Explanation Alignment: Quantifying the Correctness of Model Reasoning At Scale

Hyemin Bang[1], Angie Boggust[1], and Arvind Satyanarayan[1]

MIT CSAIL, Cambridge, MA 02139 USA
{hbang,aboggust}@csail.mit.edu, arvindsatya@mit.edu

Abstract. To improve the reliability of machine learning models, researchers have developed metrics to measure the alignment between model saliency and human explanations. Thus far, however, these saliency-based alignment metrics have been used to conduct descriptive analyses and instance-level evaluations of models and saliency methods. To enable evaluative and comparative assessments of model alignment, we extend these metrics to compute *explanation alignment* — the aggregate agreement between model and human explanations. To compute explanation alignment, we aggregate saliency-based alignment metrics over many model decisions and report the result as a performance metric that quantifies how often model decisions are made for the right reasons. Through experiments on nearly 200 image classification models, multiple saliency methods, and MNIST, CelebA, and ImageNet tasks, we find that explanation alignment automatically identifies spurious correlations, such as model bias, and uncovers behavioral differences between nearly identical models. Further, we characterize the relationship between explanation alignment and model performance, evaluating the factors that impact explanation alignment and how to interpret its results in-practice.

Keywords: explainability · AI alignment · saliency methods

1 Introduction

Saliency methods, or feature attribution methods, are a class of explainable AI techniques used to interpret machine learning model decisions [41, 58, 61] in domains from object classification [10, 11, 50] to radiology [3, 48, 57, 70]. Given an image, saliency methods explain model behavior by estimating the importance of each input feature (e.g., RGB pixel) to the model's decision, which humans compare against their expectations. However, this process is tedious, requiring manual analysis of each dataset instance. Thus, saliency interpretation is often limited to a few manually reviewable instances and can result in missed insights, cherry-picked analysis, and an incomplete understanding of model behavior [8].

To leverage saliency methods without manual inspection, researchers have designed saliency-based alignment metrics that quantify the agreement between model and human explanations [8, 57, 71]. For a given image, these metrics compare the features salient to the model against a ground truth annotation of features important to a human. The result is a quantitative value representing how well the model's decision-making process on that instance aligns with human expectations. Thus far, these metrics have been used for qualitative model evaluations [8] and evaluations of new saliency methods [15, 42, 50, 71].

While saliency-based alignment metrics have proven useful for observing model behavior on particular data instances, they have never been used to provide evaluative or comparative assessments of model alignment. As a result, they are often only invoked during qualitative assessments of model behavior [8] and are excluded from quantitative performance analysis. However, using saliency-based alignment metrics to quantify the human alignment of model behavior across many decisions could provide insight into whether the model consistently makes decisions for the right reasons. Moreover, since even highly accurate models can rely on spurious correlations [11, 44], large-scale application of these metrics could distinguish deployable models from those that are misaligned.

Building on the success of saliency-based alignment metrics, we use them to compute *explanation alignment* — the aggregate agreement between model explanations and human expectations. To do so, we aggregate the results of saliency-based alignment metrics over many model decisions and report the result as a quantitative performance metric alongside traditional task-specific performance metrics. To generate a comprehensive understanding of explanation alignment, we use two common saliency-based aligned metrics — Shared Interest [8] and The Pointing Game [71] — to measure the alignment of the model's entire explanation as well as its most important feature. The result is a quantitative alignment value, that, when used alongside traditional performance metrics, provides a more complete picture of a model's decisions *and* reasoning.

On computer vision classification tasks, explanation alignment uncovers model bias and reveals substantial reasoning differences between highly accurate models[1]. Explanation alignment automatically exposes model biases stemming from synthetic spurious correlations in MNIST [18] and naturally-occurring distributional biases in CelebA [37]. By comparing model and human explanations, it identifies biases without exhaustive validation or prior knowledge of their existence, enabling us to refine the models, remove their bias, and improve their generalizability. In settings with multiple valid human explanations, explanation alignment exposes models' reasoning processes, revealing otherwise imperceptible differences between models with nearly identical performance, architectures, and training set ups. Finally, to support the use of explanation alignment in practice, we characterize its behavior across 195 ImageNet [17] classification models using varying architectures, saliency methods, and tasks.

2 Related Work

AI alignment measures the extent to which machine learning models' behaviors and outcomes are consistent with human expectations [28, 63, 65] and is crucial

[1] Code: `https://github.com/mitvis/explanation_alignment`.

for building reliable models that safely operate in real-world applications [4, 22, 36, 60, 68]. Thus far, research measuring AI alignment has analyzed how closely a model's internal representations match human cognitive processes [31, 46, 52] or its output decisions match human errors [25, 26, 43]. Explanation alignment expands on these alignment by comparing features important to the model against human explanations, providing a complementary quantification that is efficient to compute and human-understandable.

Another line of research has focused on using model explanations to improve the alignment of AI models [53] by designing explanation-based loss terms [23, 55, 56], incorporating explanations into model architectures [35], and using interactive human feedback [24]. These efforts have established a strong foundation for using AI explanations in alignment research. However, rather than influencing model behavior directly, we introduce a scalable approach to assess how well existing models' explanations align with human reasoning across multiple tasks.

To compute explanation alignment, we compute model explanations using saliency methods [3, 9–11, 15, 20, 29, 39, 41, 48, 50, 54, 57, 58, 62, 70]. They offer an advantage by defining model explanations over the input image, making it simple to compare to existing human explanations in the form of image annotations. Further, given the diversity of saliency methods (e.g., gradient-based [29, 62, 64], black-box [12, 50], architecture-specific [10, 14, 15, 59]), we can compute explanation alignment for many modelling tasks.

Given a model explanation and a human explanation, we compute explanation alignment by leveraging existing saliency-based alignment metrics. Saliency-based alignment metrics refer to methods for comparing the overlap between a saliency map and a human explanation [8, 57, 71]. These methods help users efficiently evaluate saliency maps [8] and the localization ability of new saliency methods [9, 42, 57, 71]. While prior work has utilized these metrics to evaluate the effectiveness of explanation methods [2, 30, 45], we re-purpose them to conduct comparative and evaluative analyses of model behavior at scale, across varying datasets and model architectures.

3 Method

To compute explanation alignment, we quantify the alignment between human and model explanations and aggregate it over many decisions. We extract human explanations from ML datasets (Sec. 3.1) and compute model explanations using saliency methods (Sec. 3.2). We use these human and model explanations to compute instance-wise alignment using saliency-based alignment metrics (Sec. 3.3). Then, we aggregate these alignment metrics over an entire dataset and report the result as the model's explanation alignment (Sec. 3.4).

3.1 Representing Human Explanations

To compute the human alignment of model explanations, we need a compatible representation for human explanations on the same decision-making tasks. Since saliency methods operate over the image features, we also define human explanations on the image space. Specifically, we treat human explanations as binary masks, where image features within the mask are considered important to the human decision and features outside the mask are unimportant. While model explanations assign importance to every image feature (i.e., the color channel for each pixel), we define human explanations on the pixel level since, for humans, channel values are visually aggregated into a single perceivable color. Given an image $I \in [0, 255]^{c \times m \times n}$ where c is the number of color channels and m and n are the height and width, the human explanation is defined as $H \in \{0, 1\}^{m \times n}$. For instance, to compute the explanation alignment on MNIST digits in Sec. 4.1, the human explanation includes every pixel in the digit and excludes the black background. This representation allows us to directly compare the model explanation to the human explanation on a feature-by-feature basis.

Often, human explanations exist or can be extracted from existing datasets. For instance, our experiments use the bounding box annotations included with ImageNet [17], which define regions in the image containing the object label. Even when exact explanations do not exist, we can often infer them using available dataset information. For example, our experiments on CelebA [37] smile prediction use existing annotations of the left and right mouth points to define a human explanation region around the mouth. Similarly, since MNIST [18] images are a white foreground digit on a black background, we define the human explanation mask by thresholding the image pixel values and selecting the region corresponding to the digit. In cases where the human explanation can not be extracted or inferred, image segmentation or object localization models could extract object regions as the human explanation, or human annotators could manually annotate regions for high-stakes domains, like medical imaging.

3.2 Generating Model Explanations

To compute the model's explanation alignment, we compute its explanations using saliency methods. Saliency methods compute a continuous score for each input feature, representing its importance to the model's decision. The result is a saliency map $S \in [0, 1]^{c \times m \times n}$ that represents the model's explanation. Since saliency outputs operate over the input space, they are easily comparable to the human explanation. Further, given the variety of saliency methods, we can compute explanation alignment for a variety of models, including black-box or non-gradient-based models. In our experiments, we use Grad-CAM [59] and Vanilla Gradients [61], two prominent saliency methods.

3.3 Measuring Instance-Wise Alignment

We compute the human alignment of a model's decision by comparing its saliency to the human explanation. To do so, we leverage existing saliency-based alignment methods that quantify the relationship between the human and model explanations. While saliency-based alignment methods were originally designed to support qualitative model analysis [8] and evaluate saliency methods [71], we repurpose them to quantify the model's explanation alignment on a given image.

We use two common saliency-based alignment metrics — Shared Interest [8] and The Pointing Game [71]. Shared Interest defines alignment by quantifying the intersection-over-union (IoU) of the model and human explanations. To compute IoU, we must discretize the model's importance scores into regions (see Sec. 4 for details). We then sum the model explanation over the channel dimension to get an importance score per pixel. After discretization and aggregation, we have a model explanation $S' \in \{0,1\}^{m \times n}$ that is in the same format as the human explanation H. We compute IoU [8] for each dataset instance i:

$$\text{IoU}_i = \frac{|H_i \cap S'_i|}{|H_i \cup S'_i|} \tag{1}$$

This value represents the similarity between human and model explanations, ranging from 0 (disjoint) to 1 (identical).

