Measuring the (Un)Faithfulness of Concept-Based Explanations

Shubham Kumar
University of Illinois Urbana-Champaign
sk138@illinois.edu

Narendra Ahuja
University of Illinois Urbana-Champaign
n-ahuja@illinois.edu

Abstract

Deep vision models perform input-output computations that are hard to interpret. Concept-based explanation methods (CBEMs) increase interpretability by re-expressing parts of the model with human-understandable semantic units, or concepts. Checking if the derived explanations are faithful—that is, they represent the model's internal computation—requires a surrogate that combines concepts to compute the output. Simplifications made for interpretability inevitably reduce faithfulness, resulting in a tradeoff between the two. State-of-the-art unsupervised CBEMs (U-CBEMs) are seemingly more interpretable, while also being more faithful to the model. However, we observe that the reported improvement in faithfulness artificially results from either (1) using overly complex surrogates, which introduces an unmeasured cost to the explanation's interpretability, or (2) relying on deletion-based approaches that, as we demonstrate, do not properly measure faithfulness. We propose ***Surrogate Faithfulness (SURF)****, which (1) replaces prior complex surrogates with a simple, linear surrogate that measures faithfulness without changing the explanation's interpretability and (2) introduces well-motivated metrics that assess loss across all output classes, not just the predicted class. We validate SURF with a measure-over-measure study by proposing a simple sanity check—explanations with random concepts should be less faithful—which prior surrogates fail. SURF enables the first reliable faithfulness benchmark of U-CBEMs, revealing that many visually compelling U-CBEMs are not faithful.* ***Code is released.***

1. Introduction

Deep learning models have delivered state-of-the-art (SOTA) results across diverse tasks, yet their internal computation remain difficult to interpret [17], which is especially problematic in high-stakes domains such as healthcare and finance. This gap has given rise to explainable AI (XAI), whose goal is to produce *explanations* of model behavior. Two properties characterize explanation quality: *interpretability*—explanations that humans can understand—and *faithfulness*—explanations that reflect the model's internal computation. XAI methods must balance the natural tension between interpretability and faithfulness. While interpretability is assessed with human studies, faithfulness must be evaluated by defining a *surrogate* that maps the explanation to the model's outputs and measuring the loss between the surrogate and model. Since explanations are inevitably lossy, some discrepancy is expected. For certain XAI methods, defining a surrogate is challenging, so alternative measures, or *proxies*, attempt to evaluate faithfulness using different criteria.

One broad family of XAI methods constructs *inherently interpretable models*, injecting an explanatory structure directly into the model [3, 6]. Merging the explanation and the model eliminates the need for faithfulness checks; however, these methods often underperform black-box models by reducing model complexity (capacity) for explainability. In contrast, *feature attribution* methods express a trained model's outputs in terms of each input feature's contribution, commonly done by finding the output's sensitivity to changes in input features, either through gradients or permutations [10, 33, 37, 39, 47]. For images, attributions are visualized as pixel-level heatmaps. Although such explanations are intuitive, developing faithfulness measures has been challenging, since many attribution methods do not specify a surrogate to reconstruct the model's output from the explanation. Thus, proxies—such as checking if important pixels come from foreground regions or looking for significant model degradation after deleting important pixels—are used to study faithfulness. Despite strong performance on the proxies, clever sanity checks have revealed faithfulness problems [1, 15, 22, 41]. The resulting heatmaps also lack interpretability—they only indicate *where* the model attends, not *what* semantics it recognizes [7, 35].

Concept-based explanation methods (CBEMs) improve interpretability by explaining predictions in terms of human-understandable *concepts* (e.g., edges, colors, object parts) [4, 14, 24, 55]. To recognize concepts, supervised CBEMs require a concept-annotated image dataset, which is costly and subjective; moreover, any fixed concept vo-

cabulary may bias the explanation and miss important aspects of the model's computation, raising faithfulness concerns [38]. To avoid these issues, *unsupervised* CBEMs (U-CBEMs) discover concepts automatically as *concept activation vectors* (CAVs)—directions in the model's intermediate representation space—paired with a *concept importance* score that captures the concept's relevance to the output. Since CBEMs operate on the model's intermediate representations, defining a surrogate is more natural, paving the way for proper faithfulness evaluations.

However, a closer look shows that U-CBEMs have adopted inadequate faithfulness measures. We discover that existing evaluations rely on complex surrogates, which allow U-CBEMs to simultaneously show users interpretable explanations and report high faithfulness, yet the explanations do not clearly lead to the model's output. Furthermore, U-CBEMs adapt popular deletion-based proxies from the attribution literature, yet Sec. 4.1 outlines serious unresolved limitations that prevent them from properly measuring faithfulness. Finally, *each* work that proposes a new U-CBEM also evaluates it with a *new* faithfulness measure, with no measure-over-measure comparison, indicating a lack of consensus on faithfulness measures in the field.

We argue that these issues have misled us into believing that current U-CBEMs are faithful. To shed light on this, we make the following contributions:

1. **Organize prior U-CBEM faithfulness measures** under a common framework, allowing us to discuss their limitations and identify appropriate desiderata.
2. **Propose Surrogate Faithfulness (SURF)**, a faithfulness measure satisfying the desiderata. Our *measure-over-measure comparison* checks if faithfulness decreases as the explanation is increasingly randomized; only SURF passes this check.
3. **Conduct the first, comprehensive faithfulness benchmark of current U-CBEMs** across a variety of tasks and model architectures. SURF shows that SOTA U-CBEMs previously evaluated to be faithful are not.
4. Leverage SURF to **provide a selection criterion for the number of concepts** U-CBEMs discover, improving on prior work that sets this hyperparameter arbitrarily.

2. Preliminaries

We first introduce notation used throughout the paper, as well as preliminaries to familiarize the reader with the field.

Background: Model $\phi : \mathcal{X} \to \mathcal{Y}$ maps from an input space $\mathcal{X} \subseteq \mathbb{R}^P$ to an output space $\mathcal{Y} \subseteq \mathbb{R}^C$. Assume ϕ admits an intermediate space $\mathcal{H} \subseteq \mathbb{R}^D$. Let $g : \mathcal{X} \to \mathcal{H}$ and $f : \mathcal{H} \to \mathcal{Y}$. Thus, $\mathbf{y} = \phi(\mathbf{X}) = f(g(\mathbf{X}))$. Let $g(\mathbf{x}) = \mathbf{h} \in \mathcal{H}$ represent the embedding of $\mathbf{x}$ (e.g., image patch), and $g(\mathbf{X}) = \mathbf{H} \in \mathbb{R}^{(HW)\times D}$ denote the vectorized application of g on $\mathbf{X} = \begin{bmatrix}\mathbf{x}_1 & \cdots & \mathbf{x}_{HW}\end{bmatrix}^T$, where H, W are the spatial dimensions.

