

Explaining Black-box Model Predictions via Two-level Nested Feature Attributions with Consistency Property

Yuya Yoshikawa^{1*}, Masanari Kimura², Ryotaro Shimizu³ and Yuki Saito³

¹STAIR Lab, Chiba Institute of Technology

²School of Mathematics and Statistics, The University of Melbourne

³ZOZO Research

yoshikawa@stair.center, m.kimura@unimelb.edu.au, ryotaro.shimizu@zozo.com, yuki.saito@zozo.com

Abstract

Techniques that explain the predictions of black-box machine learning models are crucial to make the models transparent, thereby increasing trust in AI systems. The input features to the models often have a nested structure that consists of high- and low-level features, and each high-level feature is decomposed into multiple low-level features. For such inputs, both high-level feature attributions (HiFAs) and low-level feature attributions (LoFAs) are important for better understanding the model’s decision. In this paper, we propose a model-agnostic local explanation method that effectively exploits the nested structure of the input to estimate the two-level feature attributions simultaneously. A key idea of the proposed method is to introduce the consistency property that should exist between the HiFAs and LoFAs, thereby bridging the separate optimization problems for estimating them. Thanks to this consistency property, the proposed method can produce HiFAs and LoFAs that are both faithful to the black-box models and consistent with each other, using a smaller number of queries to the models. In experiments on image classification in multiple instance learning and text classification using language models, we demonstrate that the HiFAs and LoFAs estimated by the proposed method are accurate, faithful to the behaviors of the black-box models, and provide consistent explanations.

1 Introduction

The rapid increase in size and complexity of machine learning (ML) models has led to a growing concern about their *black-box* nature. Models provided as cloud services are literal black boxes, as users have no access to the models themselves or the training data used. This opacity raises numerous concerns, including issues of trust, accountability, and transparency. Consequently, techniques for explaining the predictions made by such black-box models have

*The extended version of this paper, including detailed proofs and additional experimental results, is available at: <https://arxiv.org/abs/2405.14522>.

been attracting significant attention [Danilevsky *et al.*, 2020; Došilović *et al.*, 2018; Saeed and Omlin, 2023].

Various *model-agnostic* local explanation methods have been proposed to explain the predictions of black-box models. Representative methods include, for example, local interpretable model-agnostic explanation (LIME) [Ribeiro *et al.*, 2016] and kernel Shapley additive explanations (Kernel SHAP) [Lundberg and Lee, 2017], which estimate the feature attributions for the individual prediction by approximating the model’s behavior with local linear surrogate models around the input.

In LIME and Kernel SHAP, the input to the model is generally assumed to be a flat structure, where the input features are treated as independent variables. In many realistic tasks for various domains, such as image, text, geographic, e-commerce, and social network data, however, the input features have a nested structure that consists of high- and low-level features, and each high-level feature is decomposed into multiple low-level features. A typical task with such nested features is multiple instance learning (MIL) [Ilse *et al.*, 2018] where the model is formulated as a set function [Kimura *et al.*, 2024]. In MIL, the input is a set of instances, the high-level feature is an instance in the set, and the low-level features represent the features of the instance. In addition, even if the input is not represented in a nested structure when it is fed into the model, it may be more natural to interpret it with the nested structure. For example, although a text input is usually represented as a sequence of words, it is natural to interpret it as having high-level features such as phrases, sentences, and paragraphs.

The two-level features enable us to understand the model predictions with two types of feature attributions at different levels of granularity, which we name *high-level feature attributions (HiFAs)* and *low-level feature attributions (LoFAs)*, respectively. Figure 1 shows an example of the prediction for a nested-structured input and its corresponding HiFAs and LoFAs. The HiFAs represent how much each of the high-level features contributes to the prediction. These are also referred to as *instance attributions* in the MIL literature [Early *et al.*, 2022; Javed *et al.*, 2022], which are used to reveal which instances strongly affected the model’s decision. On the other hand, the LoFAs represent how much each of the low-level features contributes to the prediction, providing a finer-grained explanation of how the components of the

instances affected the prediction. Both the HiFAs and LoFAs are important for understanding the model’s decision. *However, existing studies have focused on estimating either-level attributions, and no study has addressed estimating the HiFAs and LoFAs simultaneously.*

For the estimation of the HiFAs and LoFAs, two naive approaches can be applied. One is to estimate the HiFAs and LoFAs separately by applying existing model-agnostic local explanation methods to the high- and low-level features, respectively. The other is to estimate the LoFAs first, as in the former approach, and then estimate the HiFAs by aggregating the LoFAs. However, these approaches have two limitations in terms of leveraging the nested structure of the input. First, even though the queries to the black-box model are often limited in practice due to the computational time and request costs, the input structure is not utilized to reduce the number of queries in the estimation. Second, the former approach can produce mutually inconsistent HiFAs and LoFAs. For example, the high-level feature identified as most influential (highest HiFA) might not coincide with the high-level feature that contains the most influential low-level feature (highest LoFA).

To address these issues, we propose a model-agnostic local explanation method that effectively exploits the nested structure of the input to estimate the HiFAs and LoFAs simultaneously. A key idea of the proposed method is to introduce the consistency property that should exist between the HiFAs and LoFAs, thereby bridging the separate optimization problems for them. We solve a joint optimization problem to estimate the HiFAs and LoFAs simultaneously with the consistency constraints depicted in Figure 1 based on the alternating direction method of multipliers (ADMM) [Boyd *et al.*, 2011]. The algorithm is a general framework that can also introduce various types of regularizations and constraints for the HiFAs and LoFAs, such as the ℓ_1 and ℓ_2 regularizations and non-negative constraints, which lead to the ease of interpretability for humans.

In experiments, we quantitatively and qualitatively assess the HiFAs and LoFAs estimated by the proposed method on image classification in the MIL setting and text classification using language models. The experimental results show that the HiFAs and LoFAs estimated by the proposed method 1) satisfy the consistency property, 2) are faithful explanations to the black-box models even when the number of queries to the model is small, 3) can accurately guess the ground-truth positive instances and their features in the MIL task, and 4) are reasonable explanations visually.

The contributions of this work are summarized as follows:

- This study is the first to propose a model-agnostic local explanation method to estimate the two-level nested feature attributions simultaneously, which satisfies the consistency property between them.
- To justify the behavior of the proposed method, we provide theoretical analysis in terms of the existence and uniqueness of the solution and the convergence of the optimization algorithm.
- In the experiments on practical tasks, we demonstrate that the proposed method can produce accurate, faithful,

and consistent two-level feature attributions with a smaller number of queries to the black-box models.

2 Related Work

Numerous methods for explaining the individual predictions of black-box models have been proposed in the literature [Ribeiro *et al.*, 2016; Lundberg and Lee, 2017; Ribeiro *et al.*, 2018; Petsiuk *et al.*, 2018; Plumb *et al.*, 2018]. A versatile approach is to explain feature attributions estimated by approximating the model predictions with surrogate models around the input, such as LIME [Ribeiro *et al.*, 2016] and Kernel SHAP [Lundberg and Lee, 2017]. The proposed method is in line with this type of approach.

One example of a nested input structure is set data, where a set of multiple instances is treated as a single input. Set data appears in various ML applications, such as point cloud classification [Guo *et al.*, 2021], medical image analysis [Cheplygina *et al.*, 2019], and group recommendation [Dara *et al.*, 2020], and the explainability in those applications has also been studied in the literature [Tan and Kotthaus, 2022; van der Velden *et al.*, 2022]. Unlike our work, most such studies focus only on estimating instance attributions corresponding to those of high-level features. For example, Early *et al.* proposed to estimate instance attributions by learning surrogate models with MIL-suitable kernel functions [Early *et al.*, 2022].

