This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Explaining in Diffusion: Explaining a Classifier with Diffusion Semantics

Tahira Kazimi[†]
Virginia Tech
tahirakazimi@vt.edu

Ritika Allada[†]
Virginia Tech
ritika88@vt.edu

Pinar Yanardag
Virginia Tech
pinary@vt.edu

https://explain-in-diffusion.github.io

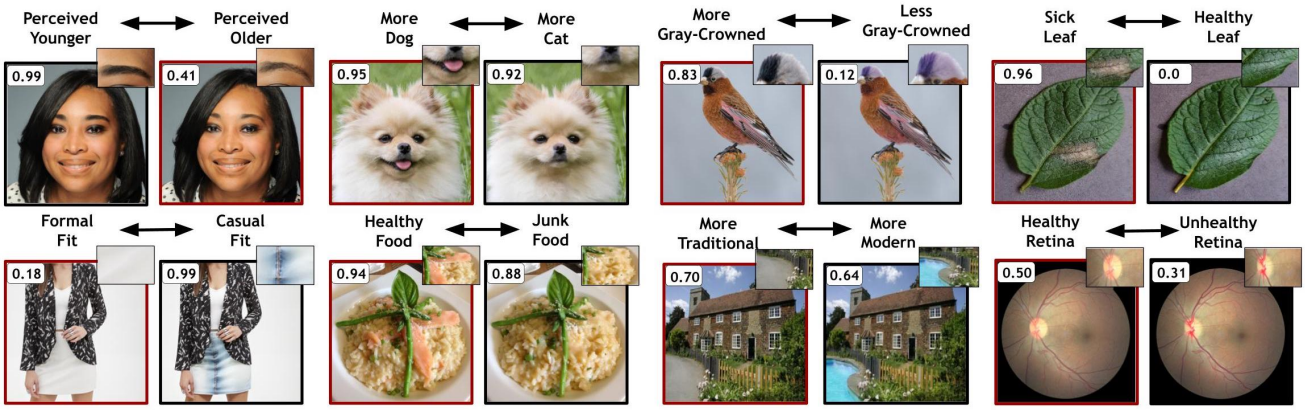

Figure 1. *DiffEx* explains the decisions of domain-specific classifiers by identifying the most influential semantics affecting their predictions. Classifier scores for each example are displayed in the top-left corner, demonstrating how classifier predictions change in response to the manipulation of different semantics (original images are shown with red borders). Our approach is capable of explaining classifiers that concentrate on individual concepts such as faces or animals (top row) as well as those that manage complex scenes involving multiple objects, such as a formal/casual fit in a fashion context (bottom row).

Abstract

Classifiers are important components in many computer vision tasks, serving as the foundational backbone of a wide variety of models employed across diverse applications. However, understanding the decision-making process of classifiers remains a significant challenge. We propose DiffEx, a novel method that leverages the capabilities of text-to-image diffusion models to explain classifier decisions. Unlike traditional GAN-based explainability models, which are limited to simple, single-concept analyses and typically require training a new model for each classifier, our approach can explain classifiers that focus on single concepts (such as faces or animals) as well as those that handle complex scenes involving multiple concepts. DiffEx employs vision-language models to create a hierarchical list of semantics, allowing users to identify not only the overarching semantic influences on classifiers (e.g., the 'beard' semantic in a facial classifier) but also their sub-types, such as 'goatee' or 'Balbo' beard. Our experiments demonstrate that DiffEx is able to cover a significantly broader spectrum of semantics compared to its GAN counterparts, providing a hierarchical tool that delivers a more detailed and fine-grained understanding of classifier decisions.

1. Introduction

Classifiers are fundamental to computer vision tasks, forming the backbone of many models used in a broad spectrum of applications [15, 21, 30, 32, 52]. Their ability to generalize across tasks and adapt to new domains with minimal retraining makes them highly transferable, and thus they are employed extensively in fields such as healthcare [37, 39, 42, 69], finance [38, 60, 62], security [2, 29, 33], and autonomous systems [5, 59, 65]. Despite their versatility and widespread utility, understanding the decision-making process of classifiers remains a significant challenge [3, 44, 50, 74, 76]. This challenge comes mainly from their "black box" nature. As images move through the deep layers of the network, the features used for classification become more abstract and harder to understand. The lack of interpretable features in such mod-

[†]Equal contributions.

