and introduces a novel framework for integrating neuronal interactions with weight patterns, it does not generalize to a quantitative measure of neuronal contribution.

3.3 Introducing Explainable Weight Perturbation

We introduce a novel attribution method, bridging the gap between pattern formation and neuronal contribution. This is achieved through a theoretically grounded framework rooted in the principles of Occlusion. We term our method **Explainable Weight Perturbation** (XWP), operating by assessing neuronal contribution as a distance measure of the model's response at the transition of values W_i from the initial to the trained model state.

Formally, we denote $f(x)$ as the output of the trained model, and $f_{W_i=W_i^o}(x)$ being the targeted replacement of parameters W_i of the trained model by those of the untrained state—denoted as W_i^o. XWP applies the following rule:

$$R_i(x) = \mathcal{I}^t * (f(x) - f_{W_i=W_i^o}(x)), \tag{3}$$

where $\mathcal{I}$ equals one when output neuron n^l matches the target t:

$$\mathcal{I}^t_{n^L} = \begin{cases} -1, & \text{if } n^L \neq t \\ 1, & \text{if } n^L = t. \end{cases} \tag{4}$$

It thus functions as a reward signal, by positively awarding increases in target neuron prediction (or decreases in non-target neurons) in this weight transition, signaling a favourable change of i's contribution to the model's prediction capacity. Any other case meets its unfavorable response.

Unlike traditional Occlusion approaches, our method follows a distinct path, with perturbations unfolding in the parameter instead of the input space. It leverages the model's initial state, considering it a blank spot, where mechanisms detecting patterns are in their early stage. The application of the distance measure permits the model's pattern formation to be expressed—through its effect on the input samples—while guiding the selection of the baseline value by the model's state. This approach introduces no human bias or ambiguous statistical rules, enabling the safe and reliable application of the Occlusion rule.

In equation 3 the model's inner parameters are selected from the trained state, while in constrast to values of W_i, their values remain unaltered. This static consideration of intermediate parameters stems from the assumption that their effect is equally shared among input neurons. This in turn permitted the consideration of input and inner representations as being decoupled, and the application of changes only to the first—further enabling the exploitation of the input's spatial structure. Despite this, this independence is deceptive, since the training algorithm considers their effect when applying the chain rule to weights W_i, rendering our method accountable for internal neuronal interactions.

Having defined a method for identifying and isolating the contribution of feature i to pattern formation, we can express its complementary rule XWP^c, as follows:

$$R_i^c(x) = \mathcal{I}^t * (f(x) - f_{W_{i'}=W_{i'}^o}(x)), \tag{5}$$

where $i' \neq i$ represents all neurons of the input layer differing from i.

To summarize the role of each rule, they jointly perturb weights in the direction indicated by the model's trained state, to derive a feature importance score based on the model's response. XWP achieves this for feature i by measuring its effect when excluding it from the trained model, while XWP^c assesses this effect by dropping the values of all other features back to their initial states. Expressed in simpler terms, a feature in XWP asks itself *"what did I achieve for the model's capabilities considering all updates"* while in XWP^c the underlying question is *"what did I achieve only by myself"*.

We move forward to testing our proposed variants for two baseline datasets in Section 4, while identifying key characteristics in their produced heatmaps and comparing them with state-of-the-art approaches.

4 Experiments

To evaluate the effectiveness of our framework, we conduct a comprehensive experimental analysis of its two variants, applied to FCNNs trained on the Typeface MNIST and Fashion MNIST datasets. We employ a standard FCNN architecture with ReLU activations in the intermediate layers and a softmax output layer. The model consists of four layers with dimensions (784, 400, 100, 10), respectively. Training continues until convergence, typically achieving test accuracy exceeding 90% within 10 epochs. All experiments are performed on an NVIDIA GeForce RTX 4060 GPU.

We then compute the attribution scores generated by our proposed variants and compare them against those produced by five state-of-the-art baseline methods: Occlusion (OCCL) [31], Shapley values (SHAP) [17], RISE [20], Integrated Gradients [26] and LRP [6]. Our evaluation combines quantitative metrics with qualitative analysis, including the visualization of computed heatmaps. The detailed results are presented in Section 4.3.