To complement IoU, we also use The Pointing Game metric (PG) [71] to compute model-human alignment. The Pointing Game defines alignment based on whether the model's most important feature is a human-important feature. Unlike IoU, which compares the similarity of the two explanations, The Pointing Game only checks if the model's most salient feature aligns with the human explanation. Following Zhang et al. [71], we compute PG as:

$$\text{PG}_i = \mathbb{1}_{H_{i_{b',c'}}=1} \quad \text{where} \quad (a',b',c') = \underset{(a,b,c)}{\arg\max}\, S_{i_{a,b,c}} \tag{2}$$

The result is either 0 or 1, where 1 indicates the model's most important feature is human-aligned and 0 indicates it is not.

Using both IoU and PG as saliency-based alignment metrics provides complementary insight into the model's behavior — IoU evaluates the entire explanation and PG focuses on specific key features. In cases where the model's explanation relies on a subset of the human important features (i.e., only part of the object), IoU will penalize the alignment for not precisely matching the human, whereas PG accounts for precise explanations. On the other hand, IoU is more robust to noisy saliency maps that mostly focus on the object but assign importance to one-off features. Using both metrics provides a clearer understanding of the model's decision-making processes.

3.4 Computing Explanation Alignment

Finally, to compute explanation alignment, we aggregate a model's instance-wise alignment over an entire dataset. As a result, explanation alignment provides a single quantitative value representing how frequently the model's behavior aligns with human expectations over many decisions. Given a model and dataset of N instances to evaluate explanation alignment on, we create human explanations H (Sec. 3.1) for each dataset instance. Next, given a saliency method, we compute the model's explanations S (Sec. 3.2) for every dataset instance. Finally, given a saliency-based alignment method A (Sec. 3.3), we compute the instance-wise alignment for every dataset instance and average the result.

$$\mathrm{EA}_A = \frac{1}{N}\sum_{i=1}^{N} A_i \tag{3}$$

We compute the explanation alignment using both Shared Interest IoU ($\mathrm{EA}_{\mathrm{IoU}}$) and The Pointing Game ($\mathrm{EA}_{\mathrm{PG}}$). The resulting metrics represent the overall alignment of the model's explanations.

4 Experiments and Results

We demonstrate how explanation alignment can reveal spurious correlations (Sec. 4.1), uncover model bias (Sec. 4.2) and expose differences in model reasoning Sec. 4.3). In Sec. 4.4, we characterize practical considerations of explanation alignment through a study on 195 image classification models.

4.1 Uncovering Spurious Correlations in a Controlled Setting

In ML datasets, spurious correlations — irrelevant features that appear causally related to the outcome — can lead to models that rely on meaningless or biased features and produce unreliable results [72]. However, spurious correlations are difficult to detect using traditional performance metrics because they are artifacts of the data, meaning models that learn them can often achieve equal or better dataset performance than models that rely on human-aligned features. To detect spurious correlations, model developers often rely on manual analysis of model explanations [11] or additional evaluations on new datasets or curated dataset splits that test for a specific spurious correlation [7, 69].

With explanation alignment, we can identify spurious correlations by quantifying the alignment between model explanations and human reasoning across an entire dataset. Unlike other approaches, applying these metrics in aggregate does not require manual analysis of model explanations or a priori knowledge of the types of spurious correlations to test for. Models with high explanation alignment scores consistently rely on human salient features, whereas low alignment scores indicate the model uses features disjoint from human reasoning. In experiments, explanation alignment identifies spurious correlation in otherwise indistinguishably accurate models on MNIST [18, 33] and CelebA [37] tasks.

To demonstrate how explanation alignment detects spurious correlations, we apply it to measure the alignment of two equally performant MNIST models [33]: one using a spurious correlation and one human-aligned. To introduce a spurious correlation, we adopt a method similar to DecoyMNIST [56], augmenting the MNIST dataset by adding a 5×5 colored square in the top-left corner of each image (see Fig. 1). Placing the color outside the digit enables us to use saliency maps to distinguish between the explanations for the digit and the spurious color. Then we create two versions of our augmented MNIST dataset: a `spurious`

Table 1: Explanation alignment helps detect spurious correlations. In an augmented MNIST setting, we train two models: `not-spurious` uses the digit to make its decision and `spurious` that learns a spurious correlation between color box and the digit. Both model's achieve similar accuracy on the test set (`spurious`); however, explanation alignment reveals that the `not-spurious` model relies heavily on the digit features, whereas the `spurious` model primarily relies on the color box correlation. We compute EA_{IoU} and EA_{PG} using Vanilla Gradients [61] explanations thresholded at one standard deviation above the mean and the MNIST digit as the human explanation.

	Test Set Accuracy		**Digit**		**Color Box**	
Model	`not-spurious`	`spurious`	$\mathbf{EA_{IoU}}$	$\mathbf{EA_{PG}}$	$\mathbf{EA_{IoU}}$	$\mathbf{EA_{PG}}$
`not-spurious`	0.981	0.981	**0.294**	**0.699**	0.001	0.000
`spurious`	0.461	0.996	0.087	0.048	**0.222**	**0.937**

dataset where square color correlates with the digit (i.e., 0s have a red square, 1s have an orange square, etc.) and a `not-spurious` dataset with randomized colors and no correlation. The `spurious` dataset simulates a real dataset we might use to train our model that contains both the human-aligned correlation (digit features) as well as spurious correlation (box color). Models trained on the `not-spurious` dataset must learn a correlation between features of the digit to make correct predictions, whereas models trained on the `spurious` dataset can learn to use either features of the digit or the color of the box. For each dataset, we train a simple CNN to classify the digits — a `spurious` model trained on `spurious` dataset and a `not-spurious` model trained on the `not-spurious` dataset (details in Appendix A.2).

First, we confirm that the models have learned their intended feature correlations by evaluating them on the `spurious` and `not-spurious` test splits in Tab. 1. Both models can classify the digits accurately and achieve over 98%, accuracy on the `spurious` dataset. However, when we synthetically remove the spurious correlation (i.e., `not-spurious` dataset), the `spurious` model experiences a 53% drop in accuracy, confirming its reliance on the spurious correlation.

Explanation alignment reveals spurious correlations automatically by testing their reliance on human salient features, unlike accuracy-based methods that require prior knowledge of the correlation to manually curate the `not-spurious` dataset. For each model, we measure its EA_{IoU} and EA_{PG} on the `spurious` dataset, which simulates a real world spurious correlation detection task. We use the MNIST digit as the ground truth region (Tab. 1) and the Vanilla Gradients saliency method [61] thresholded at one standard deviation above the mean (additional details in Appendix A.3). While the `not-spurious` model focuses on the digit in 69.9% of test instances, the `spurious` model does so in only 4.8%. This is shown in Fig. 1, where the `not-spurious` model's explanation focuses on the digit, while the `spurious` model's explanation focuses on the color block.

Explanation alignment reveals misalignment without the need to hypothesize possible spurious correlations in advance; however, when a known spurious correlation exists, explanation alignment can explicitly measure a model's reliance on it. To demonstrate this, we measure the EA_{IoU} and EA_{PG} of both models on

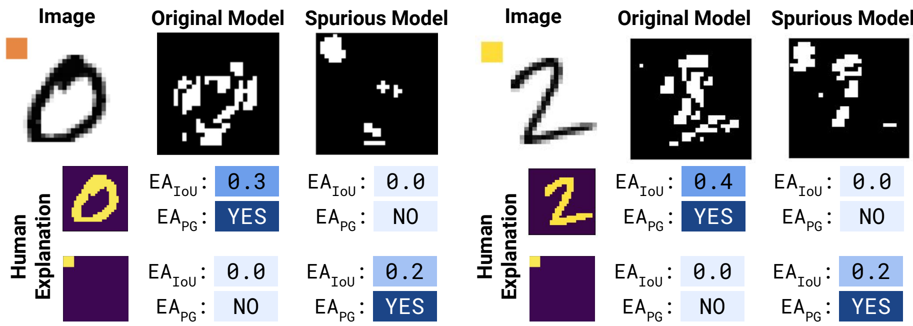

Fig. 1: Explanation alignment measures the human alignment of model decisions. In an MNIST image classification task, it quantifies the `not-spurious` model's reliance on human-aligned features of the digit and a `spurious` model's dependence on the spurious correlation between the color block and the digit. We show Vanilla Gradients [61] explanations thresholded at one standard deviation above the mean.

the `not-spurious` dataset. In this instance, we utilize the color box region as the "human explanation" to quantify how frequently the model depends on the known spurious feature (i.e., color). In Tab. 1, we see that the `not-spurious` model rarely relies on the color block features ($\text{EA}_{\text{IoU}} = 0.001$; $\text{EA}_{\text{PG}} = 0\%$), whereas the `spurious` model's most important feature is in the color box in 93.7% of instances. These results confirm that the `spurious` model's lack of human alignment stems from reliance on the color box spurious correlation, which should be removed or regularized during training.