els raises critical concerns, particularly in high-stakes environments such as medical diagnosis, where understanding the reasoning behind a model's prediction is crucial for ensuring trust, accountability, and informed decision-making [10, 34, 47, 57, 77]. Explaining classifier decisions is crucial for enhancing the transparency and reliability of these models. Prior research [31] has used generative adversarial networks (GANs) [18] to interpret classifier decisions by generating counterfactual examples that manipulate GAN latent semantics. These manipulations illustrate how changes in specific attributes, like the addition of the *eyeglasses* semantic, impact classifier outputs. However, GANs are often limited to single domains, such as facial images, and typically require training a new model per classifier, which is resource-intensive and time-consuming. Moreover, in GAN-based methods, understanding which latent semantics affect classifier decisions often requires manual intervention to identify and interpret relevant features, such as recognizing that a discovered semantic controls the *eyeglasses* attribute. This manual process is not only time-consuming but also less feasible in specialized fields like medicine, where identifying intricate attributes requires substantial domain expertise, making the approach impractical in critical scenarios.

This limitation highlights the need for more automated and versatile approaches to interpret classifier decisions. Text-to-Image (T2I) diffusion models [45] emerge as a compelling alternative, widely recognized for their ability to generate high-quality images across various domains, which makes them a promising tool for explaining classifier decisions. These models offer the potential for a richer and more diverse set of semantic features compared to existing methods. However, their ability to interpret and utilize latent space semantics remains limited in the context of diffusion models. Existing techniques for identifying meaningful semantics rely largely on supervised approaches [7, 8], which require users to craft detailed text prompts to specify particular features for editing, such as a *mustache*. This process requires domain expertise to craft appropriate prompts; thus, a broad semantic corpus is essential to improve diffusion models for classifier explanations while reducing manual effort. In this paper, we first employ Vision-Language Models (VLMs) [36] to extract a large-scale corpus covering domain-specific hierarchical semantics (see Fig. 2). Then, we introduce a training-free method, DiffEx, which leverages this hierarchical corpus and text-to-image diffusion models to explain the decision-making process of classifiers by identifying the most influential semantics. Our method provides explanations for both coarse and fine-grained semantics. For example, it can recognize a 'beard' as a coarse semantic influencing age classification scores and also demonstrate how specific beard types (such as 'Balbo' or 'Anchor' beards) impact the classifier's scores.

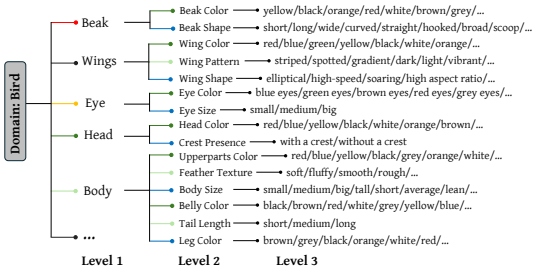

Figure 2. **Hierarchical List of Attributes for the Bird Domain.** We use VLMs to extract a hierarchical corpus of semantics within a given domain. This representation helps illustrate how different attributes are grouped within a broader domain, facilitating a better understanding of how each semantic contributes to the overall decision-making process of a classifier.

This hierarchical strategy provides users with an overview of the most significant semantics for a classifier, allowing them to dig deeper into particular fine-grained semantics that are essential for understanding classifier behavior. Our qualitative and quantitative experiments reveal that DiffEx offers considerably richer and more comprehensive explanations for binary and multi-class classifiers across various domains, including facial features, retinal health, and plant pathology. Our contributions are as follows:

- We propose `DiffEx`, a training-free approach using VLMs and T2I diffusion models to explain classifier decisions. To the best of our knowledge, this is the first hierarchical approach that explains classifier decisions.
- Our method employs a VLM to develop a comprehensive semantic corpus that spans multiple domains, with our source code publicly available to support future research.
- Unlike GAN-based methods, our approach can address classifiers that focus on single concepts (such as an 'age' classifier analyzing a headshot of a person) and also extend to classifiers for complex scenes (such as a 'modern/traditional architecture' classifier).
- Our method offers more comprehensible explanations for classifiers in applications ranging from facial recognition to retinal health compared to prior approaches and is adaptable to both binary and multiclassifiers.