4.1 Datasets

TMNIST (Typeface-MNIST). This dataset [28] is composed of 29,900 MNIST-style grayscale 28x28 images of digits from 0 to 9, rendered across a variety of Google font styles. Unlike handwritten digit datasets, the typeface-based nature of TMNIST produces clean, well-defined digit shapes with minimal noise and ambiguity. This property makes it particularly well-suited for the evaluation of attribution methods, as the clarity of the digit structures allows for a more unambiguous interpretation of the resulting attribution maps.

FashionMNIST. This is a dataset of grayscale images [30] introduced as a more challenging alternative to the original MNIST benchmark. It consists of 70,000 28x28 images spanning ten classes of clothing items, including t-shirts, trousers, and ankle boots, split into 60,000 training and 10,000 test samples. Despite its simplicity in resolution, the dataset presents non-trivial classification challenges due to inter-class visual similarity, making it a standard benchmark for evaluating models on simple image recognition tasks.

4.2 Evaluation Metrics

Deletion AUC. The Deletion AUC criterion [20] evaluates an attribution method by progressively removing input features in decreasing order of their assigned importance scores and measuring the resulting drop in model confidence. The area under the resulting confidence curve is then computed — a lower value indicates a more faithful attribution map, as removing the most important features should cause a sharp and sustained drop in confidence. This