Let E be an explanation function, and let $E(\mathbf{X}) \in \mathcal{E}$ represent the explanation for an $\mathbf{X} \in \mathcal{X}$ (e.g., image). Let $s : \mathcal{E} \to \mathcal{Y}$ be a surrogate. Motivated by Ribeiro et al. [39], E is completely faithful to ϕ if s reproduces $\phi(\mathbf{X})\ \forall\ \mathbf{X} \in \mathcal{X}$. Thus, a faithfulness measure is defined by a choice of surrogate $s(.)$ and metric $d(.)$ as:

$$Faith(E; d, s) = \int_{\mathcal{X}} d(\phi(\mathbf{X}), s(E(\mathbf{X})))d\mathbf{X} \tag{1}$$

U-CBEMs: To generate explanations, U-CBEMs find, for all outputs $i \in C$, a set of K CAVs $V_i = \{\mathbf{v}_{i,k}\}_{k=1}^K$ and concept importances $A_i = \{\alpha_{i,k}\}_{k=1}^K$. Note that while, for a given input, K can vary per output class, all prior works find the same K per output class. U-CBEMs also define a mechanism $\mathcal{P}(\mathbf{h}; V) : \mathcal{H} \to \mathbb{R}^K$ to project embedding $\mathbf{h}$ to the concept space defined by CAVs in V. The explanation for output i is:

$$\begin{aligned} E_i(\mathbf{X}) &= E_i(\mathbf{X}; g, V_i, A_i, \mathcal{P}) \\ &= \{\mathcal{P}(g(\mathbf{X}); V_i), A_i\} \end{aligned} \tag{2}$$

$E_i(\mathbf{X})$ is then conveyed to the user through some visual interface. Note that $\mathcal{P}(g(\mathbf{X}))$ denotes the vectorized application of $\mathcal{P}$ on $g(\mathbf{X}) = \mathbf{H}$.

Following prior works, we study faithfulness in the final layer, where the spatial dimensions $H = W = 1$, so $\mathbf{H} = \mathbf{h}$. Given that U-CBEMs operate on $g(\mathbf{X})$, the faithfulness measure can be simplified to:

$$\begin{aligned} &Faith_{\text{U-CBEM}}(\mathcal{P}, \{V_i\}_{i=1}^C, \{A_i\}_{i=1}^C; d, s) = \\ &\int_{\mathcal{H}} d(f(\mathbf{h}), \{s(\mathcal{P}(\mathbf{h}; V_i), A_i)\}_{i=1}^C)d\mathbf{h} \end{aligned} \tag{3}$$

3. Related Works on U-CBEMs

Wang et al. [46] find CAVs by K-Means clustering intermediate embeddings computed on a dataset. Each cluster is associated with a semantic concept (e.g., car tire, window, headlight) by tracing back the clustered embeddings to their corresponding image patches. Ghorbani et al. [18] improve this with ACE, which finds concepts at different scales by hierarchically segmenting images into "candidate" segments and clustering the embeddings of each candidate. To ensure that the discovered CAVs fully represent the model's behavior, Yeh et al. [48] propose ConceptSHAP (C-SHAP), which finds CAVs that maximize a proposed completeness score; they use Shapley Values [50] to find concept importance. ICE [53] extends ACE by replacing clustering with Non-negative Matrix Factorization (NMF), resulting in a parts-based explanation. The authors show that concepts found with NMF are more interpretable than those found by K-Means or Principal Component Analysis

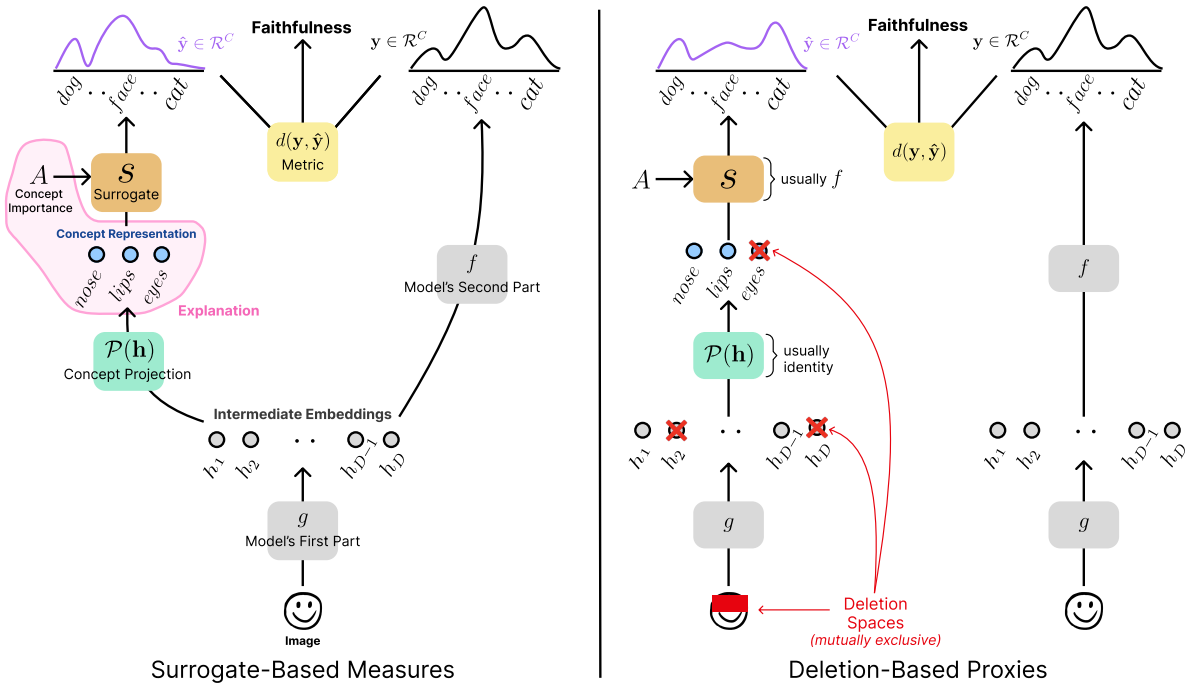

Figure 1. **Framework.** U-CBEM faithfulness measures compare (using metric d) the output $\mathbf{y}$ of the original model $\phi(.) = f(g(.))$ with the output $\hat{\mathbf{y}}$ from the explanation. To obtain the explanation, U-CBEMs transform intermediate representation $\mathbf{h} \in \mathcal{R}^D$ to the concept representation through concept projection $\mathcal{P}(\mathbf{h})$ and provide an accompanying concept importance A. The explanation is then passed through surrogate s to obtain $\hat{\mathbf{y}}$. Deletion-based proxies (right) observe model degradation after performing deletion in a deletion space. Surrogate-based measures (left) do not manipulate the computation; instead, they introduce a surrogate to directly approximate Eq. (3).