Several studies have addressed estimating feature attributions effectively by leveraging group information of input features. In the natural language processing literature, some studies estimated sentence- and phrase-level feature attributions by grouping words in the same sentence and phrase together and regarding them as a single feature [Zafar *et al.*, 2021; Mosca *et al.*, 2022]. In addition, Rychener *et al.* showed that word-level feature attributions can be improved by generating perturbations at a sentence level, mitigating out-of-distribution issues for the model and the challenges of a high-dimensional search space [Rychener *et al.*, 2023]. In the official SHAP library [shap (GitHub), 2024], by grouping input features by hierarchical clustering in advance and generating perturbations at the group level, one can reduce the number of queries to the model.

3 Proposed Method

3.1 Two-level Nested Feature Attributions with Surrogate Models

The model f to be explained is a trained black-box model that takes an arbitrary input, such as tabular, image or text, $\mathbf{x} \in \mathcal{X}$, and outputs a prediction $\mathbf{y} = f(\mathbf{x}) \in [0, 1]^C$ where \mathcal{X} is the input space, and C is the number of classes. The input \mathbf{x} consists of two-level nested features, referred to as high-level and low-level features, with each high-level feature being decomposed into multiple low-level features. In particular, the input \mathbf{x} is represented as a set or sequence of J high-level features, i.e., $\mathbf{x} = \{\mathbf{x}_j\}_{j=1}^J$ where $\mathbf{x}_j \in \mathbb{R}^{D_j}$ is a D_j -dimensional low-level feature vector representing the j -th high-level feature. One example of such input appears in image classification under the MIL setting. In this setting, the

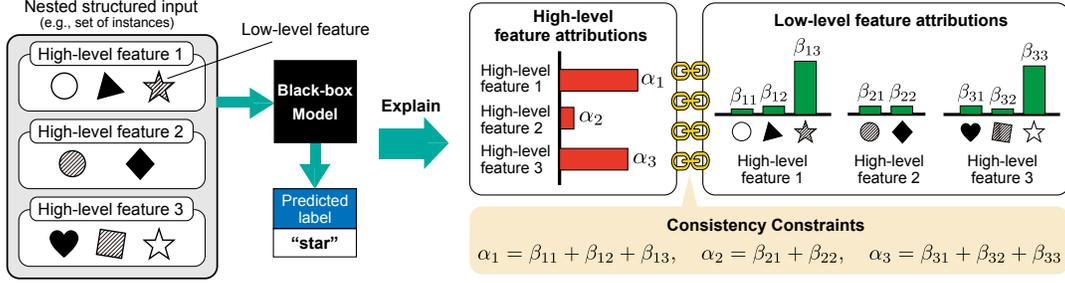


Figure 1: Example of the black-box model prediction for a nested structured input and its corresponding high- and low-level feature attributions estimated by the proposed method with consistency constraints. Objects in each high-level feature represent the low-level features.

input is a bag of images, the high-level feature is an image in the bag, and the low-level features correspond to super-pixels in the image. Another example appears in a document classification where the input is a sequence of sentences, the high-level feature is a sentence in the sequence, and the low-level features correspond to the words in the sentence.

We consider estimating the high-level feature attributions (HiFAs) and low-level feature attributions (LoFAs) that explain the prediction of the black-box model f for the input \mathbf{x} using surrogate models as with LIME and Kernel SHAP. The HiFAs and LoFAs represent how much the high- and low-level features in the input contribute to the prediction, respectively. In the aforementioned MIL setting, the HiFAs represent how much images in the input bag contribute to the prediction, which is also referred to as *instance attributions* in the literature, and the LoFAs represent how much the super-pixels in the images contribute to the prediction. To estimate the HiFAs and LoFAs, we introduce two-level local linear surrogate models for high-level and low-level features, e^H and e^L , that mimic the behaviors of the black-box model f around the input \mathbf{x} , as follows:

$$e^H(\mathbf{z}^H; \boldsymbol{\alpha}) = \sum_{j=1}^J \alpha_j z_j^H, \quad e^L(\mathbf{z}^L; \boldsymbol{\beta}) = \sum_{j=1}^J \sum_{d=1}^{D_j} \beta_{jd} z_{jd}^L, \quad (1)$$

where $\mathbf{z}^H \in \{0, 1\}^J$ and $\mathbf{z}^L = \{z_j^L\}_{j=1}^J$ with $z_j^L \in \{0, 1\}^{D_j}$ are simplified inputs associated with the input \mathbf{x} , which are used to indicate the presence or absence of the high- and low-level features in \mathbf{x} , respectively; $\boldsymbol{\alpha} \in \mathbb{R}^J$ and $\boldsymbol{\beta} = \{\beta_j\}_{j=1}^J$ with $\beta_j \in \mathbb{R}^{D_j}$ are the learnable coefficients of these surrogate models, and after learning, they will be the HiFAs and LoFAs themselves, respectively. For ease of computation, we define the concatenation of $\boldsymbol{\beta}$ and \mathbf{z}^L over the high-level features as $\boldsymbol{\beta}^\dagger = \text{concat}(\beta_1, \beta_2, \dots, \beta_J) \in \mathbb{R}^{D^\dagger}$ and $\mathbf{z}^{L\dagger} = \text{concat}(z_1^L, z_2^L, \dots, z_J^L) \in \{0, 1\}^{D^\dagger}$, where $D^\dagger = \sum_{j=1}^J D_j$.

The surrogate models are learned with the predictions of the black-box model f for perturbations around the input \mathbf{x} . The perturbations are generated by sampling the simplified inputs \mathbf{z}^H and $\mathbf{z}^{L\dagger}$ from binary uniform distributions and then constructing masked inputs $\phi_{\mathbf{x}}^H(\mathbf{z}^H), \phi_{\mathbf{x}}^L(\mathbf{z}^{L\dagger}) \in \mathcal{X}$ depending on the simplified inputs, respectively. Here, $\phi_{\mathbf{x}}^H$ and $\phi_{\mathbf{x}}^L$ are mask functions that replace the input \mathbf{x} 's di-

mensions associated with the dimensions being zero in the simplified inputs \mathbf{z}^H and $\mathbf{z}^{L\dagger}$ with uninformative values, such as zero, respectively. Let $\mathbf{Z}^H \in \{0, 1\}^{N_H \times J}$ and $\mathbf{Z}^L \in \{0, 1\}^{N_L \times D^\dagger}$ be the matrices whose rows are the generated simplified inputs for the high- and low-level features, respectively, where N_H and N_L are the numbers of perturbations used to estimate the HiFAs and LoFAs, respectively. Also, let $\tilde{\mathbf{y}}^H = [\tilde{y}_1^H, \tilde{y}_2^H, \dots, \tilde{y}_{N_H}^H]^\top \in \mathbb{R}^{N_H \times C}$ and $\tilde{\mathbf{y}}^L = [\tilde{y}_1^L, \tilde{y}_2^L, \dots, \tilde{y}_{N_L}^L]^\top \in \mathbb{R}^{N_L \times C}$ be the predictions of the black-box model for the perturbations where $\tilde{y}_n^H = f(\phi_{\mathbf{x}}^H(\mathbf{Z}_n^H))$ and $\tilde{y}_n^L = f(\phi_{\mathbf{x}}^L(\mathbf{Z}_n^L))$.