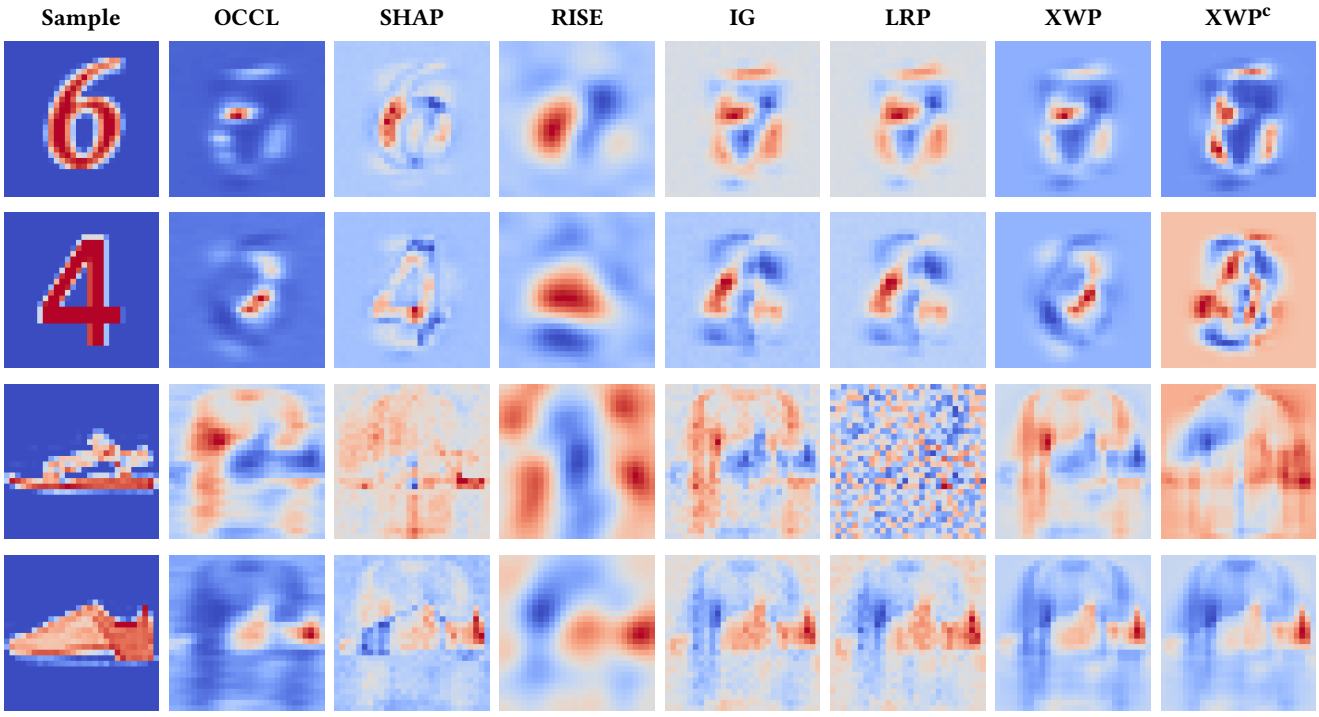

Figure 3: The importance scores of different attribution methods for the Typeface MNIST and Fashion MNIST datasets. A visual inspection suggests that XWP^c is slightly more capable than XWP in detecting patterns within the input signal, while they jointly produce the cleanest heatmaps with respect to their competitors.

metric directly tests whether the features identified as important are indeed the ones the model relies upon.

Average Drop. The Average Drop metric [8] measures the mean percentage decrease in model confidence when a percentage of the most important image regions highlighted by the attribution map are dropped. Formally, it computes the relative drop between the model's confidence on the original image and its confidence on the masked image, averaged across the dataset. A higher average drop indicates a more accurate attribution method, as the highlighted regions should hide crucial information for the model's prediction. For our experimental setup, we selected the value of 20% as the masking percentage.

4.3 Results

This section presents the results of the experiments described in Section 4. As shown in Table 1, Shapley values achieve the highest performance in TMNIST, while Integrated Gradients excel in FashionMNIST. XWP demonstrates strong performance in Deletion AUC (Figure 4), while XWP^c excels in Average Drop.

A vizualisation of the generated heatmaps (Figure 3) further clarifies the nature of the computed attribution scores. While the results generally hold, we observe cases where the individual effect of feature change does not match the signal itself, exhibiting in many cases strong similarities to Occlusion. This translates to a divergence of the objectives of XWP—targeting the model's update needs for better generalization—with the model's current dynamics in response to the input signal. This gap is filled by XWP^c, where the absent effect of all other changes increases the distance measure applied, thus premitting more changes to take effect when considering the input signal. Additional results from applying our variants to more samples from the Typeface MNIST and Fashion MNIST datasets are shown in Figures 5 and 6, respectively.

Although the selected datasets are relatively simple—thereby mitigating some of the theoretical challenges associated with baseline selection in occlusion methods—our proposed variants demonstrate greater stability across evaluations. They consistently achieve top-tier performance, while competing methods exhibit significant fluctuations As demonstrated by the heatmap visualizations, our variants strike an optimal balance between cleanliness and pattern recognition, while also effectively capturing negative patterns arising from polysemantic neurons in a clear and interpretable manner. The complementary strengths of the two methods enhance their collective ability to explain a model's decision-making process comprehensively.

In summary, these experiments confirm that the framework of weight perturbation—embodied by XWP and XWP^c—demonstrates greater robustness in model interpretability, establishing it as a reliable approach for the field of Explainable AI.

5 Conclusion

In this paper, we introduced a novel framework for model explainability through a structured approach exploiting pattern formation

Table 1: Quantitative analysis, based on Average Drop (AD) and Deletion AUC (AUC) for the datasets TMNIST and FashionMNIST (FTMNIST). Shapley values and Integrated Gradients outperform all other techniques by a small margin, yet XWP and XWP^c closely match their performance, achieving greater stability. Best results are highlighted in bold.