(PCA). Building on this, Fel et al. [12] present CRAFT, which recursively applies NMF through the model to decompose concepts into sub-concepts, and they use sensitivity analysis for finding concept importance. CDISCO [20] uses the singular value decomposition to find CAVs and a gradient-based method for finding concept importance. MCD [44] extends the notion of CAVs to a multidimensional linear subspace, rather than a single vector, allowing it to describe more of the model's behavior with fewer concepts. Finding issues with MCD's interpretability, HU-MCD [21] incorporates Segment Anything Model [27] to discover more interpretable concepts. Recently, sparse autoencoders (SAEs) have emerged as a promising, scalable U-CBEM for large language models, and recent attempts have applied them to interpret vision models [13, 19, 31].

CBEMs and U-CBEMs should not be confused with Concept Bottleneck Models (CBMs), which create an inherently explainable concept model meant to replace the original black-box model [25, 28, 36, 51].

4. Unifying U-CBEM Faithfulness Measures

Each U-CBEM mentioned in Sec. 3 introduces a *different faithfulness measure*, indicating a lack of consensus and preventing fair comparisons of faithfulness across U-CBEMs. We unify faithfulness measures under a common framework, shown in Fig. 1. Any evaluation measure compares the output of the original model with the output from the explanation (*Metric*). To obtain the output from the explanation, a *Concept Projection* transforms the intermediate representation to a concept representation, which is then passed through a *Surrogate*. Deletion-based proxies assert that removing the most important concepts will result in the greatest model degradation; more degradation is interpreted as a signal for higher faithfulness. Alternatively, surrogate-based measures avoid changing the input and introduce a surrogate to directly approximate faithfulness as in Eq. (3). Appendix Sec. A offers a discussion and comparison with faithfulness metrics introduced for SAEs.

4.1. Deletion-Based Proxies

To perform deletion, deletion-based proxies must remove concepts *(Deletion Method)* from the input's *Deletion Space* in decreasing order of importance.

Deletion Space: Where concepts are removed. This is either the image space, the model's weight space, or the concept space (shown in Fig. 1 (right)).

Deletion Method: How concepts are removed from the *Deletion Space*, usually by setting the concept to a baseline value (e.g., 0).

Table 1. **Prior deletion-based proxies (top) & surrogate-based measures (bottom)** are organized according to their differing factors: *Deletion Space* (DS), *Deletion Method* (DM), *Metric*, *Surrogate*, and *Concept Projection* (CP). Each prior work uses a different proxy or measure, with no measure-over-measure comparison.

Method	DS	DM	Metric	Surrogate	CP
ACE	Pixel	Constant (grey)	Class Accuracy	Original Model	Identity
MCD	Pixel	Inpainting	Class Accuracy	Original Model	Identity
HU-MCD	Pixel	Masking	Class Accuracy	Original Model	Identity
CDISCO (M1)	Pixel	Constant (grey)	# Classes with $> 80\%$ Accuracy Loss	Original Model	Identity
CDISCO (M2)	Weight	Constant (zero)	Class Accuracy	Original Model	Identity
CRAFT	Concept	Constant (zero)	Class Logit	Reconstruct $\rightarrow$ Original Model	U-CBEM
ICE	–	–	Normalized Target Class L1 Logit Loss Top-1 Accuracy	U-CBEM Reconstruct $\rightarrow$ Original Model	U-CBEM
C-SHAP	–	–	Normalized Top-1 Accuracy	MLP Reconstruct $\rightarrow$ Original Model	U-CBEM

Except for CRAFT, proxies set the concept projection to identity and the surrogate to $f(.)$ from the original model. Prior proxies (organized in Tab. 1) make different choices along each *italicized factor* mentioned above. **ACE** deletes concepts in the image space by setting associated pixels to a constant value (Constant Deletion). They measure the model's classification accuracy after each deletion across a dataset. **MCD** measures classification accuracy after deleting concepts in the pixel space with inpainting. **HU-MCD** similarly measures classification accuracy but performs deletions through a masking strategy which simulates running the model on the irregularly-shaped input, better ignoring masked regions. **CDISCO** has two measures. The first measure (M1) makes Constant Deletions on pixels and measures the number of classes whose accuracy degrades by more than 80%. The second measure (M2) deletes concepts by zeroing out model parameters associated with the concept, and they report classification accuracy. **CRAFT** uses the U-CBEMs defined $\mathcal{P}$ to transform $\mathbf{h}$ to a concept representation. It makes Constant Deletions in the concept space by setting the concept to zero; then, it reconstructs $\mathbf{h}$ from the perturbed concept representation and measures the change in the corresponding class's logit score.

There are two notable issues with deletion-based proxies that prevent them from accurately evaluating faithfulness. **First, it is unclear how to delete a concept.** Concepts are usually deleted by Constant Deletion, commonly with a baseline of $\mathbf{0}$. This baseline must entirely delete the presence of the concept without affecting other concepts. However, the feature attribution literature has shown that common baseline choices do not guarantee that a feature is completely deleted without affecting other features, leading to unfaithful explanations [15]. Removing features by marginalization is more accurate but requires evaluating expensive, high-dimensional expectations.

Second, there are no guarantees that representations in the Deletion Space will stay on manifold after deletion. The feature attribution literature [2, 15, 42, 43, 49] has found that the perturbed inputs used for creating explanations may be unrealistic, lying off the data manifold. In the off-manifold regions of the data space, a highly non-linear model behaves unpredictably, so the model's outputs may not be meaningfully related to input features. Thus, calculating feature importance with off-manifold inputs can lead to incorrect explanations [15]. More seriously, these explanations can be adversarially attacked; in [42], a model's off-manifold behavior was modified to hide its dependence on undesirable features used for on-manifold inputs. Such issues can be mitigated by using on-manifold input perturbations, but this requires accurate modeling of the input distribution [15, 22, 43, 45, 52]. These problems are not unique to feature attribution; they extend to deletions in the concept and weight space, which make similar manipulations in the model's intermediate representation space.

Thus, we argue that deletion-based proxies are unreliable, and we avoid using or benchmarking against them.

4.2. Surrogate-Based Measures

Any surrogate-based measure must choose the surrogate s and the metric d. For these measures, the input image, and thus the intermediate representation, is the same in both paths of Fig. 1 (left), and the concept projection is defined by the U-CBEM being evaluated.