The parameters of the surrogate models, i.e., the HiFAs $\hat{\boldsymbol{\alpha}}$ and LoFAs $\hat{\boldsymbol{\beta}}^\dagger$, can be estimated by solving the following weighted least squares separately:

$$\hat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}}{\text{argmin}} \mathcal{L}_H(\boldsymbol{\alpha}) + \lambda_H \Omega_H(\boldsymbol{\alpha}) \quad (2)$$

$$\text{where } \mathcal{L}_H(\boldsymbol{\alpha}) = \frac{1}{2} (\tilde{\mathbf{y}}^H - \mathbf{Z}^H \boldsymbol{\alpha})^\top \mathbf{W}^H (\tilde{\mathbf{y}}^H - \mathbf{Z}^H \boldsymbol{\alpha}),$$

$$\hat{\boldsymbol{\beta}}^\dagger = \underset{\boldsymbol{\beta}^\dagger}{\text{argmin}} \mathcal{L}_L(\boldsymbol{\beta}^\dagger) + \lambda_L \Omega_L(\boldsymbol{\beta}^\dagger) \quad (3)$$

$$\text{where } \mathcal{L}_L(\boldsymbol{\beta}^\dagger) = \frac{1}{2} (\tilde{\mathbf{y}}^L - \mathbf{Z}^L \boldsymbol{\beta}^\dagger)^\top \mathbf{W}^L (\tilde{\mathbf{y}}^L - \mathbf{Z}^L \boldsymbol{\beta}^\dagger),$$

where $\mathbf{W}^H \in \mathbb{R}^{N_H \times N_H}$ and $\mathbf{W}^L \in \mathbb{R}^{N_L \times N_L}$ are the diagonal matrices whose n th diagonal elements represent the sample weights for the n th perturbation; Ω_H and Ω_L are the regularizers for the HiFAs and LoFAs, respectively; $\lambda_H \geq 0$ and $\lambda_L \geq 0$ are the regularization strengths.

3.2 Optimization with Consistency Constraints

Although the HiFAs and LoFAs provide different levels of explanations, these explanations for the same black-box model should be consistent with each other. Owing to the linearity of the surrogate models and the fact that each high-level feature can be decomposed into low-level features, the following property is expected to hold:

Property 1 (Consistency between two-level feature attributions).

$$\alpha_j = \sum_{d=1}^{D_j} \beta_{jd} \quad (\forall j \in [J]). \quad (4)$$

In other words, the HiFA for each high-level feature is equal to the sum of the LoFAs of its constituent low-level features.

The two surrogate models defined in (1) that satisfy the consistency property behave equivalently for simplified inputs \mathbf{z}^H and \mathbf{z}^L under the condition if $z_j^H = 0$ then $z_j^L = \mathbf{0}_{D_j}$, and if $z_j^H = 1$ then $z_j^L = \mathbf{1}_{D_j}$ where $\mathbf{0}_{D_j}$ and $\mathbf{1}_{D_j}$ are the D_j -dimensional zero and one vectors, respectively.

The consistency property is essential for providing consistent and convincing explanations to humans. However, in practice, it is often not satisfied for two reasons. First, the number of perturbations may be insufficient to accurately estimate the feature attributions because the number of queries to the model f is often limited due to computational time and request costs. Second, in the predictions for the perturbations, the model f 's behavior can differ depending on whether the high-level features or the low-level features are masked out because of *missingness bias* [Jain *et al.*, 2022]. To overcome these problems, the proposed method estimates the HiFAs and LoFAs simultaneously by solving the following optimization with consistency constraints:

$$\begin{aligned} \hat{\alpha}, \hat{\beta}^\dagger &= \underset{\alpha, \beta^\dagger}{\operatorname{argmin}} \mathcal{L}_H(\alpha) + \mathcal{L}_L(\beta^\dagger) + \lambda_H \Omega_H(\alpha) + \lambda_L \Omega_L(\beta^\dagger) \\ \text{s.t. } \alpha_j &= \sum_{d=1}^{D_j} \beta_{j,d} \quad (\forall j \in [J]). \end{aligned} \quad (5)$$

The consistency constraints bridge the gap between the two surrogate models, forcing them to behave equivalently. This helps compensate the insufficiency of the queries to the model and mitigate the negative effects of the missingness bias on the estimation of the HiFAs and LoFAs.

We solve the optimization problem based on the alternating direction method of multipliers (ADMM) [Boyd *et al.*, 2011]. The detailed derivation of the optimization algorithm is provided in Appendix A. An advantage of employing the ADMM is that, despite the interdependence of α and β^\dagger caused by the consistency constraints, they can be estimated independently as in (2) and (3). In addition, the solution has another advantage: it allows the inclusion of various types of regularizations and constraints for α and β , such as sparsity regularization and non-negative constraints in Ω_H and Ω_L . The optimization algorithm is provided in Algorithm 1 in Appendix A.

4 Auxiliary Theoretical Analysis

In this section, we provide theoretical analysis for the proposed method, referred to as *Consistent Two-level Feature Attribution (C2FA)*, to justify its behavior. See Appendix C for proof sketches of all statements in this section. In this section, we impose the following assumptions.

Assumption 4.1 (Convexity). $\mathcal{L}_H(\cdot)$, $\mathcal{L}_L(\cdot)$, $\Omega_H(\cdot)$, $\Omega_L(\cdot)$ are convex in their respective parameters. Specifically, each loss term is weighted least squares loss, and each regularizer is a convex function.

Assumption 4.2 (Strict Feasibility). There exists $(\alpha^\circ, \beta^\circ)$ such that $\alpha_j^\circ = \sum_{d=1}^{D_j} \beta_{j,d}^\circ$ for all j , ensuring the constraint set is non-empty.

Assumption 4.3 (Non-degeneracy). The design matrices underlying \mathbf{z}^H and \mathbf{z}^L have full column rank in typical regression settings. This ensures identifiability of α and β .

Assumption 4.4 (Bounded Outputs). Black-box outputs $\{f(x_i)\}_{i=1}^N$ lie in a bounded set, or at least have finite second moments, ensuring the objective is well-defined and finite.

Assumption 4.5 (Parameter Domains). α and β lie in closed, convex domains Γ_α and Γ_β .

Define the feasible set

$$\Gamma^\circ := \left\{ (\alpha, \beta) \mid \alpha_j = \sum_{d=1}^{D_j} \beta_{j,d}, \forall j, \alpha \in \Gamma_\alpha, \beta \in \Gamma_\beta \right\},$$

and define $\mathcal{J}(\alpha, \beta) := \mathcal{L}_H(\alpha) + \mathcal{L}_L(\beta) + \lambda_H \Omega_H(\alpha) + \lambda_L \Omega_L(\beta)$. Hence the optimization problem is $\min_{(\alpha, \beta) \in \Gamma^\circ} \mathcal{J}(\alpha, \beta)$. We develop existence and uniqueness of solutions under mild conditions, as well as the precise satisfaction of $\alpha_j = \sum_{d=1}^{D_j} \beta_{j,d}$ at optimality. We first establish that a global minimizer exists. Under additional strict convexity conditions, we also show uniqueness

Lemma 4.6 (Existence of a Global Minimzer). *Under assumptions, there exists at least one global minimizer (α^*, β^*) of C2FA optimization.*

Lemma 4.7 (Strong Convexity and Uniqueness). *Suppose that the high-level and low-level losses \mathcal{L}_H and \mathcal{L}_L are strictly convex quadratic forms in α and β , respectively (e.g., weighted least-squares with full column rank), and Ω_H, Ω_L are either strictly convex or affine so as not to destroy strict convexity of the overall objective. Then any global minimizer (α^*, β^*) is unique.*

Theorem 4.8. *Let $\tilde{\alpha}$ be an unconstrained minimizer of $\mathcal{L}_H(\alpha) + \lambda_H \Omega_H(\alpha)$, and let $\tilde{\beta}$ be an unconstrained minimizer of $\mathcal{L}_L(\beta) + \lambda_L \Omega_L(\beta)$. Then, if $(\tilde{\alpha}, \tilde{\beta})$ do not already satisfy $\tilde{\alpha}_j = \sum_d \tilde{\beta}_{j,d}$ for each j , any global minimizer (α^*, β^*) of the C2FA optimization achieves strictly smaller objective value: $\mathcal{J}(\alpha^*, \beta^*) < \mathcal{J}(\tilde{\alpha}, \tilde{\beta})$.*

Next, we provide the convergence analysis based on the concentration inequality.