	OCCL		SHAP		RISE		IG		LRP		XWP		XWP^c	
Dataset	AD(↑)	AUC(↓)	AD(↑)	AUC(↓)	AD(↑)	AUC(↓)	AD(↑)	AUC(↓)	AD(↑)	AUC(↓)	AD(↑)	AUC(↓)	AD(↑)	AUC(↓)
TMNIST	0.327	0.126	**0.357**	**0.095**	0.33	0.17	0.35	0.157	0.285	0.203	0.308	0.138	0.352	0.162
FMNIST	0.636	0.489	0.633	0.491	0.622	0.564	**0.661**	**0.461**	0.645	0.551	0.647	0.473	0.655	0.524

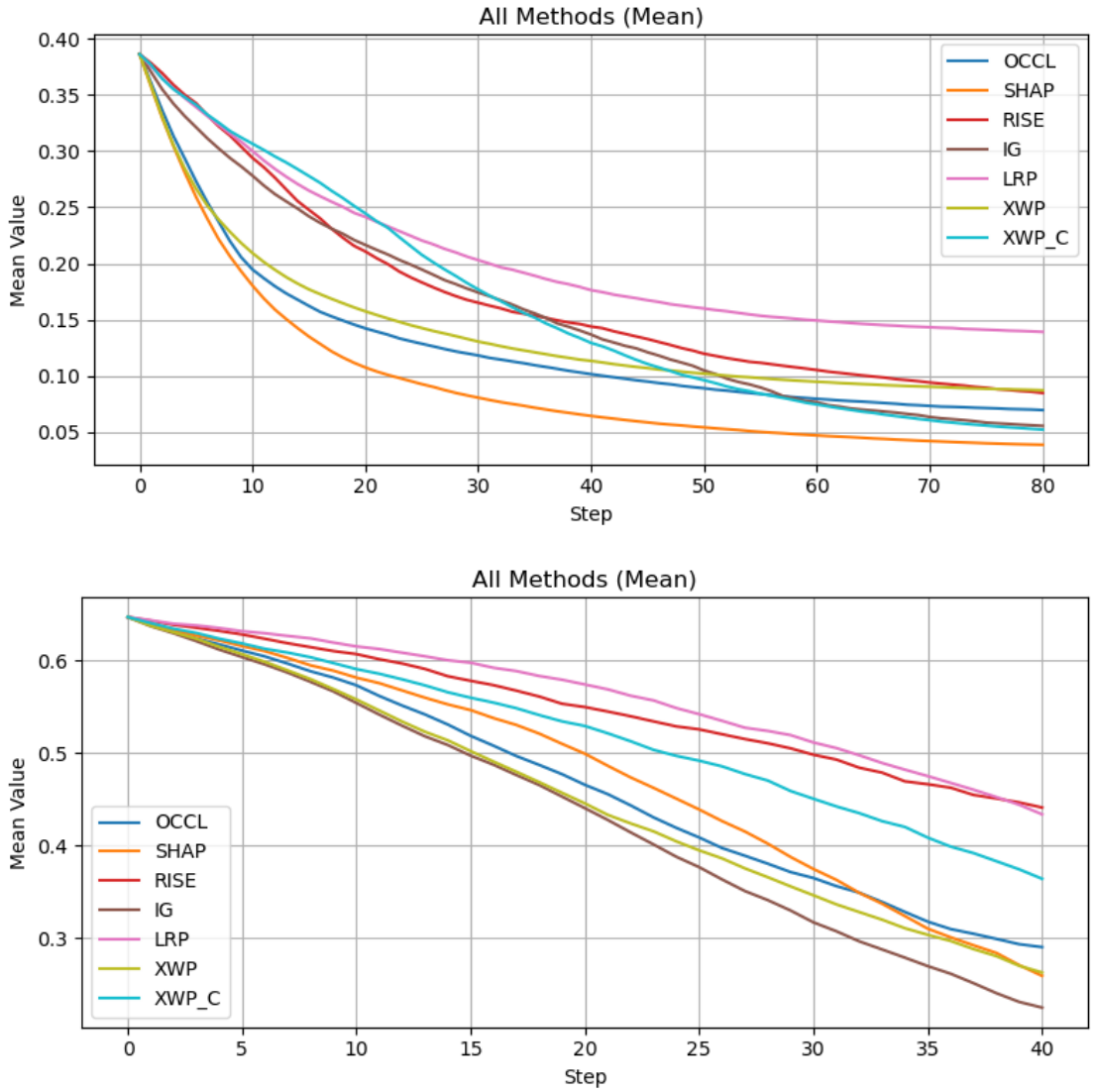

Figure 4: The evolution of Deletion AUC for different attribution methods. Shapley values and Integrated Gradients experience the sharpest drop in AUC score in TMNIST and FashionMNIST respectively, with XWP^c (denoted as XWP_C) and Occlusion to follow. The number of steps is chosen at the point just before the curves reach a plateau.