4.2.1. Desiderata

The faithfulness definition in Eq. (3) does not constrain the surrogate s, but there are certain desiderata our surrogate should satisfy. **First**, s should be as simple as possible. Recognize that s reflects the mental computation any human interpreter is expected to do when trying to connect the explanation to the model's prediction. An interpretable explanation that requires a complex s just shifts the interpretation issue downstream; the human interpreter still lacks insight into how the model reaches its prediction. **Second**, s should incorporate all components of the explanation. For example, U-CBEMs explain with both V_i and A_i, so both should be used by s. If a component is not included, its

impact on faithfulness cannot be measured. **Third**, we desire metric(s) d that rewards explanations that closely reconstruct *all* of $\mathbf{y}$ (e.g. across all classes), not just specific (e.g., predicted) classes.

4.2.2. Prior Surrogate-Based Measures

Two prior works have proposed surrogate-based measures:

ICE-Eval: ICE introduces its own evaluation measure (we term as ICE-Eval). ICE-Eval's surrogate assumes the U-CBEM's projection operation P has a reconstruction function $\tilde{\mathcal{P}}^{-1} : \mathbb{R}^K \rightarrow \mathcal{H}$. Their surrogate then is:

$$\hat{\mathbf{y}}_i = s_i(.) = f_i(\tilde{\mathcal{P}}^{-1}(\mathcal{P}(\mathbf{h}; V_i))) \;\; \forall\, i \tag{4}$$

where $f_i(.)$ denotes the i'th output from function $f(.)$. ICE-Eval uses two metrics (only applicable to classification):

$$ICE_1 = \frac{1}{|\mathcal{V}|} \sum_{\mathbf{x}, t \in \mathcal{V}} \frac{|y_t - \hat{y}_t|}{|y_t|} \tag{5}$$

where t is the index corresponding to $\mathbf{x}$'s groundtruth class and $\mathcal{V}$ is the set of inputs used to evaluate faithfulness. Letting $\mathbf{p}$ denote class probability scores, the second metric is:

$$ICE_2 = \frac{1}{|\mathcal{V}|} \sum_{\mathbf{x} \in \mathcal{V}} \mathbb{1}\{argmax(\mathbf{p}) = argmax(\hat{\mathbf{p}})\} \tag{6}$$

C-SHAP-Eval: C-SHAP introduces its own evaluation measure (we term as C-SHAP-Eval). Its surrogate learns a two-layer perceptron $\psi_{\text{C-SHAP}} : \mathbb{R}^K \rightarrow \mathcal{H}$:

$$\hat{\mathbf{y}}_i = s_i(.) = f_i(\psi_{\text{C-SHAP}}(\mathcal{P}(\mathbf{h}; V_i))) \;\; \forall\, i \tag{7}$$

The metric they define is:

$$CSHAP_1 = \frac{\sum_{\mathbf{x}, t \in \mathcal{V}} \mathbb{1}\{t = argmax(\hat{\mathbf{p}})\} - a_r}{\sum_{\mathbf{x}, t \in \mathcal{V}} \mathbb{1}\{t = argmax(\mathbf{p})\} - a_r} \tag{8}$$

where a_r is the accuracy from random predictions.

ICE-Eval and C-SHAP-Eval fail all three of our desiderata. (1) A critical piece of both surrogates is to *reconstruct* the model's original embedding, and this reconstruction may be non-linear. Additionally, C-SHAP introduces non-linearity and learnable parameters via $\psi_{\text{C-SHAP}}$. The overly complex surrogates place a large burden on the human interpreter, which is not reflected in prior interpretability and faithfulness evaluations. (2) Neither surrogate depends on concept importances ($A_i's$). Thus, unfaithful A_i's will *not* alter faithfulness scores. (3) Error is measured only on the predicted or groundtruth class, ignoring errors on the remainder of the output distribution.

5. Method: Surrogate Faithfulness (SURF)

Surrogate Faithfulness (SURF) introduces a simple, linear surrogate and two metrics; taken together, they allow SURF to satisfy the desiderata.

Table 2. **Surrogate-based desiderata** We compare prior surrogate-based measures introduced in ICE and C-SHAP to the proposed Surrogate Faithfulness (SURF) measure. According to our desiderata, we require surrogates that are simple (e.g., not reconstruction-based and not requiring additional parameters), incorporate concept importances into the surrogate, and measure errors across all outputs. Prior measures largely fail to meet these desiderata. We mark in **green** and **red** the properties that meet our desiderata. A ✓ or ✗ denotes that the method does or does not have the property.

Property	ICE	C-SHAP	SURF (Ours)
Reconstruction-Based?	✓	✓	✗
Additional Parameters?	✗	✓	✗
Uses Concept Importance?	✗	✗	✓
Errors on All Outputs?	✗	✗	✓

5.1. The SURF Surrogate

The final linear layer $\mathbf{F} = \begin{bmatrix}\mathbf{F}_1 & \cdots & \mathbf{F}_C\end{bmatrix}$ operates on representation $\mathbf{H}$ to give the output $\mathbf{y} \in \mathcal{Y}$. In terms of their components, $\mathbf{H} = \begin{bmatrix}\mathbf{h}_1 & \cdots & \mathbf{h}_{HW}\end{bmatrix}^T$ and $\mathbf{F}_i = \begin{bmatrix}\mathbf{f}_{i,1} & \cdots & \mathbf{f}_{i,HW}\end{bmatrix}^T \in \mathbb{R}^{(HW)\times D}$. Thus, output y_i of ϕ is given by (up to the bias term):

$$\begin{aligned} y_i &= \sum_{j=1}^{HW} \mathbf{h}_j^T \mathbf{f}_{i,j} = \sum_{j=1}^{HW} \mathbf{h_j}^T \frac{\mathbf{f}_{i,j}}{\|\mathbf{f}_{i,j}\|_2} \|\mathbf{f}_{i,j}\|_2 \\ &\triangleq \sum_{j=1}^{HW} \mathbf{h}_j^T \mathbf{v}_{i,j} \alpha_{i,j} \quad \text{where } \|\mathbf{v}_{i,j}\|_2 = 1 \quad \forall\, i, j \end{aligned} \tag{9}$$

where $\alpha_{i,j} = \|\mathbf{f}_{i,j}\|_2$. In words, the model's linear layer projects each embedding $\mathbf{h}_j$ onto a learned, class-specific direction (or *CAV*) $\mathbf{f}_{i,j}$. The projection is scaled by the norm (or *importance*) $\alpha_{i,j}$ of the learned direction. If the model uses global pooling to reduce the final embedding of $\mathbf{x}$ to a single vector, the summation over j above can be omitted.