Theorem 4.9 (High-Probability Convergence of C2FA). *Let $\{(\mathbf{z}_n^H, Y_n^H)\}_{n=1}^{N_H}$ and $\{(\mathbf{z}_n^L, Y_n^L)\}_{n=1}^{N_L}$ be i.i.d. samples drawn from some distribution of mask vectors (for high-level and low-level features) and black-box outputs, with sub-Gaussian or bounded noise. Form the matrices $\mathbf{Z}^H \in \mathbb{R}^{N_H \times J}$ and $\mathbf{Z}^L \in \mathbb{R}^{N_L \times D^\dagger}$ by stacking $\mathbf{z}_n^H, \mathbf{z}_n^L$, and similarly stack Y^H, Y^L . Suppose the following:*

- (i) *The random design satisfies a restricted eigenvalue property on \mathbf{Z}^H and \mathbf{Z}^L with high probability.*
- (ii) *The losses \mathcal{L}_H and \mathcal{L}_L are given by weighted least squares, and the regularizers Ω_H, Ω_L are convex.*
- (iii) *$\alpha_j = \sum_{d=1}^{D_j} \beta_{j,d}$ for each $j = 1, \dots, J$.*
- (iv) *Let (α^*, β^*) be the minimizer of the expected version of the same objective under that constraint, i.e. the best consistent linear approximation of f in expectation.*

Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$, the solution $(\widehat{\alpha}, \widehat{\beta})$ to the finite-sample C2FA problem satisfies

$$\|(\widehat{\alpha}, \widehat{\beta}) - (\alpha^*, \beta^*)\|_2 \leq C \sqrt{\frac{(J + D^\dagger) + \log(\frac{1}{\delta})}{N_H + N_L}},$$

where $C > 0$ is a constant depending on sub-Gaussian parameters, the restricted eigenvalue lower bound, and the regularization strengths.

Corollary 4.10 (Uniform Approximation Guarantee). Assume the same setting as Theorem 4.9 and let $e^H(\mathbf{z}^H) = (\mathbf{z}^H)^\top \widehat{\alpha}$ and $e^L(\mathbf{z}^L) = (\mathbf{z}^L)^\top \widehat{\beta}$ be the final C2FA surrogate models. Further suppose that \mathbf{z}^H and \mathbf{z}^L take values in some bounded set $\mathcal{Z}^H \subseteq \mathbb{R}^J$ and $\mathcal{Z}^L \subseteq \mathbb{R}^{D^\dagger}$ respectively. Then with probability at least $1 - \delta$,

$$\begin{aligned} \sup_{\mathbf{z}^H \in \mathcal{Z}^H} \left| f(\phi_{\mathbf{x}}^H(\mathbf{z}^H)) - e^H(\mathbf{z}^H) \right| + \sup_{\mathbf{z}^L \in \mathcal{Z}^L} \left| f(\phi_{\mathbf{x}}^L(\mathbf{z}^L)) - e^L(\mathbf{z}^L) \right| \\ \leq B_H \cdot \tilde{C} \sqrt{\frac{(J + D^\dagger) + \log(\frac{1}{\delta})}{N_H + N_L}} + \epsilon^*, \end{aligned}$$

for some constant $\tilde{C}, B_H, \epsilon^*$ depending on sub-Gaussian parameters, bounding sets $\mathcal{Z}^H, \mathcal{Z}^L$, and the regularization strengths.

5 Experiments

We conducted experiments on two tasks of image and text domains to evaluate the effectiveness of the proposed method, referred to as C2FA, implemented with Algorithm 1 in Appendix A. Its hyperparameters are provided in Appendix B.

Comparing Methods. We used five methods for comparison: LIME [Ribeiro *et al.*, 2016], MILLI [Early *et al.*, 2022], Bottom-Up LIME (BU-LIME), Top-Down LIME (TD-LIME), and Top-Down MILLI (TD-MILLI). With LIME, we estimated the HiFAs and LoFAs separately by solving (2) and (3), respectively, where we used the cosine kernel for the sample weights and ℓ_2 regularization for Ω_H and Ω_L . MILLI is the state-of-the-art instance attribution method in the MIL setting, which was proposed for estimating only the HiFAs. Therefore, we estimated the LoFAs in MILLI in the same way as in LIME. With BU-LIME, we first estimated the LoFAs using LIME and then calculated the HiFAs of each high-level feature by summing the LoFAs associated with the high-level feature. This method always satisfies the consistency property because the HiFAs are calculated from the LoFAs. With TD-LIME and TD-MILLI, we first estimated the HiFAs using LIME and MILLI, respectively. Then, for the j th high-level feature, we determined its LoFAs β_j by sampling from a normal distribution with mean α_j (the j th HiFA) and the standard deviation $1/D_j$. Finally, we enforced the consistency property by adjusting one randomly selected low-level feature for each high-level feature j : we set that feature’s LoFA to $\beta_{jd} = \alpha_j - \sum_{d' \in [D_j] \setminus \{d\}} \beta_{jd'}$, ensuring that $\sum_{d' \in [D_j]} \beta_{jd'} = \alpha_j$.

5.1 Multiple Instance Image Classification

Dataset. We constructed an MIL dataset from the Pascal VOC semantic segmentation dataset [Everingham *et al.*,

2015] with the ground-truth instance- and pixel-level labels. Each sample (bag) contains three to five images, randomly selected from the Pascal VOC dataset. Here, low-level features correspond to regions (super-pixels) of each image. Each bag is labeled positive if at least one image in the bag is associated with the “cat” label; otherwise the bag is negative. Also, each image pixel is labeled positive if it is associated with the “cat” label, and negative otherwise. A detailed description of the dataset is provided in Appendix B.

Black-box Model. We used a DeepSets permutation-invariant model [Zaheer *et al.*, 2017] with ResNet-50 [He *et al.*, 2016] as the black-box model f to be explained. We describe the implementation details of the model in Appendix B. Here, the test accuracy of the model was 0.945.

Quantitative Evaluation. We assessed the estimated HiFAs and LoFAs in terms of correctness, faithfulness, and consistency. Correctness is evaluated using the ground-truth instance- and pixel-level labels. Following the evaluation in the MIL study [Early *et al.*, 2022], we evaluated the estimated HiFAs with normalized discounted cumulative gain (NDCG). Additionally, by regarding the estimated LoFAs as the predictions of the pixel-level labels, we evaluated them with the area under ROC curve (AUROC) in a binary semantic segmentation setting. In the faithfulness evaluation, we assessed whether the estimated HiFAs and LoFAs are faithful to the behaviors of the model f based on insertion and deletion metrics [Petsiuk *et al.*, 2018]. For consistency, we used the following two metrics. The first one is the consistency between the estimated HiFAs and LoFAs, measured by $\|\alpha - M\beta^\dagger\|^2$, which is the same quantity used as the consistency penalty in (7). The second one is the agreement of the most important high- and low-level feature (MIHL). We ran the evaluations three times with different random seeds and reported the average scores and their standard deviations. The details of the evaluation metrics are provided in Appendix B.

Results. Figure 2a shows the NDCG and deletion scores of the estimated HiFAs over various numbers of perturbations for the LoFAs, N_L , where we fixed the number of perturbations for the HiFAs, $N_H = 5$. We found that the proposed method (C2FA) consistently achieved the best NDCG and deletion scores, and the superiority of the proposed method is especially noticeable when N_L is small. Although BU-LIME improved the scores as N_L increased, the scores were still lower than those of the proposed method. Since the other comparing methods estimate the HiFAs without the effects of the LoFAs, their scores were constant regardless of the value of N_L .