and weight perturbation. Its design stems from the existence of interpretable patterns within the model's weights, paving the way for the exploitation of spatial resolution. The identification of neuronal

interactions where this resolution was preserved, was then combined with weight perturbation and occlusion. By leveraging the model's states, this design manages to effectively circumvent persistent challenges in occlusion methods—such as out-of-distribution data and added bias. Overall, it comprises a structured framework for model interpretability.

We introduced two variants in this framework and evaluated their performance against state-of-the-art attribution methods. Our approach demonstrated greater preformance when considering both datasets, outperforming classical occlusion techniques while producing cleaner and more interpretable visualizations. These come with reduced noise and fewer artifacts, yet no exceptions in pattern recognition.

This work represents a critical first step toward addressing the theoretical challenges of Occlusion. The simplicity of the selected datasets in our experimental setup mitigates the problem of baseline selection. The method's true potential lies in its extension to more complex datasets and architectures, such as Transformers or Convolutional Neural Networks, and its integration with causal analysis—paving the way for more robust and interpretable AI systems.

References

[1] Shreyash Arya, Sukrut Rao, Moritz Böhle, and Bernt Schiele. 2024. B-cosification: Transforming Deep Neural Networks to be Inherently Interpretable. In *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Eds.), Vol. 37. Curran Associates, Inc., 62756–62786. doi:10.52202/079017-2007

[2] Beyza Nur Aydoğan and Tevfik Aytekin. 2025. An in-depth analysis of KernelSHAP and SamplingSHAP: assessing robustness, error, and efficiency. *Knowledge and Information Systems* 67 (2025), 10545 – 10579. https://api.semanticscholar.org/CorpusID:282832460

[3] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115. doi:10.1016/j.inffus.2019.12.012

[4] Leonard Bereska and Efstratios Gavves. 2024. Mechanistic Interpretability for AI Safety – A Review. arXiv:2404.14082 [cs.AI] https://arxiv.org/abs/2404.14082

[5] Alexander Binder, Sebastian Bach, Gregoire Montavon, Klaus-Robert Muller, and Wojciech Samek. 2016. Layer-Wise Relevance Propagation for Deep Neural Network Architectures. In *Information Science and Applications (ICISA) 2016*, Kuinam J. Kim and Nikolai Joukov (Eds.). Springer Singapore, Singapore, 913–922.

[6] Alexander Binder, Grégoire Montavon, Sebastian Bach, Klaus-Robert Müller, and Wojciech Samek. 2016. Layer-wise Relevance Propagation for Neural Networks with Local Renormalization Layers. arXiv:1604.00825 [cs.CV] https://arxiv.org/abs/1604.00825

[7] Ho Chan and Eduardo Veas. 2024. Importance Estimate of Features via analysis of their Weight and Gradient profile. (04 2024). doi:10.21203/rs.3.rs-4217886/v1

[8] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. 2018. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 839–847. doi:10.1109/WACV.2018.00097

[9] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2021. A Mathematical Framework for Transformer Circuits. *Transformer Circuits Thread* (2021). https://transformer-circuits.pub/2021/framework/index.html.