4.2 Revealing Model Bias in Face Classification Models

Biases can also manifest as spurious correlations, where a model learns to associate a meaningful but irrelevant feature (e.g., race) with its prediction (e.g., job offer) [5, 16]. One way bias can enter a ML pipeline is during dataset collection when one population is overrepresented, causing an unintended correlation between that population and the outcome. Like other spurious correlations, identifying bias is challenging as it often requires a priori knowledge of potential biases and manual test procedures, such as computing accuracy on different test splits that represent potential sources of bias [5, 19].

Using explanation alignment, we can identify model biases without knowing the possible biased features ahead of time. In this experiment, we use EA_{IoU} and EA_{PG} to identify bias in a CelebA smile classification model [37]. In CelebA, there is a preexisting bias between the person's hair and whether they are smiling, where people with `black` hair are more likely to be `smiling` than people with `blond` hair. To replicate this bias, we filter the CelebA dataset to images that have a `black` or `blond` hair attribute and create a `biased` dataset containing a bias towards `black` hair and `smiling`. The dataset contains equal numbers of `black` and `blond` hair images, with a 100:1 bias in the training split and a 10:1 bias in the test split. The `biased` dataset represents the original dataset we

Table 2: Explanation alignment can help detect model bias. In a CelebA smile prediction task, we train an `unbiased` model that sees equal proportions of `black` (👦) and `blond` (👱) hair that are `smiling` (😀)and `not smiling` (😐) and a `biased` model that contains a bias towards `black` hair and `smiling`. Both models achieve similar accuracy on the test set (`biased`). However, explanation alignment reveals that the `biased` model almost never relies on the human-aligned mouth features. We compute $\mathrm{EA}_{\mathrm{IoU}}$ and $\mathrm{EA}_{\mathrm{PG}}$ using GradCAM [59] model explanations thresholded at 0.5 and the mouth annotation as the human explanation.

	Test Set Accuracy							
Model	`biased`	`unbiased`	👦 😀	👦 😐	👱 😀	👱 😐	$\mathbf{EA_{IoU}}$	$\mathbf{EA_{PG}}$
`unbiased`	0.918	0.924	0.908	0.950	0.920	0.912	**0.175**	**0.263**
`biased`	0.936	0.662	0.997	0.150	0.470	0.998	0.005	0.000

would use to train and test our models, where a bias exists that the model may learn. In addition, we create an `unbiased` dataset where `black` and `blond` images are depicted as `smiling` and `not smiling` in equal proportions, representing a curated test set we might use to test bias in our models or train a model that is unbiased. We train two, equally performant models on these datasets, creating a `biased` model and an `unbiased` model. For both models, we finetune an ImageNet [17] pre-trained ResNet50 [27] on the CelebA smile prediction task [37]. Both models achieve over 90% accuracy on our `biased` test set (Tab. 2).

In bias identification task, model developers test models on datasets without potential biases. In this setting, we know a correlation exists between hair color and smiling, so we can evaluate models on an `unbiased` dataset and intersectional data splits. In Tab. 2, we see that while both the `unbiased` and `biased` models achieve similar performance on our original dataset (`biased`), the `biased` model has learned to make predictions using the hair color bias. It achieves near perfect accuracy on our high frequency subgroups, (👦 😀 and 👱 😐); however, it is worse than random guessing on the low frequency subgroups (👱 😀 and 👦 😐).

However, while identifying bias through subgroup accuracy required us to hypothesize the biased variable and create dataset splits, explanation alignment can reveal a bias problem without additional labor. We compute the explanation alignment by comparing the models' explanations against features known to be important to smile prediction (i.e., a person's mouth). We use the CelebA mouth annotation to create a ground truth region and compute Grad-CAM saliency [59] towards the predicted class (Fig. 2) for each instance. We compute explanation alignment across the entire `biased` test set and report the results in Tab. 2. Despite achieving 93.6% accuracy, the `biased` model never relies on the features of the mouth (PG = 0%) to predict whether a person is smiling, suggesting it is using a biased or spurious correlation to make its predictions. On the other hand, the `unbiased` model's most important feature is within the mouth region in 26.3% of instances, suggesting it has learned some causal features between mouths and smiling.

Confirming the numerical results, examples from the dataset in Fig. 2 show that the `biased` model's explanations often contain features related to hair color,

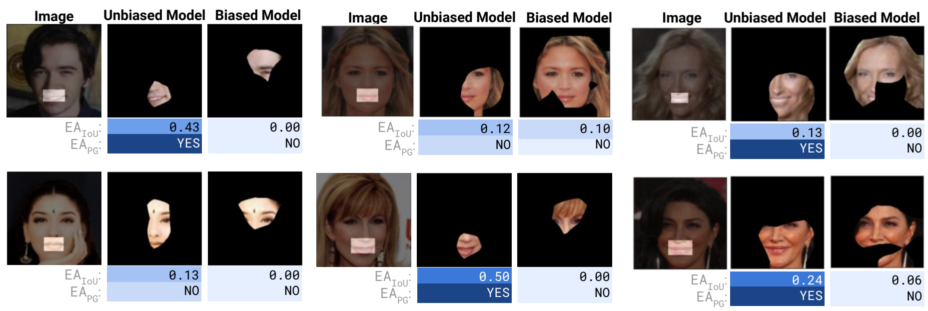

Fig. 2: Explanation alignment can identify model biases. In a CelebA smile prediction task it reveals that the `biased` model has learned the dataset bias between `smiling` and `hair color`, whereas the `unbiased` model has not. We show GradCAM [59] explanations thresholded at 0.5.

like a person's hair or eyebrow, whereas the `unbiased` model primarily focuses on features from the mouth. However, the `unbiased` model's explanation also contains parts of a person's cheeks and eyes, suggesting where it may be looking in the other 73.7% of instances. While cheeks and eyes are not mechanically related to a smile the way a mouth is, they are equally causal and as humans we can determine if someone is smiling by looking at the rest of their face. If we would like to expand the notion of a smiling ground truth in future iterations of analysis, we could include a person's entire face in the ground truth region. Or, if mouth features are particularly important to the task, such as emotion prediction in people with facial paralysis, then we may want to further improve our model to enforce mouth features as the only ones that are causal.

4.3 Exposing Behavioral Differences in Highly Accurate Models

We want to ensure that our model uses human-aligned features to make its decisions; however, there are often many possible correlations a model can learn that align with human reasoning. For instance, as humans, we can detect that someone is smiling by looking at their mouth, eyes, cheeks or a combination of those features. Evaluating a model's alignment against all possible human explanations tells us more about our model and ensures that we do not inadvertently penalize it for relying on different but equally human-aligned features.

To represent explanation alignment's ability to provide a comprehensive overview of model behavior, we apply it to a setting with multiple human explanations. Following our experimental set up in Sec. 4.2, we train three model replicates on the `unbiased` CelebA dataset [37]. We compute model explanations using Grad-CAM [59] towards the model's predicted class and threshold it at 0.5. However, this time we compute the alignment metrics with respect to five possible explanations from CelebAMask-HQ [34] — the person's *hair*, *eyes*, *mouth*, *nose*, and *skin*. We report the $\mathrm{EA}_{\mathrm{IoU}}$ and $\mathrm{EA}_{\mathrm{PG}}$ on the CelebA test set for each human explanation and report the results in Tab. 3.

Table 3: Explanation alignment uncovers model behavior differences obscured by accuracy. On three CelebA smile prediction models with similar accuracy, explanation alignment reveals that `A` relies on the mouth, `B` focuses on the nose, and `C` uses both. We compute $\mathrm{EA}_{\mathrm{IoU}}$ and $\mathrm{EA}_{\mathrm{PG}}$ using GradCAM [59] thresholded at 0.5 and compare to multiple face region annotations from CelebAMask-HQ [34].

		$\mathrm{EA}_{\mathrm{IoU}}$					$\mathrm{EA}_{\mathrm{PG}}$				
Model	**Accuracy**	**Hair**	**Eye**	**Mouth**	**Nose**	**Skin**	**Hair**	**Eye**	**Mouth**	**Nose**	**Skin**
A	0.93	0.0024	0.0000	**0.3120**	0.0802	0.1310	0.0009	0.0000	**0.8402**	0.0197	0.9937
B	0.92	0.0007	0.0607	0.0215	**0.2370**	0.2360	0.0015	0.0244	0.0018	**0.5434**	0.9997
C	0.93	0.0058	0.0332	0.1510	0.1720	**0.2010**	0.0020	0.0125	**0.3896**	0.2978	0.9994

While the models are indistinguishable by accuracy (each achieving 92–93%), explanation alignment reveals that they use significantly different facial features to predict whether a person is smiling. While `model A` relies on features of the mouth, `model B` almost never relies on the features of the mouth, instead focusing more on the person's nose, and `model C` uses both features of the mouth and nose. These findings are further supported by visual examples (Fig. 3), where we see that `model A`'s saliency map highlights the mouth, `model B's` focuses on the nose, and `model C` relies on the majority of the face. While all three models have high alignment with the skin and low alignment with the eyes, this is likely due to the size of those ground truth features. For instance, given the skin is a superset of the regions, $\mathrm{EA}_{\mathrm{PG}}$ will count alignment with the skin region even if the feature was within a more specific region like mouth.

By highlighting the differences in the model's behavior, alignment metrics can help us make more informed decisions between the models. If we were applying this model in a setting where we expect people to be wearing masks, then we may want to choose a model that relies on facial features besides the mouth. It can also provide an opportunity to assemble an ensemble of models, each focusing on a unique valid ground truth feature, resulting in a more effective and resilient model against facial obfuscations.