Figure 2b shows the AUROC and deletion scores of the estimated LoFAs over various values of N_L where we fixed $N_H = 20$. When N_L is small, we found that the proposed method achieved significantly better AUROC and deletion scores. In particular, the AUROC score of the proposed method at $N_L = 50$ was comparable to that of the second-best methods, LIME, MILLI, and BU-LIME, at $N_L = 150$, and the deletion score of the proposed method at $N_L = 50$ was much the same as that of the second-best methods at $N_L = 100$. These results show that the proposed method is very efficient in terms of the number of queries to the model

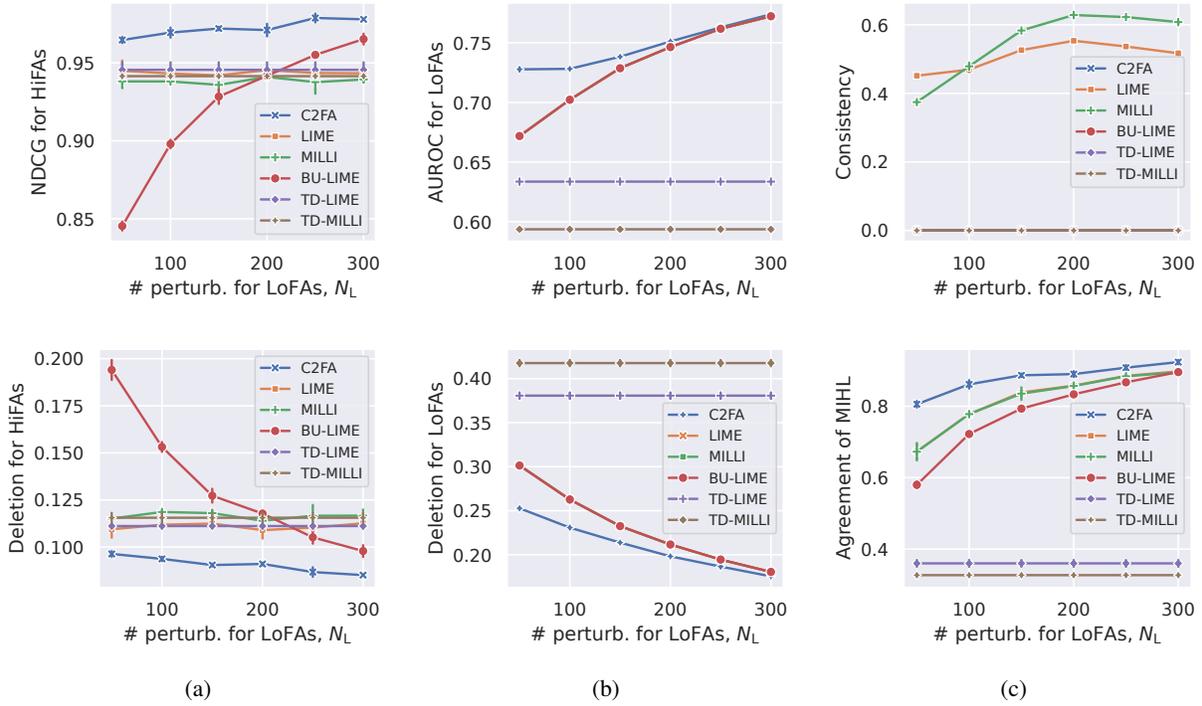


Figure 2: Quantitative evaluation on the image classification task. (a) NDCG (higher is better) and deletion scores (lower is better) of the estimated HiFAs. (b) AUROC (higher is better) and deletion scores (lower is better) of the estimated LoFAs. (c) Consistency scores (lower is better) and the agreement scores of MIHL (higher is better). The error bars represent the standard deviations of the scores over three runs with different random seeds.

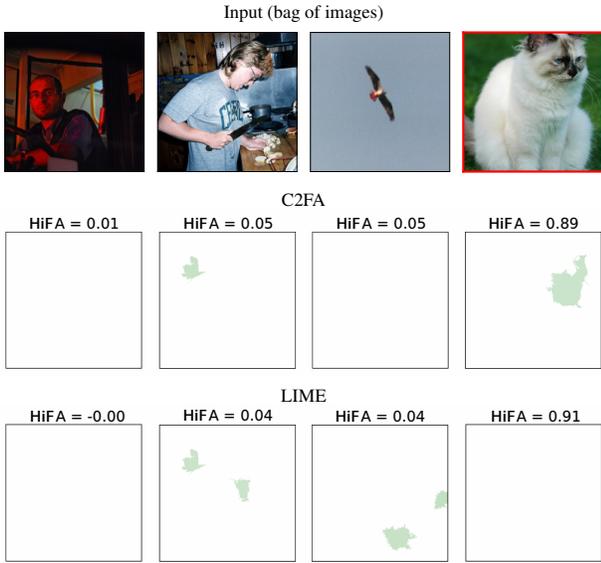


Figure 3: Example of the estimated HiFAs and LoFAs on the image classification task when $N_H = 20$ and $N_L = 50$. The input is shown on the first row, where the image with the red border is the positive instance. The LoFAs of super-pixels estimated by the proposed method and LIME are shown in the second and third rows, respectively. In each case, super-pixels are highlighted in green, with the intensity of the green color indicating the magnitude of the LoFA.

f , thanks to the simultaneous estimation of the HiFAs and LoFAs.

Figure 2c shows the consistency scores and the agreement scores of MIHL over various values of N_L where we fixed $N_H = 20$. Here, the consistency scores of BU-LIME, TD-LIME, and TD-MILLI are always zero by definition. We found that the consistency scores of LIME and MILLI were worse because they estimated the HiFAs and LoFAs separately. On the other hand, those of the proposed method were nearly zero, which means that the estimated HiFAs and LoFAs satisfied the consistency property. With the agreement scores of MIHL, we found that the proposed method outperformed the other methods regardless of the values of N_L , and the differences in the scores were especially noticeable for smaller N_L values (i.e., $N_L \leq 150$).

We visualize an example of the estimated HiFAs and LoFAs by the proposed method and the best-performing baseline, LIME, in Figure 3. Here, we only display the LoFAs larger than 0.1 for clarity. The figure shows that the proposed method assigned a high LoFA to the super-pixel in the high-level feature which was the positive instance and had the highest HiFA (HiFA = 0.89), although LIME assigned high LoFAs to the super-pixels in the negative instances. The critical difference between the two methods is whether the HiFAs and LoFAs are estimated simultaneously or separately. Since both the proposed method and LIME assigned the highest HiFA to the positive instance correctly, the result indicates that estimating the HiFAs and LoFAs simultaneously is effective.

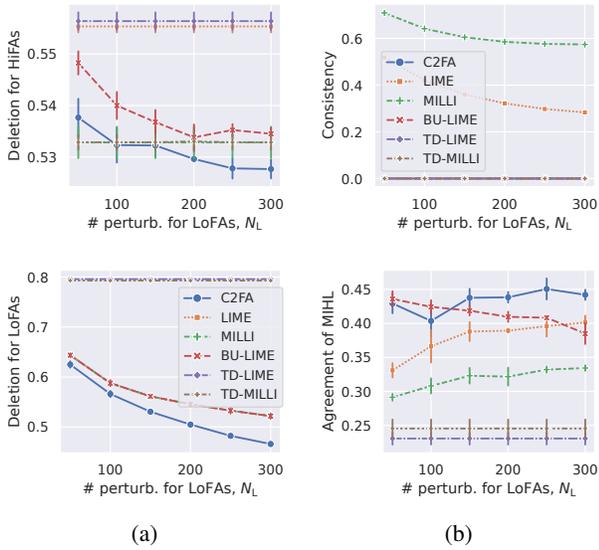


Figure 4: Quantitative evaluation on the text classification task. (a) Deletion scores of the HiFAs and LoFAs (lower is better). (b) Consistency scores (lower is better) and agreement scores of MIHL (higher is better).

tive.

5.2 Text Classification Using Language Models

Another practical application of the proposed method is to explain the attributions of sentences and the words in text classification with language models.