[10] Thomas Fel, Melanie Ducoffe, David Vigouroux, Remi Cadene, Mikael Capelle, Claire Nicodeme, and Thomas Serre. 2023. Don't Lie to Me! Robust and Efficient Explainability with Verified Perturbation Analysis. arXiv:2202.07728 [cs.CV] https://arxiv.org/abs/2202.07728

[11] Ruth C. Fong and Andrea Vedaldi. 2017. Interpretable Explanations of Black Boxes by Meaningful Perturbation. In *2017 IEEE International Conference on Computer Vision (ICCV)*. Association for Computing Machinery, 3449–3457. doi:10.1109/ICCV.2017.371

[12] Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. 2025. Causal Abstraction: A Theoretical Foundation for Mechanistic Interpretability. arXiv:2301.04709 [cs.AI] https://arxiv.org/abs/2301.04709

[13] Tristan Gomez, Thomas Fréour, and Harold Mouchère. 2022. Metrics for saliency map evaluation of deep learning explanation methods. arXiv:2201.13291 [cs.CV] https://arxiv.org/abs/2201.13291

[14] Cheng-Yu Hsieh, Chih-Kuan Yeh, Xuanqing Liu, Pradeep Ravikumar, Seungyeon Kim, Sanjiv Kumar, and Cho-Jui Hsieh. 2021. Evaluations and Methods for Explanation through Robustness Analysis. arXiv:2006.00442 [cs.LG] https://arxiv.org/abs/2006.00442

[15] Jean Kaddour, Aengus Lynch, Qi Liu, Matt J. Kusner, and Ricardo Silva. 2022. Causal Machine Learning: A Survey and Open Problems. arXiv:2206.15475 [cs.LG] https://arxiv.org/abs/2206.15475

[16] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris B. Kotsiantis. 2020. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy* 23 (2020), 45. https://api.semanticscholar.org/CorpusID:229722844

[17] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 4768–4777.

[18] Mariusz Karol Nowak and Kamil Lelowicz. 2021. Weight Perturbation as a Method for Improving Performance of Deep Neural Networks. In *2021 25th International Conference on Methods and Models in Automation and Robotics (MMAR)*. 127–132. doi:10.1109/MMAR49549.2021.9528460

[19] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. Zoom In: An Introduction to Circuits. *Distill* (2020). doi:10.23915/distill.00024.001 https://distill.pub/2020/circuits/zoom-in.

[20] Vitali Petsiuk, Abir Das, and Kate Saenko. 2018. RISE: Randomized Input Sampling for Explanation of Black-box Models. arXiv:1806.07421 [cs.CV] https://arxiv.org/abs/1806.07421

[21] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1135–1144. doi:10.1145/2939672.2939778

[22] Manar D. Samad, Rahim Hossain, and Khan M. Iftekharuddin. 2021. Dynamic Perturbation of Weights for Improved Data Reconstruction in Unsupervised Learning. In *2021 International Joint Conference on Neural Networks (IJCNN)*. 1–7. doi:10.1109/IJCNN52387.2021.9533539

[23] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70* (Sydney, NSW, Australia) *(ICML'17)*. JMLR.org, 3145–3153.

[24] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *CoRR* abs/1312.6034 (2013). https://api.semanticscholar.org/CorpusID:1450294

[25] Pascal Sturmfels, Scott Lundberg, and Su-In Lee. 2020. Visualizing the Impact of Feature Attribution Baselines. *Distill* (2020). doi:10.23915/distill.00022 https://distill.pub/2020/attribution-baselines.

[26] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70* (Sydney, NSW, Australia) *(ICML'17)*. JMLR.org, 3319–3328.

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 6000–6010.

[28] Saurabh Vyawahare. 2024. TMNIST (Typeface MNIST). Kaggle. https://www.kaggle.com/datasets/saurabhvyawahare/tmnist-typeface-mnist

[29] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. 2020. Adversarial Weight Perturbation Helps Robust Generalization. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 2958–2969. https://proceedings.neurips.cc/paper_files/paper/2020/file/1ef91c212e30e14bf125e9374262401f-Paper.pdf

[30] Han Xiao, Kashif Rasul, and Roland Vollgraf. 2017. *Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms*. arXiv:cs.LG/1708.07747 [cs.LG]

[31] Matthew D. Zeiler and Rob Fergus. 2014. Visualizing and Understanding Convolutional Networks. In *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer International Publishing, Cham, 818–833.