4.4 Characterizing Explanation Alignment

While our previous experiments demonstrate how explanation alignment can be used to evaluate and compare model reasoning, this experiment focuses on characterizing the factors that influence explanation alignment. In particular, we compute explanation alignment on 195 image classification models, evaluating how choice of saliency method, model architecture, explanation alignment metric, and evaluative task influence the results. In doing so, we identify important considerations when interpreting explanation alignment in practice.

To analyze explanation alignment at scale, we compute the explanation alignment of 195 TIMM[2] ImageNet [17] classification models with varying architectures (e.g., CNNs, Transformers), sizes (ranging from 1–200 million parameters), and performance levels (>25% accuracy range). We calculate $\mathrm{EA}_{\mathrm{IoU}}$ and $\mathrm{EA}_{\mathrm{PG}}$ for each model using the ImageNet validation set and its bounding box explanations [17]. We use Vanilla Gradients [61] thresholded at one standard deviation

[2] `https://timm.fast.ai/`

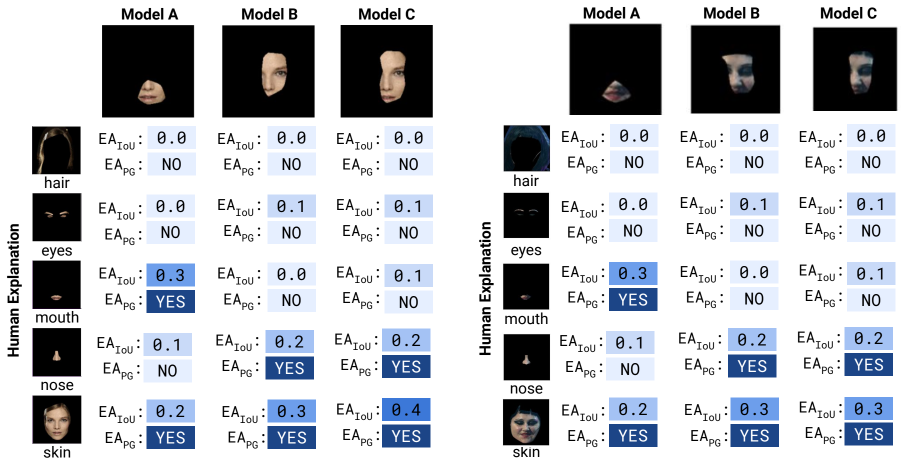

Fig. 3: Measuring explanation alignment using multiple human explanations shows that similarly accurate models use different underlying facial features. We compare GradCAM [59] explanations thresholded at 0.5 against 5 facial annotations [34].

above the mean across all models and Grad-CAM [59] thresholded at 0.5 on the 150 models containing convolutional layers. We compare the explanation alignment against the model's accuracy on the ImageNet validation set and its transfer learning performance on CIFAR-100 [32]. We perform 1-shot transfer learning via a logistic regression that takes in the models' penultimate layer embeddings and predicts the CIFAR-100 labels. We report our results in Fig. 4.

Explanation alignment differs based on model architecture. Across settings, the range of explanation alignment values differ based on model architecture. Transformers [67] have lower explanation alignment scores than CNNs [21]. For a direct comparison, we use the same saliency method (Vanilla Gradients) to compute explanation alignment for all models, regardless of architecture. However, due to the patch-based tokenization procedure of Image Transformers [47], Vanilla Gradients often highlights rectangular image regions as opposed to continuous saliency distributions we see in CNNs (Fig. A2). Since the model explanations have a different distribution for Transformers than CNNs, the explanation alignment scores are not directly comparable between model architectures.

Explanation alignment is sensitive to the underlying saliency method. Differences in saliency methods result in differences in explanation alignment values. Computing explanation alignment with Vanilla Gradients results in $\mathrm{EA}_{\mathrm{IoU}}$ scores in the range 0–0.2, whereas with Grad-CAM scores range from 0.1–0.6. Comparing models with explanation alignment should use the same saliency method to prevent confounding differences in alignment due to the saliency method with differences in alignment due to the model's behavior. Further, it is important to select a saliency method relevant to the model and task. As we saw in the previous take-away, Vanilla Gradients produces patch-based explanations for Transformers that skews the range of explanation alignment values. This signals the

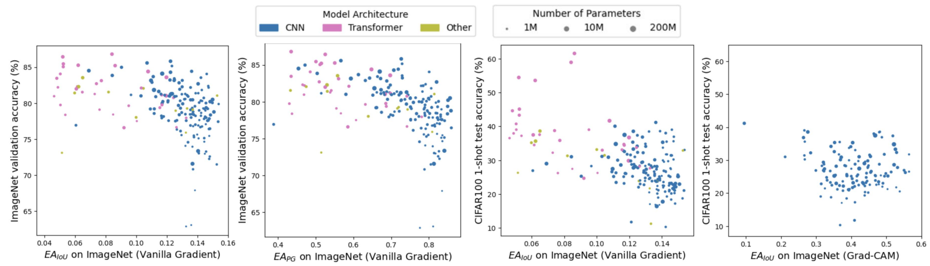

Fig. 4: We compare the explanation alignment of 195 models across saliency methods (Vanilla Gradients and Grad-CAM), explanation alignment metrics ($\mathrm{EA}_{\mathrm{IoU}}$ and $\mathrm{EA}_{\mathrm{PG}}$), and tasks (ImageNet classification and CIFAR-100 transfer learning). In each plot, color indicates architecture type and size encodes number of model parameters.

importance of computing explanation alignment with a task-appropriate saliency method, such as a method designed specifically for Transformers [10, 13, 15].

EA_{IoU} and EA_{PG} are interchangeable for relative model comparisons. Both explanation alignment measures ($\mathrm{EA}_{\mathrm{IoU}}$ and $\mathrm{EA}_{\mathrm{PG}}$) result in similar model rankings (Spearman's rank correlation coefficient $\rho = 0.902$, $p < 0.001$). Unlike explanation alignment's sensitivity to saliency method, the relative explanation alignment between models does not change substantially based on the underlying saliency-based alignment metric. While the absolute value $\mathrm{EA}_{\mathrm{IoU}}$ and $\mathrm{EA}_{\mathrm{PG}}$ measure a specific aspect of the model's alignment, they can be interchanged when measuring the relative alignment difference between models.

Accurate models can have low explanation alignment and vice versa. Confirming our prior experimental results, we find that highly accurate models can have low explanation alignment, since learning misaligned correlations can still result in correct decisions within a dataset. However, we also find that aligned models can have low task accuracy. One hypothesis for this is that spurious correlations can still occur within the object of interest. For instance, even a model that relies on pixels of the apple to predict `apple` could do so in unaligned ways, such as only looking at color due to a bias that all apples are red. This signals the importance of measuring explanation alignment alongside accuracy to ensure models are both correct and human-aligned.

Explanation alignment does not predict ImageNet to CIFAR-100 transferability. We would expect that models with greater explanation alignment would be better able to transfer to new domains because they have learned the same reasoning processes humans use to generalize between tasks. However, we do not find a correlation between the explanation alignment of an ImageNet model and its 1-shot learning performance on CIFAR-100. On one hand, this could suggest that while explanation alignment can reveal differences in model behavior (e.g., bias), it is not predictive of model generalizability. On the other hand, it could also be the case that explanation alignment on ImageNet is not necessary to generalize to the simple and small CIFAR-100 images which typically only contain a single object. Future work may consider larger-scale analysis and benchmarks to measure the relationship between different types of alignment and model generalization.

5 Conclusion and Discussion

We present explanation alignment, a method to quantify the agreement between model explanations with human reasoning. Using saliency methods [59, 61], we generate model explanations and human-defined ground truth from existing datasets [17, 34]. We aggregate the results of saliency-based alignment metrics [8, 71] over many data instances to quantify the model's alignment over many decisions. Through experiments on ImageNet [17], MNIST [18], and CelebA [37] datasets, we demonstrate that explanation alignment can reveal biases and behavioral differences between models with similar performance metrics. Our findings highlight the importance of aligning model explanations with human expectations to improve transparency, trustworthiness, and performance.

To compute explanation alignment, we leverage saliency methods to generate model explanations. Saliency methods are valuable to our computation because they quantify the importance of each image feature, making them directly comparable to human explanations that are also defined on the image pixels. However, research has demonstrated that saliency methods can generate inconsistent explanations, highlight irrelevant features, and produce misleading explanations [1, 6]. While explanation alignment accounts for one-off saliency mistakes by aggregating over many model decisions, future work could explore more robust saliency methods or ways to compute explanation alignment without saliency methods, such as through concept-based or counterfacutal explanations.

Relatedly, explanation alignment requires human explanations in the form of annotations of important image regions. In our experiments, we found that many research datasets have associated human explanations [17, 34] or that explanations can be derived from existing metadata [18, 37]. However, in settings where human explanations can not be derived, image segmentation [40] or object localization [66] models could identify important image regions as human-like explanations. Further, defining a human explanation is inherently subjective, and, as we saw in Sec. 4.3, there may exist many possible human explanations for a given decision. As a result, future work could explore alternate representations for human explanations, such as human studies to understand how humans select and combine features to make their decisions.

Successful use of explanation alignment metrics suggests incorporating them into model training to enforce explanation alignment during development. While traditional training procedures emphasize correctness, explanation alignment provides an opportunity to update model parameters based on their reasoning processes and alignment with human reasoning. Incorporating these metrics could enable developers to enhance models beyond accuracy benchmarks, emphasizing the importance of how the model made its decision. Such models would be not only trustworthy and reliable but could improve robustness to new and unseen data.