Dataset. We constructed a dataset in which the validation and test sets contain 500 and 1,000 product reviews, respectively, randomly sampled from the Amazon reviews dataset [Zhang *et al.*, 2015], respectively. Each sample consists of multiple sentences (high-level features), where each sentence is a sequence of words (low-level features). The label of each sample corresponds to the review’s sentiment polarity (positive or negative).

Black-box Model. We used BERT [Devlin *et al.*, 2018] fine-tuned on the original Amazon reviews dataset, which is provided on Hugging Face [fabricyhc (Hugging Face), 2022]. The test accuracy of the model is 0.947. When masking a word in the input to generate perturbed inputs, we replaced the word with the predefined mask token [MASK]. Similarly, when masking a sentence, we replaced all the words in the sentence with the mask token.

Quantitative Evaluation. Because no ground-truth labels for HiFAs and LoFAs are available in the dataset, we evaluated the estimated HiFAs and LoFAs only in terms of faithfulness and consistency, as in Section 5.1.

Results. Figure 4a shows the deletion scores of the estimated HiFAs and LoFAs over various values of N_L where we fixed $N_H = 5$ and 50, respectively. With the deletion scores of the HiFAs, although the scores of the proposed method were equal to or worse than those of MILLI and TD-MILLI at $N_L \leq 150$, the proposed method achieved the best deletion

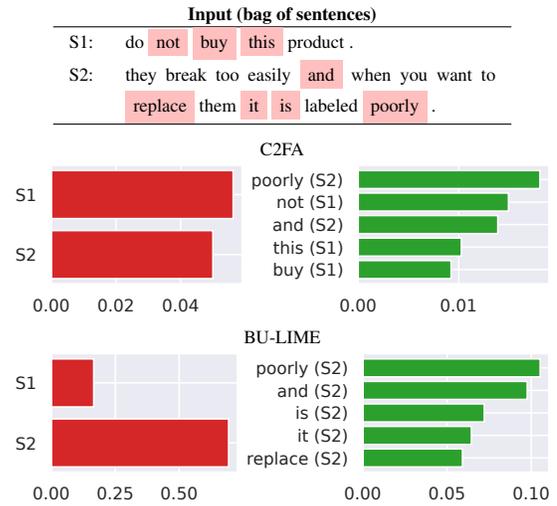


Figure 5: Example of the estimated HiFAs and LoFAs for a negative review text when $N_H = 50$ and $N_L = 50$. The review text is shown at the top, and the HiFAs (left) and the top-5 LoFAs (right) estimated by each method are shown at the bottom. Here, the words highlighted with a pink background in the review correspond to those top-5 LoFAs in the chart

score at $N_L \geq 200$. We found that in this task, the LIME-based methods, including the proposed method, were worse than the MILLI-based methods at small N_L values. As N_L increased, the proposed method benefited from the consistency constraints and became the only LIME-based method that outperformed the MILLI-based methods. With the deletion scores of the LoFAs, the proposed method outperformed the other methods regardless of the values of N_L .

Figure 4b shows the consistency scores and the agreement scores of MIHL over various values of N_L where we fixed $N_H = 50$. Again, in this task, the consistency scores of the proposed method were nearly zero regardless of the values of N_L . With the agreement scores of MIHL, the proposed method maintained high scores regardless of the values of N_L , although the scores of BU-LIME were slightly better than the proposed method at $N_L \leq 100$.

Figure 5 shows an example of the HiFAs and LoFAs estimated by the proposed method and the second-best method, BU-LIME. In the example, we fixed $N_H = 50$ and $N_L = 50$; that is, N_L is insufficient to estimate the LoFAs accurately. We found that although the baseline method assigned higher LoFAs to the words in the second sentence (S2), the proposed method assigned higher LoFAs to the words in the first sentence (S1). This is because the proposed method regularizes the LoFAs via the consistency constraints, exploiting the fact that S1 has a high HiFA.

6 Conclusion

We hypothesized that the consistency property, which is naturally derived from the characteristics of surrogate models, is essential for producing explanations that are accurate, faithful, and consistent between HiFAs and LoFAs. This property also enables the method to require fewer queries to the model.

To address this, we proposed a model-agnostic local explanation method for nested-structured inputs, that is capable of explaining two-level feature attributions. Our experiments on image and text classification tasks demonstrated that the proposed method can efficiently generate high-quality explanations while using fewer queries. Future work includes extending the method to handle more complex nested structures and exploring additional application domains.

A Optimization Algorithm

Our objective to estimate the HiFAs and LoFAs simultaneously is as follows:

$$\hat{\alpha}, \hat{\beta}^\dagger = \underset{\alpha, \beta^\dagger}{\operatorname{argmin}} \mathcal{L}_H(\alpha) + \mathcal{L}_L(\beta^\dagger) + \lambda_H \Omega_H(\alpha) + \lambda_L \Omega_L(\beta^\dagger)$$

$$\text{s.t. } \alpha_j = \sum_{d=1}^{D_j} \beta_{jd} \quad (\forall j \in [J]). \quad (6)$$

We solve the optimization based on the alternating direction method of multipliers (ADMM) [Boyd *et al.*, 2011]. By introducing auxiliary variables $\bar{\alpha} \in \mathbb{R}^J$ and $\bar{\beta}^\dagger \in \mathbb{R}^{D^\dagger}$ and Lagrange multipliers $v_1 \in \mathbb{R}^J$, $v_2 \in \mathbb{R}^{D^\dagger}$, and $v_3 \in \mathbb{R}^J$ based on the ADMM manner, our objective is rewritten as follows:

$$\hat{\alpha}, \hat{\beta}^\dagger = \underset{\alpha, \beta^\dagger}{\operatorname{argmin}} \mathcal{L}_H(\alpha) + \mathcal{L}_L(\beta^\dagger) + \lambda_H \Omega_H(\bar{\alpha}) + \lambda_L \Omega_L(\bar{\beta}^\dagger)$$

$$+ v_1^\top \mathbf{h}_1(\alpha, \bar{\alpha}) + v_2^\top \mathbf{h}_2(\beta^\dagger, \bar{\beta}^\dagger) + v_3^\top \mathbf{h}_3(\alpha, \beta^\dagger)$$

$$+ \frac{\mu_1}{2} \{ \|\mathbf{h}_1(\alpha, \bar{\alpha})\|^2 + \|\mathbf{h}_2(\beta^\dagger, \bar{\beta}^\dagger)\|^2 \}$$

$$+ \frac{\mu_2}{2} \|\mathbf{h}_3(\alpha, \beta^\dagger)\|^2, \quad (7)$$

where $\mathbf{h}_1(\alpha, \bar{\alpha}) = \alpha - \bar{\alpha}$, $\mathbf{h}_2(\beta^\dagger, \bar{\beta}^\dagger) = \beta^\dagger - \bar{\beta}^\dagger$, $\mathbf{h}_3(\alpha, \beta^\dagger) = \alpha - M\beta^\dagger$. Here, $M \in \{0, 1\}^{J \times D^\dagger}$ is a binary matrix to add up the LoFAs associated with the same high-level feature where we set to $M_{jd} = 1$ if the d th feature of the concatenated input x^\dagger belongs to the j th high-level feature x_j , and $M_{jd} = 0$ otherwise. The hyperparameters $\mu_1 \geq 0$ and $\mu_2 \geq 0$ are the penalty parameters for the regularization and the consistency constraint, respectively.

The optimization of (7) is performed by alternating the updates of the variables iteratively. We summarize the optimization algorithm using ℓ_2 regularization for Ω_H and Ω_L in Algorithm 1. Here, I_J and I_{D^\dagger} are identity matrices of size J and D^\dagger , respectively, and in Line 9, the variables at the zeroth step are initialized with zero. The algorithm is terminated when $\|\bar{\alpha}^{t-1} - \bar{\alpha}^t\|^2 + \|\bar{\beta}^{\dagger t-1} - \bar{\beta}^{\dagger t}\|^2 < \epsilon_1$ and $\|\mathbf{h}_1(\alpha^t, \bar{\alpha}^t)\|^2 + \|\mathbf{h}_2(\beta^{\dagger t}, \bar{\beta}^{\dagger t})\|^2 + \|\mathbf{h}_3(\alpha^t, \beta^{\dagger t})\|^2 < \epsilon_2$ where $\epsilon_1, \epsilon_2 \geq 0$ are hyperparameters. The other hyperparameters of the algorithm are λ_H , λ_L , μ_1 , and μ_2 .