We define SURF's surrogate s to take the form of (Eq. (9)), replacing $\mathbf{F}$ with the concept representation and A_i's obtained from any U-CBEM. The surrogate is:

$$\hat{\mathbf{y}}_i = s_i(.) = \sum_{j=1}^{HW} \sum_{k=1}^{K} \alpha_{i,k} \mathcal{P}(\mathbf{h}_j; V_i)_k \quad \forall\, i \tag{10}$$

where $\mathcal{P}(.)_k$ denotes the k'th element. Following prior U-CBEMs and evaluation measures, SURF also operates only on the model's final linear layer. Thus, its surrogate fully represents the model's computation, while not introducing any trainable parameters and greatly reducing complexity compared to prior faithfulness surrogates.

5.2. The SURF Metrics

The SURF metrics measure the difference in model and surrogate outputs. For classification models, this means the

Table 3. **Measure-over-measure comparison.** We compare surrogate-based measures across three explanation settings (*Perfect*, *Rand Imp*, and *Full Rand*). We expect evaluations in the *Perfect* setting to give perfect faithfulness scores, and we expect progressively worse faithfulness evaluations as we increase randomness (just importances in *Rand Imp* and fully random in *Full Rand*). We include SURF_{EMD} and SURF_{MAE} along with metrics used by prior works. The SURF metrics behaves as expected, whereas C-SHAP-Eval reports higher faithfulness in the *Full Rand* setting and ICE-Eval reports perfect faithfulness scores in the *Rand Imp* setting.

	Surrogate	Top-1 (%) (↑)	Rank Corr (↑)	Norm L1 (↓)	SURF_{MAE} (↓)	SURF_{EMD} (↓)	Params Learnt (↓)	FLOPs (↓)
Perfect	C-SHAP-Eval (CEL)	9.02	-0.02	1.27	1.97	0.865	1M	205M
	C-SHAP-Eval (L1)	6.13	0.08	2.15	0.54	0.883	1M	205M
	ICE-Eval	100	1.00	0.00	0.00	0.000	0	614K
	SURF (Ours)	100	1.00	0.00	0.00	0.000	0	200
Rand Imp	C-SHAP-Eval (CEL)	9.02	-0.02	1.27	1.97	0.865	1M	205M
	C-SHAP-Eval (L1)	6.13	0.08	2.15	0.54	0.883	1M	205M
	ICE-Eval	100	1.00	0.00	0.00	0.000	0	614K
	SURF (Ours)	97.5	0.13	0.83	2.70	0.862	0	200
Full Rand	C-SHAP-Eval (CEL)	97.6	0.02	181.7	168.2	0.125	1M	205M
	C-SHAP-Eval (L1)	6.1	0.08	3.59	1.721	0.883	1M	205M
	ICE-Eval	3.3	0.00	1.00	3.17	0.882	0	614K
	SURF (Ours)	1.3	0.00	1.01	3.17	0.883	0	200

logits. However, a model's logit-space is unconstrained, varying drastically between models. This makes logit-space metrics difficult to interpret. Furthermore, metrics in the logit-space dilute the importance of the predicted class by aggregating errors over all classes. On the other hand, the probability-space (after softmax) normalizes the logits, allowing for a bounded, interpretable metric. However, low error in the probability-space *does not* always imply low error in the logit-space. Due to the softmax, the probability-space emphasizes the predicted class and diminishes the other classes, so one could achieve a low error by accurately reproducing the predicted class and ignoring the others. Since each space addresses the other's flaws, SURF measures errors in both spaces.

If ϕ is a classification model, then $\mathbf{y}$ denotes class logits. **To measure errors in the logit-space**, we use the mean absolute error between the logits:

$$\text{SURF}_{\text{MAE}} = \frac{1}{|\mathcal{V}|C} \sum_{\mathbf{x} \in \mathcal{V}} \sum_{i=1}^{C} |y_i - \hat{y}_i| \tag{11}$$

To measure errors in the probability-space, we find the Earth Mover's Distance using a constant distance cost.

$$\text{SURF}_{\text{EMD}} = \frac{1}{2|\mathcal{V}|} \sum_{\mathbf{x} \in \mathcal{V}} \sum_{i=1}^{C} |p_i - \hat{p}_i| \tag{12}$$

We choose these specific metrics as they are easily interpretable. A faithful U-CBEM will have explanations with low SURF_{EMD} *and* SURF_{MAE} errors. For regression, we only use SURF_{MAE}.

In contrast to prior surrogate-based measures, SURF meets all of our desiderata. (1) It introduces a simple surrogate with no additional learnable parameters: the output for any class i is a linear combination of the concept representation, weighted by the concept importances (A_i). The surrogate *is not reconstruction-based*; it tries to linearly predict the **next** representation (e.g., class logits), instead of reconstructing embedding $\mathbf{h}$. Thus, only concepts useful for the prediction are needed, which may be a subset of the concepts needed for reconstructing $\mathbf{h}$ from $\mathcal{P}(\mathbf{h}; V)$. (2) SURF's surrogate incorporates A_i, so inaccuracies in A_i impact faithfulness scores. (3) SURF defines two metrics d that measure errors across the entire output space, instead of specific classes, to comprehensively assess faithfulness. These differences are summarized in Tab. 2.

6. Experiments

We first perform a measure-over-measure study with a basic sanity check: do faithfulness scores decrease as we progressively randomize the explanation? Only SURF passes this check. Then, we benchmark prior U-CBEMs across three tasks, revealing that SOTA methods are not faithful. Finally, we leverage SURF to analyze the tradeoff between the number of concepts in an explanation and faithfulness.

6.1. Measure-over-Measure Comparison

The fundamental challenge in evaluating faithfulness measures is that there is no "groundtruth" of faithfulness to compare evaluation results to. To address this, we propose a simple sanity check, which introduces three manufactured settings where the relative faithfulness of explanations is known a priori. Then, we check if the faithfulness measure preserves the relative ordering of scores across settings.

Concretely, we look at faithfulness results for a *perfect* explanation and two *randomized* explanations. To generate the *Perfect* explanation, observe that the most faithful explanation of the model is the model itself. Thus, we set

Table 4. **Benchmarking U-CBEM faithfulness.** We apply SURF to evaluate explanations from prior U-CBEMs in three tasks. Along with the SURF_{MAE} and SURF_{EMD}, we report other metrics to serve as a comparison. We find that prior U-CBEMs are not faithful, as indicated by large errors in the logit and probability space.