Bibliography

[1] Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I.J., Hardt, M., Kim, B.: Sanity checks for saliency maps. In: Bengio, S., Wallach, H.M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada, pp. 9525–9536 (2018), URL `https://proceedings.neurips.cc/paper/2018/hash/294a8ed24b1ad22ec2e7efea049b8737-Abstract.html` 14

[2] Adebayo, J., Muelly, M., Abelson, H., Kim, B.: Post hoc explanations may be ineffective for detecting unknown spurious correlation. In: The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022, OpenReview.net (2022), URL `https://openreview.net/forum?id=xNOVfCCvDpM` 3

[3] Aggarwal, M., Arun, N.T., Gupta, S., Vaswani, A., Chen, B., Li, M.D., Chang, K., Patel, J.B., Höbel, K., Gidwani, M., Kalpathy-Cramer, J., Singh, P.: Towards trainable saliency maps in medical imaging. CoRR **abs/2011.07482** (2020), URL `https://arxiv.org/abs/2011.07482` 1, 3

[4] Amodei, D., Olah, C., Steinhardt, J., Christiano, P.F., Schulman, J., Mané, D.: Concrete problems in AI safety. CoRR **abs/1606.06565** (2016), URL `http://arxiv.org/abs/1606.06565` 3

[5] Angwin, J., Larson, J., Mattu, S., Kirchner, L.: Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. (2016), URL `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`, accessed: 2024-08-20 8

[6] Arun, N.T., Gaw, N., Singh, P., Chang, K., Aggarwal, M., Chen, B., Hoebel, K., Gupta, S., Patel, J.B., Gidwani, M., Adebayo, J., Li, M.D., Kalpathy-Cramer, J.: Assessing the (un)trustworthiness of saliency maps for localizing abnormalities in medical imaging. CoRR **abs/2008.02766** (2020), URL `https://arxiv.org/abs/2008.02766` 14

[7] Bellamy, R.K.E., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K.N., Richards, J.T., Saha, D., Sattigeri, P., Singh, M., Varshney, K.R., Zhang, Y.: AI fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM J. Res. Dev. **63**(4/5), 4:1–4:15 (2019), `https://doi.org/10.1147/JRD.2019.2942287`, URL `https://doi.org/10.1147/JRD.2019.2942287` 6

[8] Boggust, A., Hoover, B., Satyanarayan, A., Strobelt, H.: Shared interest: Measuring human-ai alignment to identify recurring patterns in model behavior. In: Barbosa, S.D.J., Lampe, C., Appert, C., Shamma, D.A., Drucker, S.M., Williamson, J.R., Yatani, K. (eds.) CHI '22: CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 29 April 2022

- 5 May 2022, pp. 10:1–10:17, ACM (2022), https://doi.org/10.1145/3491102.3501965, URL https://doi.org/10.1145/3491102.3501965 1, 2, 3, 5, 14, 26

[9] Boggust, A.W., Suresh, H., Strobelt, H., Guttag, J.V., Satyanarayan, A.: Saliency cards: A framework to characterize and compare saliency methods. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2023, Chicago, IL, USA, June 12-15, 2023, pp. 285–296, ACM (2023), https://doi.org/10.1145/3593013.3593997, URL https://doi.org/10.1145/3593013.3593997 3

[10] Bousselham, W., Boggust, A.W., Chaybouti, S., Strobelt, H., Kuehne, H.: Legrad: An explainability method for vision transformers via feature formation sensitivity. CoRR **abs/2404.03214** (2024), https://doi.org/10.48550/ARXIV.2404.03214, URL https://doi.org/10.48550/arXiv.2404.03214 1, 3, 13

[11] Carter, B., Jain, S., Mueller, J., Gifford, D.: Overinterpretation reveals image classification model pathologies. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 15395–15407 (2021) 1, 2, 3, 6

[12] Carter, B., Mueller, J., Jain, S., Gifford, D.K.: What made you do this? understanding black-box decisions with sufficient input subsets. In: Chaudhuri, K., Sugiyama, M. (eds.) The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan, Proceedings of Machine Learning Research, vol. 89, pp. 567–576, PMLR (2019), URL http://proceedings.mlr.press/v89/carter19a.html 3

[13] Chang, C., Creager, E., Goldenberg, A., Duvenaud, D.: Explaining image classifiers by counterfactual generation. In: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019, OpenReview.net (2019), URL https://openreview.net/forum?id=B1MXz20cYQ 13

[14] Chefer, H., Gur, S., Wolf, L.: Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021, pp. 387–396, IEEE (2021), https://doi.org/10.1109/ICCV48922.2021.00045, URL https://doi.org/10.1109/ICCV48922.2021.00045 3

[15] Chefer, H., Gur, S., Wolf, L.: Transformer interpretability beyond attention visualization. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021, pp. 782–791, Computer Vision Foundation / IEEE (2021), https://doi.org/10.1109/CVPR46437.2021.00084, URL https://openaccess.thecvf.com/content/CVPR2021/html/Chefer_Transformer_Interpretability_Beyond_Attention_Visualization_CVPR_2021_paper.html 2, 3, 13

[16] Dastin, J.: Amazon scraps secret ai recruiting tool that showed bias against women (2018), URL https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/, accessed: 2024-08-20 8

[17] Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA, pp. 248–255, IEEE Computer Society (2009), `https://doi.org/10.1109/CVPR.2009.5206848`, URL `https://doi.org/10.1109/CVPR.2009.5206848` 2, 4, 9, 11, 14, 25, 26

[18] Deng, L.: The MNIST database of handwritten digit images for machine learning research [best of the web]. IEEE Signal Process. Mag. **29**(6), 141–142 (2012), `https://doi.org/10.1109/MSP.2012.2211477`, URL `https://doi.org/10.1109/MSP.2012.2211477` 2, 4, 6, 14, 25

[19] Dooley, S., Downing, R., Wei, G.Z., Shankar, N., Thymes, B., Thorkelsdottir, G., Kurtz-Miott, T., Mattson, R., Obiwumi, O., Cherepanova, V., Goldblum, M., Dickerson, J.P., Goldstein, T.: Comparing human and machine bias in face recognition. CoRR **abs/2110.08396** (2021), URL `https://arxiv.org/abs/2110.08396` 8

[20] Fong, R.C., Vedaldi, A.: Interpretable explanations of black boxes by meaningful perturbation. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, pp. 3449–3457, IEEE Computer Society (2017), `https://doi.org/10.1109/ICCV.2017.371`, URL `https://doi.org/10.1109/ICCV.2017.371` 3

[21] Fukushima, K.: Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological cybernetics **36**(4), 193–202 (1980) 12

[22] Gabriel, I.: Artificial intelligence, values, and alignment. Minds Mach. **30**(3), 411–437 (2020), `https://doi.org/10.1007/S11023-020-09539-2`, URL `https://doi.org/10.1007/s11023-020-09539-2` 3

[23] Gao, Y., Sun, T.S., Bai, G., Gu, S., Hong, S.R., Zhao, L.: RES: A robust framework for guiding visual explanation. In: Zhang, A., Rangwala, H. (eds.) KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022, pp. 432–442, ACM (2022), `https://doi.org/10.1145/3534678.3539419`, URL `https://doi.org/10.1145/3534678.3539419` 3

[24] Gao, Y., Sun, T.S., Zhao, L., Hong, S.R.: Aligning eyes between humans and deep neural network through interactive attention alignment. Proc. ACM Hum. Comput. Interact. **6**(CSCW2), 1–28 (2022), `https://doi.org/10.1145/3555590`, URL `https://doi.org/10.1145/3555590` 3

[25] Geirhos, R., Meding, K., Wichmann, F.A.: Beyond accuracy: quantifying trial-by-trial behaviour of cnns and humans by measuring error consistency. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual (2020), URL `https://proceedings.neurips.cc/paper/2020/hash/9f6992966d4c363ea0162a056cb45fe5-Abstract.html` 3

[26] Geirhos, R., Temme, C.R.M., Rauber, J., Schütt, H.H., Bethge, M., Wichmann, F.A.: Generalisation in humans and deep neural networks. In: Bengio, S., Wallach, H.M., Larochelle, H., Grauman, K., Cesa-Bianchi,

N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada, pp. 7549–7561 (2018), URL https://proceedings.neurips.cc/paper/2018/hash/0937fb5864ed06ffb59ae5f9b5ed67a9-Abstract.html 3

[27] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, pp. 770–778, IEEE Computer Society (2016), https://doi.org/10.1109/CVPR.2016.90, URL https://doi.org/10.1109/CVPR.2016.90 9, 26

[28] Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., Duan, Y., He, Z., Zhou, J., Zhang, Z., Zeng, F., Ng, K.Y., Dai, J., Pan, X., O'Gara, A., Lei, Y., Xu, H., Tse, B., Fu, J., McAleer, S., Yang, Y., Wang, Y., Zhu, S., Guo, Y., Gao, W.: AI alignment: A comprehensive survey. CoRR **abs/2310.19852** (2023), https://doi.org/10.48550/ARXIV.2310.19852, URL https://doi.org/10.48550/arXiv.2310.19852 2

[29] Kapishnikov, A., Bolukbasi, T., Viégas, F.B., Terry, M.: XRAI: better attributions through regions. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, pp. 4947–4956, IEEE (2019), https://doi.org/10.1109/ICCV.2019.00505, URL https://doi.org/10.1109/ICCV.2019.00505 3