2. Related Work

Traditional research focuses mainly on heat maps and patch-based extractions to explain classifier decisions. Specifically, class activation and saliency maps attempt to emulate human visual strategies by focusing on the image regions most relevant to a specific class [43, 49, 53, 56, 72, 75, 78]. While these maps highlight the object or part of an image that most influences the classifier's decision, they fail to reveal which specific, fine-grained object attributes (e.g., color, pattern, or texture) impact the classification. Other approaches try to explain classifier outputs by an-

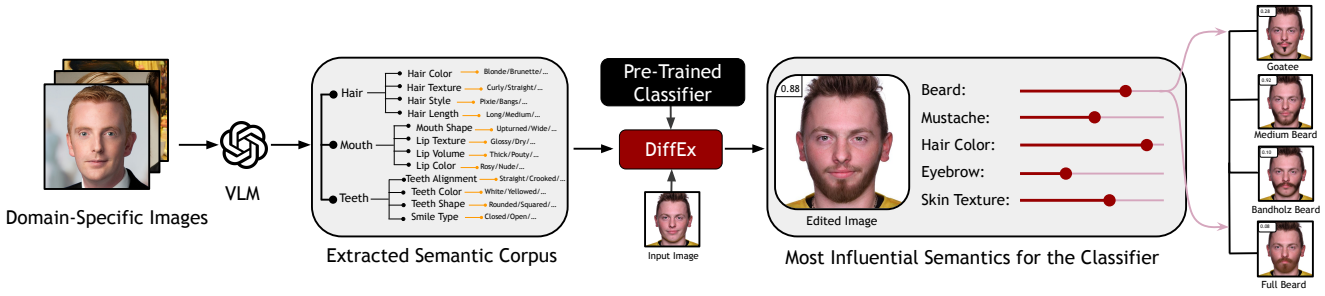

Figure 3. **An Overview of DiffEx.** Our pipeline processes a set of sample domain-specific images and a text prompt using a VLM to generate a hierarchical semantic corpus of attributes relevant to a specific domain. Based on this corpus, DiffEx identifies and ranks the most influential features affecting the classifier's decisions, sorting them from most to least impactful (rightmost image). The hierarchical explanation of semantics (like a "beard" and its subtypes) provides a fine-grained understanding of which features drive classifier outputs.

alyzing extracted image patches [16, 73]; however, these methods can only reveal spatially localized attributes. Previous studies have leveraged generative models such as variational autoencoders (VAEs) [19] or GANs [17, 48, 54, 55]. The most similar research to ours, StylEx [31], introduces a GAN-based approach to identify various attributes that influence a classifier's decisions. However, this method has several limitations. Firstly, it requires training a new GAN for each classifier, which can be resource-intensive and time-consuming. Secondly, each identified attribute requires manual labeling, which can require domain expertise. Lastly, StylEx only uncovers a limited range of semantic attributes, potentially overlooking others that might significantly affect a classifier's scores. Additionally, a notable limitation of the StylEx method and similar GAN models is their focus on single concepts, such as *human* or *animal* faces, or individual objects like *leaves*, rather than entire scenes. In contrast, diffusion models, known for their robust capability to generate complex and detailed entire scenes, offer a significant advantage. Our approach leverages the power of diffusion models to cover a wider spectrum of visual contexts, and enhances the versatility and applicability of our method across various classifiers that assess not only individual elements but also the interaction and composition of entire scenes.

Recent approaches have begun using diffusion models to generate counterfactual examples. One method utilizes shortcut learning to generate counterfactual images but fails to make semantically meaningful edits for certain attributes [68]. Another study explores modifying the diffusion process via adaptive parametrization and cone regularization to produce realistic counterfactual images; however, this approach depends on a robust model, which can be difficult to train [4]. [26] explored counterfactual image generation, however their approach is computationally demanding and uses DDPM [22] models trained on single domains. As a result, it does not take advantage of large-scale latent diffusion models like Stable Diffusion [46], which can handle more complex scenes. Furthermore, even though some studies have leveraged diffusion models to generate realistic counterfactuals in high-stakes domains [24], such as the medical field, they use a less efficient base model for the image generation process and focus predominantly on single-attribute modifications. Furthermore, to the best of our knowledge, there is no existing research that explores a hierarchical explanation of classifiers. Such an approach would systematically unpack the layers of influence that different semantic levels have on a classifier's scoring mechanism. This gap highlights a significant opportunity to enhance understanding by detailing how various semantic categories and their sub-types contribute to the decisions made by classifiers.

3. Methodology

We first discuss the background on identifying attributes that influence classifier decisions through changes in logits. Next, we introduce our method, which includes curating hierarchical attributes using VLMs and a novel algorithm inspired by beam search to pinpoint attributes that affect classifier scores. Our pipeline is detailed in Fig. 3.

3.1. Background

StylEx [31] identifies semantics that meaningfully influence classifier decisions by ranking each attribute based on its impact on the classifier's logit outputs. This ranking process aims to identify and select attributes that best explain the classifier's behavior in a given context, such as understanding the factors influencing *age* classification.

Given a semantic corpus $\mathcal{S}$ and a set of N images, counterfactual images are generated to analyze the influence of semantic attributes on classifier decisions. For each original image x_i, an edited version $x'_i = g(x_i, s)$ is generated by applying a semantic attribute $s \in \mathcal{S}$ through a transformation function g. This transformation highlights the effect of each semantic attribute on a classifier's output. Moreover, the logit difference for each attribute measures how the classifier's score changes due to the presence of s.