	(a) Object classification (ResNet-50)				(b) Multi-attribute prediction (MobileNetV2)		(c) Age (ViT)
U-CBEM	**SURF_{MAE} (↓)**	**SURF_{EMD} (↓)**	**Top-1 (%) (↑)**	**Rank Corr (↑)**	**SURF_{MAE}**	**Attr-Acc (%) (↑)**	**SURF_{MAE}**
CDISCO	3.40	0.932	0.2	0.002	6.77	50.7	32.6
ICE	3.33	0.628	98.9	0.093	5.55	76.1	–
CRAFT	3.19	0.878	90.6	0.068	6.87	19.6	–
C-SHAP	3.28	0.882	6.3	0.005	7.75	51.0	–
MCD	2.60	0.426	99.4	**0.145**	**2.83**	**96.6**	–
HU-MCD	1.97	0.384	**99.7**	0.149	–	–	–
SAE	**1.04**	**0.195**	99.2	0.366	3.16	81.9	**3.67**

$\{A_i, V_i\}_{i=1}^{C}$ according to the weights of the linear classification layer. In our second setting (termed *Rand Imp*), we keep the perfect CAVs fixed and randomly sample the concept importances. Our final setting (termed *Full Rand*) additionally randomly samples the CAVs. In the *Perfect* setting, we expect to obtain no faithfulness error, regardless of the evaluation measure. Then, we expect to obtain increasingly less faithful scores as we progressively increase randomness (from *Rand Imp* to *Full Rand*).

We apply this sanity check on SURF, ICE-Eval, and C-SHAP Eval for an ImageNet-pretrained ResNet-50 [23] finetuned on the Caltech-101 dataset [30]. Along with the SURF metrics, we report Top-1 Accuracy (**Top-1**) defined in Eq. (6) and normalized L1 logit error (**Norm L1**) defined in Eq. (5). As a complementary metric to SURF_{EMD}, we measure Spearman's rank correlation (**Rank Corr**) between surrogate and model outputs. Rank Corr ignores magnitude differences but penalizes fine-grained changes that result in a different prediction ordering. Finally, we measure surrogate complexity by its learnable parameters and FLOPs from concept representation to output. Since C-SHAP-Eval's surrogate was originally trained with cross entropy loss (CEL), we also train it with L1 loss, as this may align better with the SURF metrics. More details are in Appendix Sec. B.

Results are reported in Tab. 3, averaged over 10 seeds. In the *Perfect* setting, both ICE-Eval and SURF report perfect faithfulness, as expected. However, both C-SHAP-Eval variants do not achieve perfect faithfulness across any metric, showing C-SHAP-Eval's limitations. In the *Rand Imp* setting, we observe that C-SHAP-Eval and ICE-Eval do not report differences when compared to the *Perfect* setting; because their surrogates do not use concept importances, using random concept importances do not change results from the *Perfect* setting. Contrast this with SURF, which clearly finds a deterioration in faithfulness across *all* metrics. Finally, in the *Full Rand* setting, C-SHAP-Eval unintuitively reports an improvement in faithfulness compared to the *Perfect* setting. ICE-Eval and SURF report similarly poor faithfulness scores for this setting, as expected.

Next, we examine the metrics. SURF's results in the *Rand Imp* setting highlight a failure case for Top-1. The Top-1 score is strong; however, the other metrics (especially SURF_{EMD}) reveal that the surrogate output has severe errors on the remainder of the output distribution. Norm L1 also comes with issues; notice how C-SHAP-Eval (CEL) reports a lower error compared to C-SHAP-Eval (L1) in the *Perfect* setting, despite C-SHAP-Eval (L1) directly minimizing the logit L1 error. We attribute this inconsistency to Norm L1's emphasis on *only the groundtruth class*; while the L1 error across *all* classes (as measured by SURF_{MAE}) has decreased, the normalized error for the *groundtruth* class increased, giving an incomplete picture of faithfulness.

Of the faithfulness measures, only SURF passes the sanity check while having the lowest surrogate complexity (fewest FLOPs and no learnable parameters). The remainder of the paper uses SURF to study prior U-CBEMs.

6.2. Benchmarking U-CBEMs

We evaluate seven prior U-CBEMs (CDISCO, ICE, CRAFT, C-SHAP, MCD, HU-MCD, SAE) on a varied range of tasks and models to demonstrate the applicability of SURF. Specifically, our tasks are: 1) *Object Classification*, 2) *Multi-Attribute Prediction*, and 3) *Age Regression*. In all tasks, U-CBEMs discover 5 CAVs (or subspaces for MCD and HU-MCD) per output. CAVs and importances are found on the training set; the resulting explanations are evaluated on the test set. Dataset, finetuning, U-CBEM implementation details, and good-faith efforts made to adapt U-CBEMs to new tasks are in Appendix Sec. C.

We evaluate on three tasks. **(1) Object Classification:** We use the Caltech-101-finetuned ResNet-50 and report SURF_{MAE}, SURF_{EMD}, Top-1, and Rank Corr. HU-MCD requires architectural modifications, so we can only evaluate it on this task, which uses a compatible ResNet-50. Results on a VGG and InceptionV3 are in Appendix Sec. D. **(2) Multi-Attribute Prediction:** Given an image, we finetune a MobileNetV2 [40] to predict the presence/absence of at-

tributes on the CelebA dataset [32]. We report SURF_{MAE} and attribute prediction accuracy (Attr-Acc), which is introduced as a task-specific replacement for Top-1. **(3) Age Estimation:** Given a human face, we finetune a ViT [9] to predict their age (in years) on the UTK-Face dataset [54]. Because ViTs have negative activations, U-CBEMs with non-negativity assumptions (i.e., ICE and CRAFT) cannot be used. C-SHAP is omitted as it is only applicable to classification, and MCD is incompatible with ViTs.

Benchmark results are reported in Tab. 4. In the evaluated settings, we observe that **no prior U-CBEM is faithful to the original model**. In the *Object Classification* task, the most faithful U-CBEM is SAE with an SURF_{EMD} of 0.195, denoting significant errors in the probability-space. Only half of the tested U-CBEMs (ICE, MCD, and SAE) perform significantly above random chance in the *Multi-Attribute Prediction* task. SAE exhibits an average error of 3.67 years in the *Age Estimation* task. Note that this evaluation *solely* focuses on faithfulness and does not make any judgment on the interpretability side of the tradeoff.