[30] Kim, S.S.Y., Meister, N., Ramaswamy, V.V., Fong, R., Russakovsky, O.: HIVE: evaluating the human interpretability of visual explanations. In: Avidan, S., Brostow, G.J., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XII, Lecture Notes in Computer Science, vol. 13672, pp. 280–298, Springer (2022), https://doi.org/10.1007/978-3-031-19775-8_17, URL https://doi.org/10.1007/978-3-031-19775-8_17 3

[31] Kornblith, S., Norouzi, M., Lee, H., Hinton, G.E.: Similarity of neural network representations revisited. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, Proceedings of Machine Learning Research, vol. 97, pp. 3519–3529, PMLR (2019), URL http://proceedings.mlr.press/v97/kornblith19a.html 3

[32] Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009) 12

[33] LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998), https://doi.org/10.1109/5.726791, URL https://doi.org/10.1109/5.726791 6

[34] Lee, C.H., Liu, Z., Wu, L., Luo, P.: Maskgan: Towards diverse and interactive facial image manipulation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 10, 11, 12, 14

[35] Li, K., Wu, Z., Peng, K., Ernst, J., Fu, Y.: Tell me where to look: Guided attention inference network. In: 2018 IEEE Conference on Computer Vi-

sion and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, pp. 9215–9223, Computer Vision Foundation / IEEE Computer Society (2018), https://doi.org/10.1109/CVPR.2018.00960, URL http://openaccess.thecvf.com/content_cvpr_2018/html/Li_Tell_Me_Where_CVPR_2018_paper.html 3

[36] Linardatos, P., Papastefanopoulos, V., Kotsiantis, S.: Explainable AI: A review of machine learning interpretability methods. Entropy **23**(1), 18 (2021), https://doi.org/10.3390/E23010018, URL https://doi.org/10.3390/e23010018 3

[37] Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015, pp. 3730–3738, IEEE Computer Society (2015), https://doi.org/10.1109/ICCV.2015.425, URL https://doi.org/10.1109/ICCV.2015.425 2, 4, 6, 8, 9, 10, 14, 25

[38] Liu, Z., Mao, H., Wu, C., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022, pp. 11966–11976, IEEE (2022), https://doi.org/10.1109/CVPR52688.2022.01167, URL https://doi.org/10.1109/CVPR52688.2022.01167 26

[39] Lundberg, S.M., Lee, S.: A unified approach to interpreting model predictions. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pp. 4765–4774 (2017), URL https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html 3

[40] Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., Terzopoulos, D.: Image segmentation using deep learning: A survey. CoRR **abs/2001.05566** (2020), URL https://arxiv.org/abs/2001.05566 14

[41] Molnar, C.: Interpretable Machine Learning. 2 edn. (2022), URL https://christophm.github.io/interpretable-ml-book 1, 3

[42] Morrison, K., Mehra, A., Perer, A.: Shared interest...sometimes: Understanding the alignment between human perception, vision architectures, and saliency map techniques. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023 - Workshops, Vancouver, BC, Canada, June 17-24, 2023, pp. 3776–3781, IEEE (2023), https://doi.org/10.1109/CVPRW59228.2023.00391, URL https://doi.org/10.1109/CVPRW59228.2023.00391 2, 3

[43] Muttenthaler, L., Dippel, J., Linhardt, L., Vandermeulen, R.A., Kornblith, S.: Human alignment of neural network representations. In: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023, OpenReview.net (2023), URL https://openreview.net/forum?id=ReDQ1OUQR0X 3

[44] Narla, A., Kuprel, B., Sarin, K., Novoa, R., Ko, J.: Automated classification of skin lesions: from pixels to practice. Journal of Investigative Dermatology **138**(10), 2108–2110 (2018) 2

[45] Nguyen, G., Kim, D., Nguyen, A.: The effectiveness of feature attribution methods and its correlation with automatic evaluation scores. In: Ranzato, M., Beygelzimer, A., Dauphin, Y.N., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual, pp. 26422–26436 (2021), URL https://proceedings.neurips.cc/paper/2021/hash/de043a5e421240eb846da8effe472ff1-Abstract.html 3

[46] Nguyen, T., Raghu, M., Kornblith, S.: Do wide and deep networks learn the same things? uncovering how neural network representations vary with width and depth. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021, OpenReview.net (2021), URL https://openreview.net/forum?id=KJNcAkY8tY4 3

[47] Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N., Ku, A., Tran, D.: Image transformer. In: Dy, J.G., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018, Proceedings of Machine Learning Research, vol. 80, pp. 4052–4061, PMLR (2018), URL http://proceedings.mlr.press/v80/parmar18a.html 12

[48] Pasa, F., Golkov, V., Pfeiffer, F., Cremers, D., Pfeiffer, D.: Efficient deep network architectures for fast chest x-ray tuberculosis screening and visualization. Scientific reports **9**(1), 6268 (2019) 1, 3

[49] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E.Z., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, pp. 8024–8035 (2019), URL https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html 26

[50] Petsiuk, V., Das, A., Saenko, K.: RISE: randomized input sampling for explanation of black-box models. In: British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018, p. 151, BMVA Press (2018), URL http://bmvc2018.org/contents/papers/1064.pdf 1, 2, 3

[51] Radosavovic, I., Kosaraju, R.P., Girshick, R.B., He, K., Dollár, P.: Designing network design spaces. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, pp. 10425–10433, Computer Vision Foundation / IEEE (2020), https://doi.org/10.1109/CVPR42600.2020.01044, URL https://openaccess.thecvf.com/content_CVPR_2020/html/Radosavovic_Designing_Network_Design_Spaces_CVPR_2020_paper.html 26

[52] Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., Dosovitskiy, A.: Do vision transformers see like convolutional neural networks? In: Ranzato, M., Beygelzimer, A., Dauphin, Y.N., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual, pp. 12116–12128 (2021), URL https://proceedings.neurips.cc/paper/2021/hash/652cf38361a209088302ba2b8b7f51e0-Abstract.html 3

[53] Rao, S., Böhle, M., Parchami-Araghi, A., Schiele, B.: Studying how to efficiently and effectively guide models with explanations. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023, pp. 1922–1933, IEEE (2023), https://doi.org/10.1109/ICCV51070.2023.00184, URL https://doi.org/10.1109/ICCV51070.2023.00184 3

[54] Ribeiro, M.T., Singh, S., Guestrin, C.: "why should I trust you?": Explaining the predictions of any classifier. In: Krishnapuram, B., Shah, M., Smola, A.J., Aggarwal, C.C., Shen, D., Rastogi, R. (eds.) Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016, pp. 1135–1144, ACM (2016), https://doi.org/10.1145/2939672.2939778, URL https://doi.org/10.1145/2939672.2939778 3

[55] Rieger, L., Singh, C., Murdoch, W.J., Yu, B.: Interpretations are useful: Penalizing explanations to align neural networks with prior knowledge. In: Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event, Proceedings of Machine Learning Research, vol. 119, pp. 8116–8126, PMLR (2020), URL http://proceedings.mlr.press/v119/rieger20a.html 3

[56] Ross, A.S., Hughes, M.C., Doshi-Velez, F.: Right for the right reasons: Training differentiable models by constraining their explanations. In: Sierra, C. (ed.) Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017, pp. 2662–2670, ijcai.org (2017), https://doi.org/10.24963/IJCAI.2017/371, URL https://doi.org/10.24963/ijcai.2017/371 3, 6

[57] Saporta, A., Gui, X., Agrawal, A., Pareek, A., Truong, S.Q.H., Nguyen, C.D.T., Ngo, V.D., Seekins, J., Blankenberg, F.G., Ng, A.Y., Lungren, M.P., Rajpurkar, P.: Benchmarking saliency methods for chest x-ray interpretation. Nat. Mac. Intell. **4**(10), 867–878 (2022), https://doi.org/10.1038/S42256-022-00536-X, URL https://doi.org/10.1038/s42256-022-00536-x 1, 3

[58] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. vol. 128, pp. 336–359 (2020), https://doi.org/10.1007/S11263-019-01228-7, URL https://doi.org/10.1007/s11263-019-01228-7 1, 3