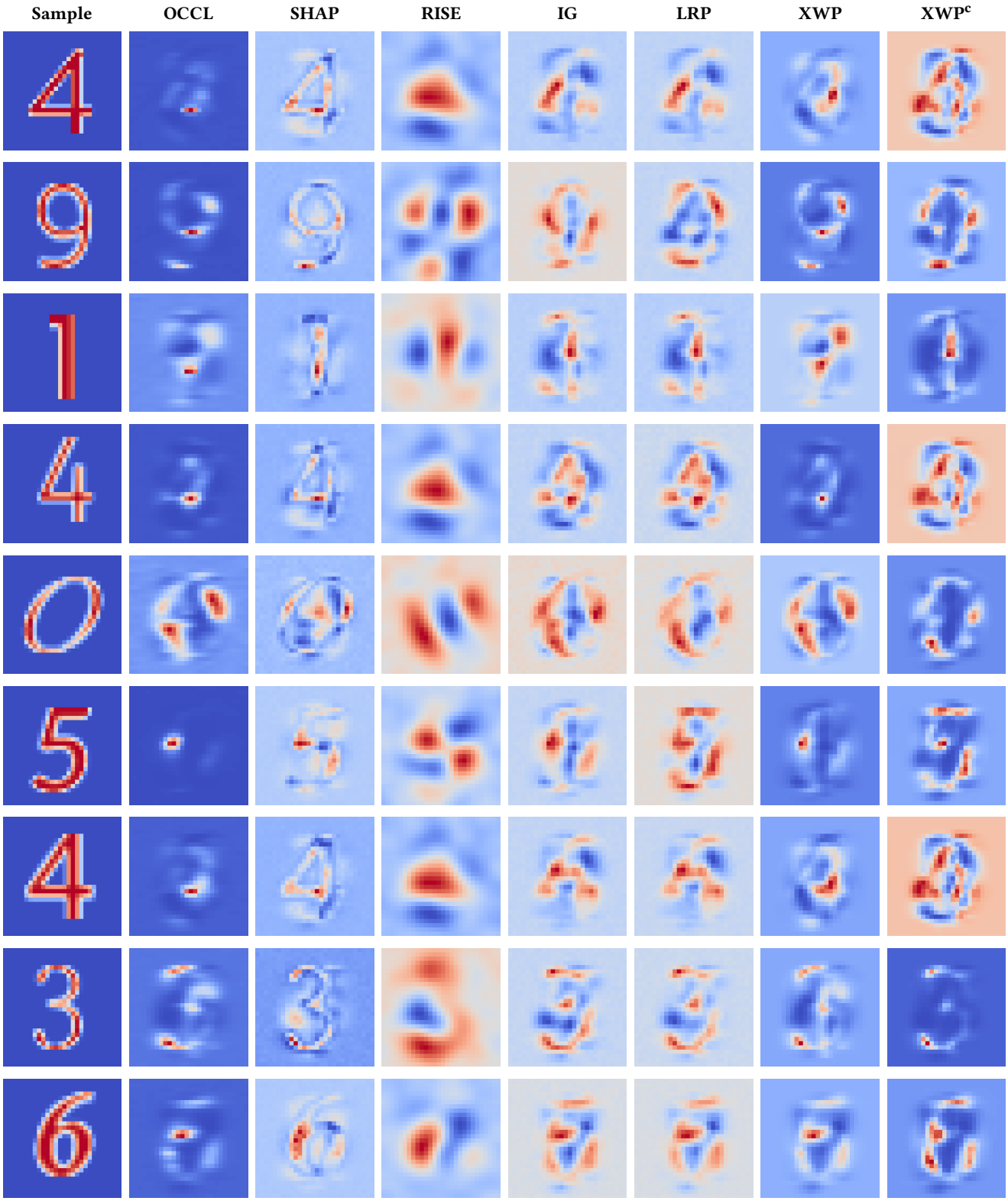

Figure 5: The importance scores of different attribution methods for the Typeface MNIST datasets. XWP^c recognizes the input signal with greater confidence XWP, while the latter produces more focused heatmaps across different digits.