We believe that U-CBEMs are unfaithful for two reasons: (1) Other than C-SHAP and SAE, all methods discover class-specific CAVs to reconstruct the embedding. Crucially, the class-specific CAVs are found *only* on class-specific images, so using class c_1's CAVs to reconstruct an image embedding of class c_2 will have large reconstruction error (e.g., the U-CBEM is operating out-of-distribution), and therefore, large SURF errors. This issue is exacerbated for CRAFT and ICE, which non-linearly project to the concept space. Though MCD and HU-MCD also find class-specific concepts, they achieve superior faithfulness by using concept subspaces; thus they faithfully explain more of the model with the same number of concepts. (2) Most methods do not satisfy a logit-based completeness criterion, preventing a targeted reproduction of model outputs. C-SHAP, instead, satisfies an accuracy-based completeness criterion, and ICE, MCD, and HU-MCD satisfy the logit-based criterion by including uninterpretable residuals, which cannot be shown in the user explanation.

6.3. Faithfulness vs. Parsimony

The *most important hyperparameter* for any U-CBEM is the number of concepts to be discovered. Intuitively, having more concepts should result in a more complete (i.e., faithful) explanation. However, explanations containing too many concepts may be difficult to interpret. Cognitive psychology studies show that humans can hold a limited number of items in their working memory at a time [8, 34]. Thus, simple, or *parsimonious*, explanations are preferred, which communicate the bulk of the model's computation with a few concepts. However, inadequate faithfulness measures make the faithfulness-parsimony tradeoff hard to analyze, so prior U-CBEMs commonly set this hyperparameter arbitrarily (10 in ICE, 25 in CRAFT & ACE). SURF allows us to intelligently analyze this tradeoff.

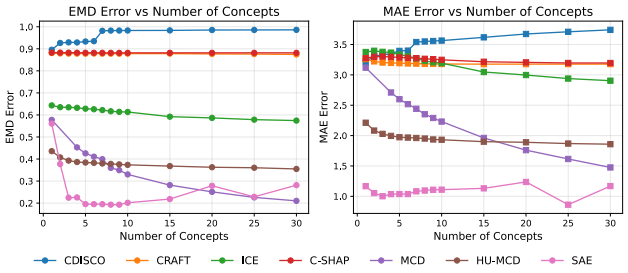

Figure 2. We fit U-CBEMs on Sec. 6.2's *Object Classification* task with an increasing number of concepts and evaluate their faithfulness with SURF. Some U-CBEMs do not improve as the number of concepts increase. U-CBEMs that improve quickly plateau.

Using the *Object Classification* task (Sec. 6.2), we fit U-CBEMs with an increasing number of concepts, each evaluated with the SURF measure and visualized in Fig. 2. Some U-CBEMs (CDISCO, CRAFT, C-SHAP, ICE) either marginally improve or perform worse on one or both metrics as they discover more concepts. SAEs initially improve but then oscillates. Only MCD and HU-MCD uniformly improve as the number of concepts increases, and they exhibit a plateauing effect, marking a natural choice for the number of concepts. Interestingly, HU-MCD is more faithful than MCD only when both discover few concepts; as the number of concepts grows, MCD becomes more faithful. This adds a nuanced touch to the findings from the HU-MCD paper, which claimed superior faithfulness over MCD based on results from a deletion-based proxy.

7. Conclusion

This paper argues that we lack a clear view on how faithful current U-CBEMs really are, largely because of measures that deviate from the formal definition of faithfulness. In accordance with this definition and associated desiderata, we propose SURF as a simple, principled measure for faithfulness. Among other things, we find that **SOTA U-CBEMs do not faithfully explain the final output layer**. While SURF accurately measures U-CBEM faithfulness, we emphasize that it should be paired with an interpretability analysis to judge the U-CBEM's overall quality and value. We urge future work on concept-based explanations to adopt SURF as a standard faithfulness measure and to report SURF scores along their interpretability claims. U-CBEMs are predominantly applied on the final layer, but there is increasing interest in interpreting intermediate layers. SURF, as formulated, does not apply to this case, given the non-linear relationship between the explanation and the model's output. Extending SURF to evaluate U-CBEMs for intermediate layers is important future work.

8. Acknowledgment

The support of the Office of Naval Research under grant N00014-24-1-2169, and IBM-Illinois Discovery Accelerator Institute (IIDAI), and USDA National Institute of Food and Agriculture under grant AFRI 2020-67021-32799/1024178 are gratefully acknowledged. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE 21-46756. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Shubham Kumar also gratefully acknowledges the support of UIUC ECE's Distinguished Research Fellowship and Promise of Excellence Fellowship.

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian J. Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Adv. Neural Inform. Process. Syst.*, 2018.

[2] Zeynep Akata, Ulrike von Luxburg, Uddeshya Upadhyay, and Sebastian Bordt. The manifold hypothesis for gradient-based explanations. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, 2022.

[3] David Alvarez-Melis and T. Jaakkola. Towards robust interpretability with self-explaining neural networks. *Adv. Neural Inform. Process. Syst.*, abs/1806.07538, 2018.

[4] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3319–3327, 2017.

[5] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

[6] Chaofan Chen, Oscar Li, Alina Jade Barnett, Jonathan Su, and Cynthia Rudin. This looks like that: deep learning for interpretable image recognition. In *Adv. Neural Inform. Process. Syst.*, 2019.

[7] Julien Colin, Thomas Fel, Remi Cadene, and Thomas Serre. What i cannot predict, i do not understand: A human-centered evaluation framework for explainability methods. In *Adv. Neural Inform. Process. Syst.*, 2022.

[8] Nelson Cowan. The magical mystery four. *Current Directions in Psychological Science*, 19:51 – 57, 2010.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Int. Conf. Learn. Represent.*, 2021.

[10] Thomas FEL, Remi Cadene, Mathieu Chalvidal, Matthieu Cord, David Vigouroux, and Thomas Serre. Look at the variance! efficient black-box explanations with sobol-based sensitivity analysis. In *Adv. Neural Inform. Process. Syst.*, 2021.

[11] Thomas Fel, Lucas Hervier, David Vigouroux, Antonin Poche, Justin Plakoo, Remi Cadene, Mathieu Chalvidal, Julien Colin, Thibaut Boissin, Louis Bethune, Agustin Picard, Claire Nicodeme, Laurent Gardes, Gregory Flandin, and Thomas Serre. Xplique: A deep learning explainability toolbox. *Workshop on Explainable Artificial Intelligence for Computer Vision (CVPR)*, 2022.

[12] Thomas Fel, Agustin Picard, Louis Béthune, Thibaut Boissin, David Vigouroux, Julien Colin, Rémi Cadène, and Thomas Serre. Craft: Concept recursive activation factorization for explainability. *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2711–2721, 2022.

[13] Thomas Fel, Ekdeep Singh Lubana, Jacob S. Prince, Matthew Kowal, Victor Boutin, Isabel Papadimitriou, Binxu Wang, Martin Wattenberg, Demba E. Ba, and Talia Konkle. Archetypal SAE: Adaptive and stable dictionary learning for concept extraction in large vision models. In *Int. Conf. Machine Learn.*, 2025.