[59] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based

localization. Int. J. Comput. Vis. **128**(2), 336–359 (2020), https://doi.org/10.1007/S11263-019-01228-7, URL https://doi.org/10.1007/s11263-019-01228-7 3, 4, 9, 10, 11, 12, 14, 26
[60] Shahriari, K., Shahriari, M.: IEEE standard review - ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems. In: IEEE Canada International Humanitarian Technology Conference, IHTC 2017, Toronto, ON, Canada, July 21-22, 2017, pp. 197–201, IEEE (2017), https://doi.org/10.1109/IHTC.2017.8058187, URL https://doi.org/10.1109/IHTC.2017.8058187 3
[61] Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. In: Bengio, Y., LeCun, Y. (eds.) 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings (2014), URL http://arxiv.org/abs/1312.6034 1, 4, 7, 8, 11, 14, 26
[62] Smilkov, D., Thorat, N., Kim, B., Viégas, F.B., Wattenberg, M.: Smoothgrad: removing noise by adding noise. CoRR **abs/1706.03825** (2017), URL http://arxiv.org/abs/1706.03825 3
[63] Sucholutsky, I., Muttenthaler, L., Weller, A., Peng, A., Bobu, A., Kim, B., Love, B.C., Grant, E., Achterberg, J., Tenenbaum, J.B., Collins, K.M., Hermann, K.L., Oktar, K., Greff, K., Hebart, M.N., Jacoby, N., Zhang, Q., Marjieh, R., Geirhos, R., Chen, S., Kornblith, S., Rane, S., Konkle, T., O'Connell, T.P., Unterthiner, T., Lampinen, A.K., Müller, K., Toneva, M., Griffiths, T.L.: Getting aligned on representational alignment. CoRR **abs/2310.13018** (2023), https://doi.org/10.48550/ARXIV.2310.13018, URL https://doi.org/10.48550/arXiv.2310.13018 2
[64] Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017, Proceedings of Machine Learning Research, vol. 70, pp. 3319–3328, PMLR (2017), URL http://proceedings.mlr.press/v70/sundararajan17a.html 3
[65] Terry, M., Kulkarni, C., Wattenberg, M., Dixon, L., Morris, M.R.: AI alignment in the design of interactive AI: specification alignment, process alignment, and evaluation support. CoRR **abs/2311.00710** (2023), https://doi.org/10.48550/ARXIV.2311.00710, URL https://doi.org/10.48550/arXiv.2311.00710 2
[66] Tompson, J., Goroshin, R., Jain, A., LeCun, Y., Bregler, C.: Efficient object localization using convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, pp. 648–656, IEEE Computer Society (2015), https://doi.org/10.1109/CVPR.2015.7298664, URL https://doi.org/10.1109/CVPR.2015.7298664 14
[67] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan,

S.V.N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pp. 5998–6008 (2017), URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html 12

[68] Vellido, A., Martín-Guerrero, J.D., Lisboa, P.J.G.: Making machine learning models interpretable. In: 20th European Symposium on Artificial Neural Networks, ESANN 2012, Bruges, Belgium, April 25-27, 2012 (2012), URL https://www.esann.org/sites/default/files/proceedings/legacy/es2012-7.pdf 3

[69] Wang, Z., Qinami, K., Karakozis, I.C., Genova, K., Nair, P., Hata, K., Russakovsky, O.: Towards fairness in visual recognition: Effective strategies for bias mitigation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, pp. 8916–8925, Computer Vision Foundation / IEEE (2020), https://doi.org/10.1109/CVPR42600.2020.00894, URL https://openaccess.thecvf.com/content_CVPR_2020/html/Wang_Towards_Fairness_in_Visual_Recognition_Effective_Strategies_for_Bias_Mitigation_CVPR_2020_paper.html 6

[70] Wollek, A., Graf, R., Cecatka, S., Fink, N., Willem, T., Sabel, B.O., Lasser, T.: Attention-based saliency maps improve interpretability of pneumothorax classification. CoRR **abs/2303.01871** (2023), https://doi.org/10.48550/ARXIV.2303.01871, URL https://doi.org/10.48550/arXiv.2303.01871 1, 3

[71] Zhang, J., Bargal, S.A., Lin, Z., Brandt, J., Shen, X., Sclaroff, S.: Top-down neural attention by excitation backprop. Int. J. Comput. Vis. **126**(10), 1084–1102 (2018), https://doi.org/10.1007/S11263-017-1059-X, URL https://doi.org/10.1007/s11263-017-1059-x 1, 2, 3, 5, 14

[72] Zhou, C., Ma, X., Michel, P., Neubig, G.: Examining and combating spurious features under distribution shift. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, Proceedings of Machine Learning Research, vol. 139, pp. 12857–12867, PMLR (2021), URL http://proceedings.mlr.press/v139/zhou21g.html 6

A Appendix

A.1 Additional Examples

Fig. A1 and Fig. A2 illustrate the experimental setup used to evaluate the explanation alignment between model predictions and human expectations in ImageNet image classification tasks, providing visual context for the comparison of models with different model architectures and saliency methods.

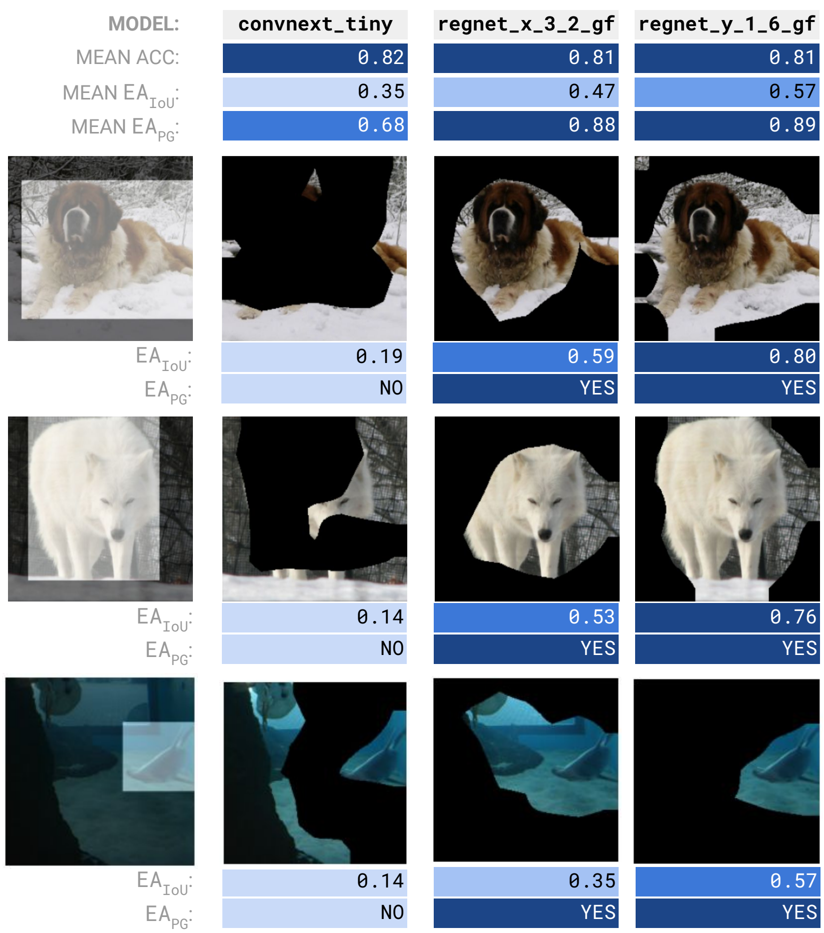

Fig. A1: Explanation alignment reveals that similarly accurate ImageNet models have different model reasoning. Despite each model (`convnext_tiny`, `regnet_x_3_2_gf`, and `regnet_y_1_6_gf`) achieving $81 - 82\%$ accuracy, their $\mathrm{EA}_{\mathrm{IoU}}$ and $\mathrm{EA}_{\mathrm{PG}}$ vary by 20%, reflecting the differences in human alignment visible in their saliency maps.

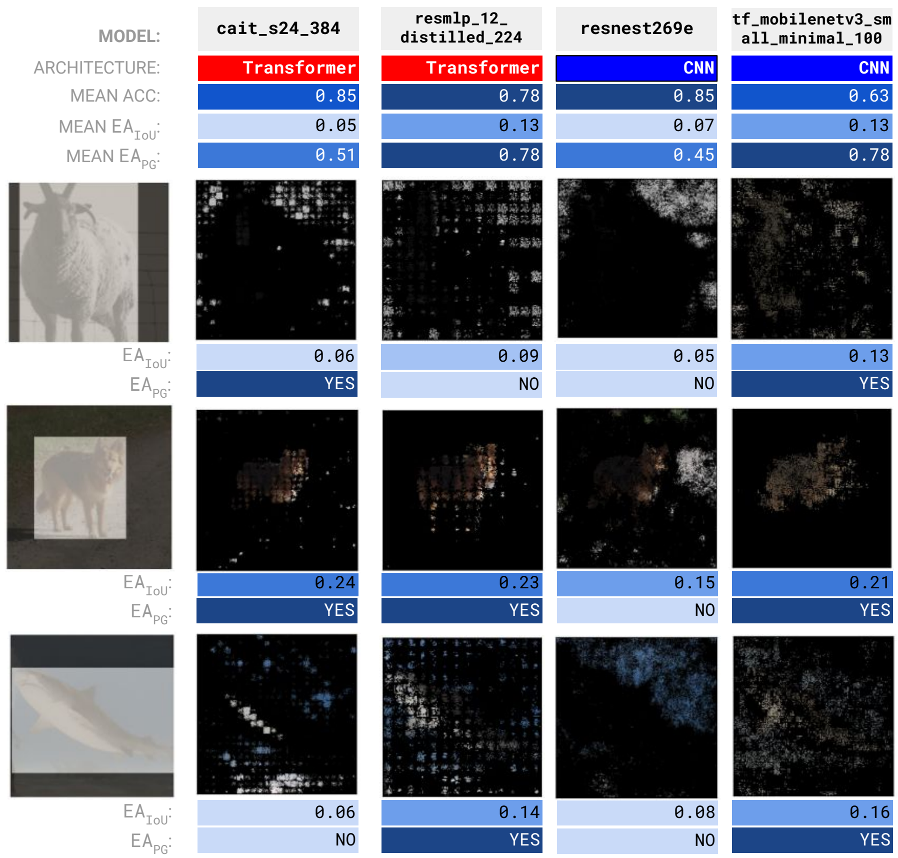

Fig. A2: The saliency maps of `cait_s24_384` and `resmlp_12_distilled_224` show more scattered and noisy patterns due to their transformer-based architecture, whereas `resnest269e` and `tf_mobilenetv3_small_minimal_100` exhibit more continuous maps, due to their convolutional network designs. Interestingly, `tf_mobilenetv3_small_minimal_100` yields higher mean explanation alignment despite less focused saliency maps, likely because the human explanation annotations in ImageNet tend to cover larger regions. This highlights the impact of the choice of saliency method, model architecture, and human explanations on measuring explanation alignment.