B Experiments on Image Classification in Multiple Instance Learning

Detailed Description of Dataset. We constructed an MIL dataset from the Pascal VOC semantic segmentation

Algorithm 1 Estimating consistent two-level feature attributions (C2FA) with ℓ_2 regularization

- 1: Generate binary random matrices Z^H and Z^L
- 2: Obtain perturbed inputs $\{\tilde{X}_n^H\}_{n=1}^{N_H}$ and $\{\tilde{X}_n^L\}_{n=1}^{N_L}$ using ϕ_H and ϕ_L
- 3: Obtain predictions \tilde{y}^H and \tilde{y}^L from the perturbed inputs
- 4: Obtain weight matrices W^H and W^L
- 5: $A = (Z^{H^\top} W^H Z^H + (\mu_1 + \mu_2) I_J)^{-1}$
- 6: $B = A Z^{H^\top} W^H \tilde{y}^H$
- 7: $C = (Z^{L^\top} W^L Z^L + \mu_1 I_{D^\dagger} + \mu_2 M^\top M)^{-1}$
- 8: $D = C Z^{L^\top} W^L \tilde{y}^L$
- 9: Initialize $\alpha^0, \bar{\alpha}^0, \beta^{\dagger 0}, \bar{\beta}^{\dagger 0}, v_1^0, v_2^0, v_3^0$ with zero
- 10: $t = 0$
- 11: **repeat**
- 12: $\alpha^{t+1} = B + A(\mu_2 M \beta^{\dagger t} - \mu_1 \bar{\alpha}^t - v_1^t - v_3^t)$
- 13: $\bar{\alpha}^{t+1} = (\mu_1 + 2\lambda_H)^{-1}(v_1^t + \mu_1 \alpha^{t+1})$
- 14: $\beta^{\dagger t+1} = D + C(M^\top v_3^t + \mu_1 \bar{\beta}^{\dagger t} + \mu_2 M^\top \alpha^{t+1} - v_2^t)$
- 15: $\bar{\beta}^{\dagger t+1} = (\mu_1 + 2\lambda_L)^{-1}(v_2^t + \mu_1 \beta^{\dagger t+1})$
- 16: $v_1^{t+1} = v_1^t + \mu_1(\alpha^{t+1} - \bar{\alpha}^{t+1})$
- 17: $v_2^{t+1} = v_2^t + \mu_1(\beta^{\dagger t+1} - \bar{\beta}^{\dagger t+1})$
- 18: $v_3^{t+1} = v_3^t + \mu_2(\alpha^{t+1} - M\beta^{\dagger t+1})$
- 19: $t = t + 1$
- 20: **until** stop criterion is met
- 21: **return:** $\bar{\alpha}^t, \bar{\beta}^{\dagger t}$

dataset [Everingham *et al.*, 2015] that allows us to evaluate the estimated HiFAs and LoFAs with the ground-truth instance- and pixel-level labels. With the training subset of the dataset, each bag contains **three to five images**, randomly selected from the Pascal VOC dataset. Here, low-level features correspond to regions (super-pixels) of each image, which are obtained by the quick shift algorithm [Vedaldi and Soatto, 2008]. Each bag is labeled positive if at least one image in the bag is associated with the ‘‘cat’’ label; otherwise the bag is negative. Also, each image pixel is labeled positive if the pixel is associated with ‘‘cat’’ label and negative otherwise. We used the instance- and pixel-level supervision only for evaluation. Similarly, we constructed validation and test subsets whose samples contain images from the training and test subsets of the Pascal VOC, respectively. The number of samples in training, validation, and test subsets is 5,000, 1,000, and 2,000, respectively, and the positive and negative samples ratio is equal.

Detailed Description of Dataset. We defined the DeepSets permutation-invariant model [Zaheer *et al.*, 2017] as a black-box model f to be explained. According to [Zaheer *et al.*, 2017], the model f comprises two components: a representation function that transforms each instance, ϕ , and a non-linear network that produces predictions from the extracted representation, ρ . We used ResNet-50 [He *et al.*, 2016] pre-trained on ImageNet as the representation function ϕ and two-layer multi-layer perceptron (MLP) as the non-linear network ρ . Here, in ρ , we used the ReLU activation function for the first layer and the softmax function for the second layer. Also, the number of hidden units in the MLP was set to 1,024. The

model first extracts the representation of each instance using ϕ , then adds them up into a single representation, and finally, outputs a prediction by applying ρ to the aggregated single representation. We trained the model using our MIL image classification dataset with Adam optimizer [Kingma and Ba, 2015] with a learning rate of 0.001, a batch size of 32, and a maximum epoch of 300. The test accuracy of the model was 0.945.

The hyperparameters of C2FA, λ_H , λ_L , and μ_1 , were tuned using the validation subset of each dataset within the following ranges: $\lambda_H, \lambda_L \in \{0.1, 1\}$, and $\mu_2 \in \{0.001, 0.01, 0.1\}$. The remaining hyperparameters were set to $\mu_1 = 0.1$ and $\epsilon_1 = \epsilon_2 = 10^{-4}$. The experiments were conducted on a server with an Intel Xeon Gold 6148 CPU and an NVIDIA Tesla V100 GPU.

Detailed Description of Quantitative Evaluation Metrics.

We assessed the estimated HiFAs and LoFAs in terms of correctness, faithfulness, and consistency. The correctness is evaluated using the ground-truth instance- and pixel-level labels. Following the evaluation in the MIL study [Early *et al.*, 2022], we evaluated the estimated HiFAs using normalized discounted cumulative gain (NDCG). For the estimated LoFAs, as with the evaluation of the LoFAs for single image classification [Sampaio and Cordeiro, 2023], we evaluated them as the predictions of the pixel-level labels by the area under ROC curve (AUROC) in the binary semantic segmentation manner. For faithfulness, we assessed whether the estimated HiFAs and LoFAs are faithful to the behaviors of the model f using insertion and deletion metrics [Petsiuk *et al.*, 2018]. The insertion and deletion metrics evaluate the change in the predictions of the model f when features deemed important in the LoFAs are gradually added and removed from the sample, respectively. In our experiments, we gradually add and remove the low-level features across all the high-level features in descending order of their LoFAs. Also, for the HiFAs, we add and remove the high-level features instead of the low-level ones, respectively. In terms of the consistency evaluation, we used the following two metrics. The first one is the consistency between the estimated HiFAs and LoFAs, which is calculated with $\|\alpha - M\beta^\dagger\|^2$ used to calculate the penalty for the consistency constraints in (7). The second one is the agreement of the most important high- and low-level feature (MIHL), which is calculated by the ratio that the high-level feature of the highest HiFA is identical to the one associated with the low-level feature of the highest LoFA.

We evaluated the above metrics using only the samples with the positive bag label because we could not evaluate the correctness of those with the negative bag label. We ran the evaluations three times with different random seeds and reported the average scores and their standard deviation.