[14] Ruth Fong and Andrea Vedaldi. Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8730–8738, 2018.

[15] Christopher Frye, Damien de Mijolla, Laurence Cowton, Megan Stanley, and Ilya Feige. Shapley-based explainability on the data manifold. *Int. Conf. Learn. Represent.*, abs/2006.01272, 2021.

[16] Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *The Thirteenth International Conference on Learning Representations*, 2025.

[17] Julie Gerlings, Arisa Shollo, and Ioanna D. Constantiou. Reviewing the need for explainable artificial intelligence (xai). In *Hawaii Int. Conf. on System Sciences*, 2020.

[18] Amirata Ghorbani, James Wexler, James Y. Zou, and Been Kim. Towards automatic concept-based explanations. In *Adv. Neural Inform. Process. Syst.*, 2019.

[19] Liv Gorton. The missing curve detectors of inceptionv1: Applying sparse autoencoders to inceptionv1 early vision. In *ICML 2024 Workshop on Mechanistic Interpretability*, 2024.

[20] Mara Graziani, An phi Nguyen, Laura O'Mahony, Henning Müller, and Vincent Andrearczyk. Concept discovery and dataset exploration with singular value decomposition. In *ICLR 2023 Workshop on Pitfalls of limited data and computation for Trustworthy ML*, 2023.

[21] Arne Grobrügge, Niklas Kühl, Gerhard Satzger, and Philipp Spitzer. Towards human-understandable multi-dimensional concept discovery. *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 20018–20027, 2025.

[22] Peter Hase, Harry Xie, and Mohit Bansal. The out-of-distribution problem in explainability and search methods for feature importance explanations. In *Adv. Neural Inform. Process. Syst.*, 2021.

[23] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016.

[24] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie J. Cai, James Wexler, Fernanda B. Viégas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *Int. Conf. Machine Learn.*, 2017.

[25] Eunji Kim, Dahuin Jung, Sangha Park, Siwon Kim, and Sung-Hoon Yoon. Probabilistic concept bottleneck models. In *International Conference on Machine Learning*, 2023.

[26] Diederik P Kingma. Adam: A method for stochastic optimization. In *Int. Conf. Learn. Represent.*, 2015.

[27] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Int. Conf. Comput. Vis.*, pages 4015–4026, 2023.

[28] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning*, pages 5338–5348. PMLR, 2020.

[29] Neehar Kondapaneni, Oisin Mac Aodha, and Pietro Perona. Representational similarity via interpretable visual concepts. In *The Thirteenth International Conference on Learning Representations*, 2025.

[30] Fei-Fei Li, Marco Andreeto, Marc'Aurelio Ranzato, and Pietro Perona. Caltech 101, 2022.

[31] Hyesu Lim, Jinho Choi, Jaegul Choo, and Steffen Schneider. Sparse autoencoders reveal selective remapping of visual concepts during adaptation. In *Int. Conf. Learn. Represent.*, 2025.

[32] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

[33] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Adv. Neural Inform. Process. Syst.*, 2017.

[34] George A. Miller. The magical number seven plus or minus two: some limits on our capacity for processing information. *Psychological review*, 63 2:81–97, 1956.

[35] Giang Nguyen, Daeyoung Kim, and Anh Totti Nguyen. The effectiveness of feature attribution methods and its correlation with automatic evaluation scores. In *Adv. Neural Inform. Process. Syst.*, 2021.

[36] Tuomas Oikarinen, Subhro Das, Lam M Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. In *International Conference on Learning Representations*, 2023.

[37] 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 (BMVC)*, 2018.

[38] Vikram V. Ramaswamy, Sunnie S. Y. Kim, Ruth C. Fong, and Olga Russakovsky. Overlooked factors in concept-based explanations: Dataset choice, concept learnability, and human capability. *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10932–10941, 2023.

[39] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. *Proc. of the 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, 2016.

[40] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4510–4520, 2018.

[41] Leon Sixt, Maximilian Granz, and Tim Landgraf. When explanations lie: Why many modified bp attributions fail. In *Int. Conf. Machine Learn.*, 2020.

[42] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2020.

[43] Muhammad Faaiz Taufiq, Patrick Blöbaum, and Lenon Minorics. Manifold restricted interventional shapley values. In *Int. Conf. on Artificial Intelligence and Statistics*, pages 5079–5106. PMLR, 2023.

[44] Johanna Vielhaben, Stefan Bluecher, and Nils Strodthoff. Multi-dimensional concept discovery (MCD): A unifying framework with completeness guarantees. *Transactions on Machine Learning Research*, 2023.

[45] Minh N Vu, Huy Q Mai, and My T Thai. Emap: Explainable ai with manifold-based perturbations. *arXiv preprint arXiv:2209.08453*, 2022.

[46] Jianyu Wang, Zhishuai Zhang, Cihang Xie, Vittal Premachandran, and Alan Loddon Yuille. Unsupervised learning of object semantic parts from internal states of cnns by population encoding. *arXiv: Learning*, 2015.

[47] Yongjie Wang, Tong Zhang, Xu Guo, and Zhiqi Shen. Gradient based feature attribution in explainable ai: A technical review. *arXiv preprint arXiv:2403.10415*, 2024.

[48] Chih-Kuan Yeh, Been Kim, Sercan Ö. Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. On completeness-aware concept-based explanations in deep neural networks. *Adv. Neural Inform. Process. Syst.*, 2020.

[49] Chih-Kuan Yeh, Kuan-Yun Lee, Frederick Liu, and Pradeep Ravikumar. Threading the needle of on and off-manifold value functions for shapley explanations. In *Int. Conf. on Artificial Intelligence and Statistics*, pages 1485–1502. PMLR, 2022.

[50] H. Peyton Young. Monotonic solutions of cooperative games. *International Journal of Game Theory*, 14:65–72, 1985.

[51] Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. In *The Eleventh International Conference on Learning Representations*, 2023.

[52] Eslam Zaher, Maciej Trzaskowski, Quan Nguyen, and Fred Roosta. Manifold integrated gradients: Riemannian geometry for feature attribution. *Int. Conf. Machine Learn.*, 2024.

[53] Ruihan Zhang, Prashan Madumal, Tim Miller, Krista A. Ehinger, and Benjamin I. P. Rubinstein. Invertible concept-based explanations for cnn models with non-negative concept activation vectors. In *AAAI*, 2020.

[54] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conf. Comput. Vis. Pattern Recog.* IEEE, 2017.

[55] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene cnns. *Int. Conf. Learn. Represent.*, 2015.