A.2 Model Training Details

In this paper, we conduct experiments on three distinct datasets: ImageNet [17], MNIST [18], and CelebA [37], using pretrained models and custom architectures to evaluate their performance across different image classification tasks. These experiments are executed on a GPU-enhanced, high-performance Power 9 system, featuring 64 nodes with 1TB memory each, equipped with four NVIDIA

V100 32GB GPUs per node, interconnected by NVLink2 for high-speed GPU communication and a 100Gb/s Infiniband network for cluster connectivity.

ImageNet We evaluate pretrained ImageNet [17] models provided by PyTorch [49]. Among them, we use three CNN models: ConvNeXt Tiny, RegNetX_3.2GF, and RegNetY_1.6GF. ConvNeXt, designed by updating a standard ResNet to mimic a Vision Transformer (ViT), results in similar accuracy to ViT but maintains the simplicity of standard ConvNets [38]. RegNet, a network design space rather than a single architecture, presents a variety of model architectures characterized by distinct parameters [51]. The key difference between RegNetX and RegNetY models is the inclusion of the Squeeze and Excitation layer in RegNetY.

MNIST To perform the experiment in Sec. 4.1, we design a neural network architecture for image classification. This architecture processes 3-channel input images through two convolutional layers with ReLU activations, incorporates max pooling and dropout layers to reduce overfitting, and concludes with two fully connected layers utilizing softmax activation for class probability output. This custom model is trained on a modified version of the MNIST dataset, described in Sec. 4.1, over four epochs with negative log likelihood loss. With the batch size of 64, the training of each epoch took approximately one minute, totaling four minutes per model.

CelebA We use the pretrained ImageNet [17] pretrained ResNet50 [27] model for image recognition. ResNet50 model, a part of the Residual Network (ResNet) series, is a deep convolutional neural network (CNN) architecture, with 50 layers in the network. The dataset is divided into biased and unbiased sets based on hair color and smiling attributes, further described in Sec. 4.2 for training and testing. Models are trained over five epochs using cross-entropy loss, with a batch size of 128. Each epoch took, in average, 16 minutes, totaling in 80 minutes per model.

A.3 Metric Computation

Saliency Method and Implementation After comparing different existing saliency methods, we use Grad-CAM [59] and Vanilla Gradients [61] for their effectiveness in producing representative explanations of model reasoning. Grad-Cam excels at localizing relevant areas in the image for the model's decision, whereas Vanilla Gradient demonstrate the sensitivity of each pixel on its decision. We use the publicly available implementations of these saliency methods from the Shared Interest paper [8].

Thresholding Technique For Grad-CAM, we select a threshold of 0.5, chosen after observing a range in average saliency values across models, which varied from nearly zero to almost one. This decision is validated through analyses across diverse settings, including differences in ground truth and saliency map sizes. This thresholding technique produces a consistent and balanced representation of the model's explanation across these settings.

For Vanilla Gradients, we apply a threshold set at one standard deviation above the mean, chosen through our analysis of saliency maps' focus and specificity. Due to its tendency to be noisier, thresholding based on a single value is insufficient. Also, thresholding at the mean results in broad, unfocused maps lacking in detailed model explanation, whereas thresholding at two standard deviations above the mean produced overly narrow map that are sometimes too limited for effective metric evaluation. The chosen threshold produces a balanced representation, capturing the essence of model explanations in a way that is both focused and sufficiently detailed.

List of 195 models used in experiments

1. adv_inception_v3
2. bat_resnext26ts
3. beit_base_patch16_224
4. beit_base_patch16_384
5. botnet26t_256
6. cait_s24_224
7. cait_s24_384
8. cait_s36_384
9. cait_xs24_384
10. cait_xxs24_224
11. cait_xxs24_384
12. cait_xxs36_224
13. cait_xxs36_384
14. coat_lite_mini
15. coat_lite_small
16. coat_lite_tiny
17. coat_mini
18. coat_tiny
19. convit_base
20. convit_small
21. convit_tiny
22. convmixer_1024_20_ks9_p14
23. convnext_base
24. cspdarknet53
25. cspresnet50
26. cspresnext50
27. deit_base_patch16_224
28. densenet121
29. dla60
30. dla60_res2net

31. dla60_res2next
32. dla60x
33. dla60x_c
34. dm_nfnet_f2
35. dpn68
36. efficientnetv2_rw_t
37. ens_adv_inception_resnet_v2
38. ese_vovnet19b_dw
39. ese_vovnet39b
40. fbnetc_100
41. fbnetv3_d
42. fbnetv3_g
43. gc_efficientnetv2_rw_t
44. gcresnet33ts
45. gcresnet50t
46. gcresnext26ts
47. gcresnext50ts
48. gernet_l
49. gernet_m
50. gernet_s
51. ghostnet_100
52. gluon_inception_v3
53. gluon_resnet101_v1b
54. gluon_resnet152_v1b
55. gluon_resnet152_v1c
56. gluon_resnet152_v1d
57. gluon_resnet152_v1s
58. gluon_resnet18_v1b
59. gluon_resnet34_v1b
60. gluon_resnet50_v1b
61. gluon_resnet50_v1c
62. gluon_resnet50_v1d
63. gluon_resnet50_v1s
64. gluon_resnext101_32x4d
65. gluon_senet154
66. gluon_seresnext101_32x4d
67. gmixer_24_224
68. gmlp_s16_224
69. halo2botnet50ts_256
70. halonet26t
71. haloregnetz_b
72. hardcorenas_a
73. hrnet_w18
74. ig_resnext101_32x16d
75. inception_resnet_v2
76. inception_v3

77. jx_nest_base
78. lambda_resnet26rpt_256
79. lamhalobotnet50ts_256
80. lcnet_050
81. legacy_senet154
82. legacy_seresnet101
83. legacy_seresnext101_32x4d
84. mixer_b16_224
85. mixnet_l
86. mnasnet_100
87. regnety_080
88. regnety_120
89. regnety_160
90. regnety_320
91. regnetz_b16
92. regnetz_c16
93. regnetz_d32
94. regnetz_d8
95. regnetz_e8
96. repvgg_a2
97. repvgg_b0
98. repvgg_b1
99. repvgg_b1g4
100. repvgg_b2
101. repvgg_b2g4
102. repvgg_b3
103. repvgg_b3g4
104. res2net101_26w_4s
105. res2net50_14w_8s
106. res2net50_26w_4s
107. res2net50_26w_6s
108. res2net50_26w_8s
109. res2net50_48w_2s
110. res2next50
111. resmlp_12_224
112. resmlp_12_distilled_224
113. resmlp_24_224
114. resmlp_24_distilled_224
115. resmlp_36_224
116. resmlp_36_distilled_224
117. resmlp_big_24_224
118. resmlp_big_24_224_in22ft1k
119. resmlp_big_24_distilled_224
120. resnest101e
121. resnest14d
122. resnest200e

123. resnest269e
124. resnest26d
125. resnest50d
126. resnest50d_1s4x24d
127. resnest50d_4s2x40d
128. resnet101
129. resnet101d
130. resnet152
131. resnet152d
132. resnet18
133. resnet18d
134. resnet200d
135. resnet26
136. resnet26d
137. resnet26t
138. resnet32ts
139. resnet33ts
140. resnet34
141. resnext26ts
142. resnext50_32x4d
143. sebotnet33ts_256
144. sehalonet33ts
145. selecsls60b
146. semnasnet_075
147. semnasnet_100
148. seresnet152d
149. seresnet33ts
150. seresnet50
151. seresnext26d_32x4d
152. seresnext26t_32x4d
153. seresnext26ts
154. seresnext50_32x4d
155. skresnet18
156. skresnet34
157. skresnext50_32x4d
158. spnasnet_100
159. ssl_resnet18
160. ssl_resnet50
161. ssl_resnext101_32x16d
162. ssl_resnext101_32x4d
163. ssl_resnext101_32x8d
164. ssl_resnext50_32x4d
165. swin_base_patch4_window12_384
166. swin_base_patch4_window7_224
167. swsl_resnet18
168. swsl_resnext101_32x4d

169. swsl_resnext101_32x8d
170. swsl_resnext50_32x4d
171. tf_efficientnet_b0
172. tf_efficientnet_b0_ap
173. tf_efficientnet_b0_ns
174. tf_efficientnet_b1
175. tf_efficientnet_b1_ap
176. tf_efficientnet_b1_ns
177. tf_efficientnet_b2
178. tf_efficientnet_b2_ap
179. tf_efficientnet_b2_ns
180. tf_efficientnet_b3
181. tf_efficientnet_b3_ap
182. tf_efficientnet_b3_ns
183. tf_efficientnet_b4
184. tf_inception_v3
185. tf_mixnet_s
186. tf_mobilenetv3_small_minimal_100
187. tinynet_a
188. tnt_s_patch16_224
189. tv_resnet50
190. tv_resnext50_32x4d
191. twins_pcpvt_base
192. twins_svt_base
193. vgg16
194. vgg16_bn
195. visformer_small