C Proof Sketches

Sketch of Proof for Lemma 4.6. By assumptions, the loss functions $\mathcal{L}_H(\cdot)$, $\mathcal{L}_L(\cdot)$, as well as the regularizers $\Omega_H(\cdot)$, $\Omega_L(\cdot)$ are convex in their respective parameters. Since the objective function \mathcal{J} is convex, lower bounded, and the feasible set Γ° is non-empty, closed and convex, by the Weierstrass extreme value theorem, there exists at least one global

minimizer $(\alpha^*, \beta^*) \in \Gamma^\circ$. \square

Sketch of Proof for Lemma 4.7. Under Assumption 4.3 each quadratic loss is strictly convex. Adding any convex regulariser keeps strict convexity. The sum of two strictly convex functions on a linear subspace is strictly convex, so the constrained optimisation has a single global minimiser. \square

Sketch of Proof for Theorem 4.8. Let $\tilde{\alpha}$ and $\tilde{\beta}$ be the separate ridge/LIME minimisers. By assumption they violate at least one consistency equation. Project them onto Γ° by defining $\alpha_j^\circ := \sum_d \tilde{\beta}_{jd}$, $\beta^\circ := \tilde{\beta}$. Owing to strict convexity, the affine projection strictly decreases the joint objective unless the constraint already held. The global constrained minimiser therefore attains strictly lower objective value than the naive concatenation $(\alpha^\circ, \beta^\circ)$. \square

Sketch of Proof for Theorem 4.9. Treat the joint parameter vector $\theta = (\alpha, \beta) \in \mathbb{R}^{J+D^\dagger}$, lying in a linear subspace of rank $J + D^\dagger - J$. With sub-Gaussian masks, empirical Gram matrices for both high- and low-level designs satisfy a restricted-eigenvalue (RE) bound with probability $1 - \delta$ as $\lambda_{\min}(\hat{\Sigma})$. Standard stochastic-convex-optimisation argument gives excess-risk and constant C absorbs RE constant, variance proxy, and regularization strengths. \square

Sketch of Proof for Corollary 4.10. Split the pointwise approximation error into (population) linear-model bias plus parameter-estimation error, and under bounded design $\|z^H\| \leq B_H$ the first term is bounded by $B_H \|\hat{\alpha} - \alpha^*\|_2$. Insert the rate from Theorem 4.8 and add the analogous low-level term, yielding the stated uniform bound with probability $1 - \delta$. \square

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 22K17953.

References

- [Boyd *et al.*, 2011] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers*. Now Foundations and Trends, 2011.
- [Cheplygina *et al.*, 2019] Veronika Cheplygina, Marleen de Bruijne, and Josien P.W. Pluim. Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical Image Analysis*, 54:280–296, 2019.
- [Danilevsky *et al.*, 2020] Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. A Survey of the State of Explainable AI for Natural Language Processing. In *Proceedings of the Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, pages 447–459, 2020.

- [Dara *et al.*, 2020] Sriharsha Dara, C Ravindranath Chowdary, and Chintoo Kumar. A survey on group recommender systems. *Journal of Intelligent Information Systems*, 54(2):271–295, 2020.
- [Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [Došilović *et al.*, 2018] Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. Explainable artificial intelligence: A survey. In *Proceedings of the International Convention on Information and Communication Technology, Electronics and Microelectronics*, pages 210–215, 2018.
- [Early *et al.*, 2022] Joseph Early, Christine Evers, and Sarvapali Ramchurn. In *Proceedings of the International Conference on Learning Representations*, 2022.
- [Everingham *et al.*, 2015] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes Challenge: A Retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015.
- [fabriceyh (Hugging Face), 2022] fabriceyh (Hugging Face). `fabriceyh/bert-base-uncased-amazon_polarity`, 2022. Accessed: 2024-2-2.
- [Guo *et al.*, 2021] Yulan Guo, Hanyun Wang, Qingyong Hu, Hao Liu, Li Liu, and Mohammed Bennis. Deep Learning for 3D Point Clouds: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12):4338–4364, 2021.
- [He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.
- [Ilse *et al.*, 2018] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based Deep Multiple Instance Learning. In *Proceedings of the International Conference on Machine Learning*, volume 80, pages 2127–2136, 2018.
- [Jain *et al.*, 2022] Saachi Jain, Hadi Salman, Eric Wong, Pengchuan Zhang, Vibhav Vineet, Sai Vemprala, and Aleksander Madry. Missingness Bias in Model Debugging. In *Proceedings of the International Conference on Learning Representations*, 2022.
- [Javed *et al.*, 2022] Syed Ashar Javed, Dinkar Juyal, Harshith Padigela, Amaro Taylor-Weiner, Limin Yu, and Aaditya Prakash. Additive MIL: intrinsically interpretable multiple instance learning for pathology. *Advances in Neural Information Processing Systems*, pages 20689–20702, 2022.
- [Kimura *et al.*, 2024] Masanari Kimura, Ryotaro Shimizu, Yuki Hirakawa, Ryosuke Goto, and Yuki Saito. On permutation-invariant neural networks. *arXiv preprint arXiv:2403.17410*, 2024.
- [Kingma and Ba, 2015] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*, 2015.
- [Lundberg and Lee, 2017] Scott M Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, pages 4765–4774, 2017.
- [Mosca *et al.*, 2022] Edoardo Mosca, Defne Demirtürk, Luca Mülln, Fabio Raffagnato, and Georg Groh. GramSHAP: An efficient model-agnostic and structure-aware NLP explainer. In *Proceedings of the ACL Workshop on Learning with Natural Language Supervision*, pages 10–16, 2022.
- [Petsiuk *et al.*, 2018] Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized Input Sampling for Explanation of Black-box Models. In *Proceedings of the British Machine Vision Conference*, 2018.
- [Plumb *et al.*, 2018] Gregory Plumb, Denali Molitor, and Ameeth S Talwalkar. Model Agnostic Supervised Local Explanations. *Advances in Neural Information Processing Systems*, pages 4768–4777, 2018.
- [Ribeiro *et al.*, 2016] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “Why Should I Trust You?” Explaining the Predictions of Any Classifier. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.
- [Ribeiro *et al.*, 2018] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-Precision Model-Agnostic Explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32(1), 2018.
- [Rychener *et al.*, 2023] Yves Rychener, Xavier Renard, Djamé Seddah, Pascal Frossard, and Marcin Detryniecki. On the Granularity of Explanations in Model Agnostic NLP Interpretability. In *Proceedings of the Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pages 498–512, 2023.
- [Saeed and Omlin, 2023] Waddah Saeed and Christian Omlin. Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*, 263:110273, 2023.
- [Sampaio and Cordeiro, 2023] Vicente Sampaio and Filipe R. Cordeiro. Improving Mass Detection in Mammography Images: A Study of Weakly Supervised Learning and Class Activation Map Methods. In *Proceedings of the Conference on Graphics, Patterns and Images*, pages 139–144, 2023.
- [shap (GitHub), 2024] shap (GitHub). `shap`: A game theoretic approach to explain the output of any machine learning model, 2024. Accessed: 2024-2-2.
- [Tan and Kotthaus, 2022] Hanxiao Tan and Helena Kotthaus. Surrogate model-based explainability methods for point cloud NNs. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2927–2936, 2022.

- [van der Velden *et al.*, 2022] Bas H M van der Velden, Hugo J Kuijf, Kenneth G A Gilhuijs, and Max A Viergever. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Medical image analysis*, 79:102470, 2022.
- [Vedaldi and Soatto, 2008] Andrea Vedaldi and Stefano Soatto. Quick Shift and Kernel Methods for Mode Seeking. In *Proceedings of the European Conference on Computer Vision*, pages 705–718, 2008.
- [Zafar *et al.*, 2021] Muhammad Bilal Zafar, Philipp Schmidt, Michele Donini, Cédric Archambeau, Felix Biessmann, Sanjiv Ranjan Das, and Krishnaram Kenthapadi. More Than Words: Towards Better Quality Interpretations of Text Classifiers. *arXiv preprint arXiv:2112.12444*, 2021.
- [Zaheer *et al.*, 2017] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan R Salakhutdinov, and Alexander J Smola. Deep Sets. *Advances in Neural Information Processing Systems*, pages 3391–3401, 2017.
- [Zhang *et al.*, 2015] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. *Advances in Neural Information Processing Systems*, 28:649–657, 2015.