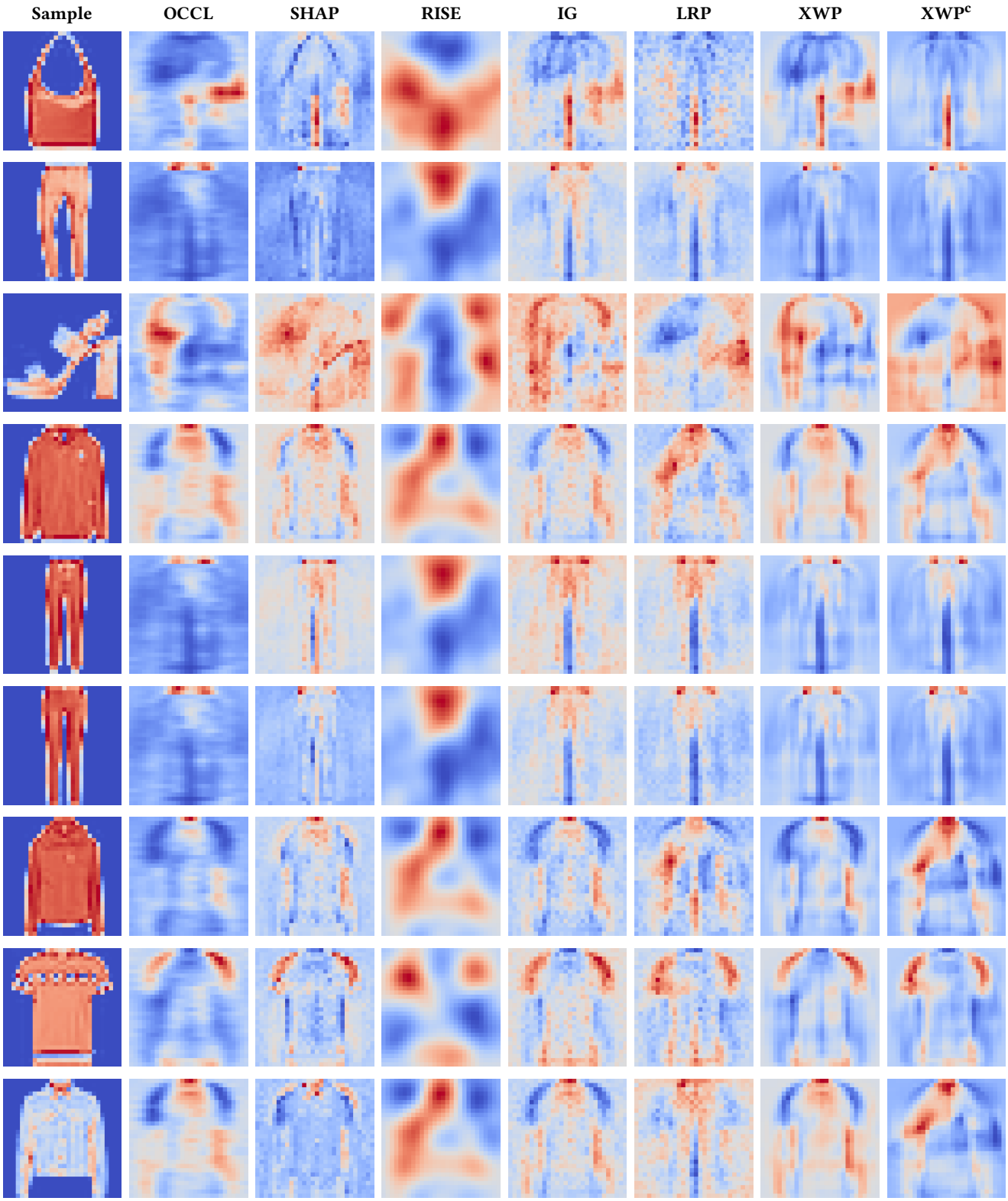

Figure 6: The importance scores of different attribution methods for the Fashion MNIST datasets. Both XWP and XWP^c align with other pattern recognition techniques, yet they produce heatmaps that are significantly cleaner and more interpretable.