The influence of each attribute s is quantified as the aver-

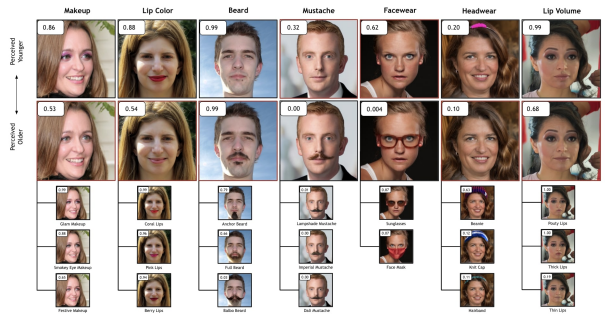

Figure 4. **Top-7 Facial Attributes for Age Classification.** DiffEx identifies key attributes and their top subtypes for the age classifier. Edited images are arranged hierarchically, with scores for the "young" label shown at the top-left of each image.

age logit difference between the original and edited images across a set of sample images, defined as:

$$I(s) = |f(X', s, y) - \frac{1}{N}\sum_{i=1}^{N} C(x_i, y)| \tag{1}$$

$$f(X, s, y) = \frac{1}{N}\sum_{i=1}^{N} C(g(x_i, s), y) \tag{2}$$

where $f(X, s, y)$ denotes the average classifier's logit score for target class y on image set X edited with attribute s. Here, y represents the unedited images' target labels, such as the "young" label for an age classifier. This influence score $I(s)$ captures the average impact of each semantic attribute on classifier decisions by reflecting the logit score change due to attribute manipulation.

3.2. Our Method

We first employ Vision-Language Models (VLMs) [36] to extract a large-scale corpus covering domain-specific hierarchical semantics. Then, we introduce a training-free method, *DiffEx*, which leverages this hierarchical corpus and text-to-image diffusion models to explain the decision-making process of classifiers by identifying the most influential semantics.

3.2.1. VLM-Based Semantic Space

While the StylEx method outlined in Section 3.1 employs a logit-based approach to identify semantics, it requires manually labeling each attribute extracted from the trained GAN model. Moreover, this method does not support the explanation of hierarchical attributes. Therefore, we first compile a large-scale set of semantics using VLMs. Given a domain d, such as the 'facial domain,' our objective is to identify a comprehensive set of domain-specific attributes from a collection of domain-specific images, denoted as N_d. To accomplish this, we utilize a Vision-Language Model (VLM) [36] to extract a range of relevant features, represented by $\mathcal{H}$. Employing in-context learning [9, 58], we prompt the VLM with a small set of images N_d, along with a detailed task description and examples of desired outputs. This process allows us to generate a substantial semantic corpus of keywords, $\mathcal{H}$, that captures the fine-grained attributes relevant to the domain. The resulting corpus, $\mathcal{H}$, comprises a comprehensive collection of keywords that covers full spectrum of the domain's attributes. We use this dataset to hierarchically explain how different attributes influence classifier decisions.

Algorithm 1 DiffEx

Require: Hierarchical structure $\mathcal{H}$ with semantic groups and features, beam width B, classifier or scoring function f, scoring threshold δ
Ensure: Optimal semantics maximizing I

```
Initialize S ← root-level groups in H {Initial candidate set at top-level groups}
Initialize beam B ← ∅
Score each candidate s ∈ S using the scoring function f(X, s, y) for y class label
Select top B candidates with I(b) ≥ δ and store in beam B and S {Apply thresholding to filter relevant candidates}
while B ≠ ∅ do
    Initialize S_next ← ∅
    for each candidate b ∈ B do
        Expand b by adding sub-features from its next level in H to form new candidates
        for each new combination b' expanded from b do
            if I(b') > I(b) then
                Add b' to S_next
            end if
        end for
    end for
    Set B ← S_next
    Append S_next to S
end while
Return highest-scoring combination from final S as the optimal joint semantic combination
```

3.2.2. DiffEx

Considering the extensive number of semantics identified using a VLM, it is computationally expensive to evaluate every possible semantic or combination of attributes. We introduce *DiffEx*, an efficient approach inspired by beam search to explain classifier decisions (see Algorithm 1). This method leverages our hierarchical semantic corpus to streamline the process. Our approach iteratively refines candidate attributes by expanding only the most impactful semantic paths at each hierarchical level, with each path's relevance guided by a scoring function that assesses the clas-

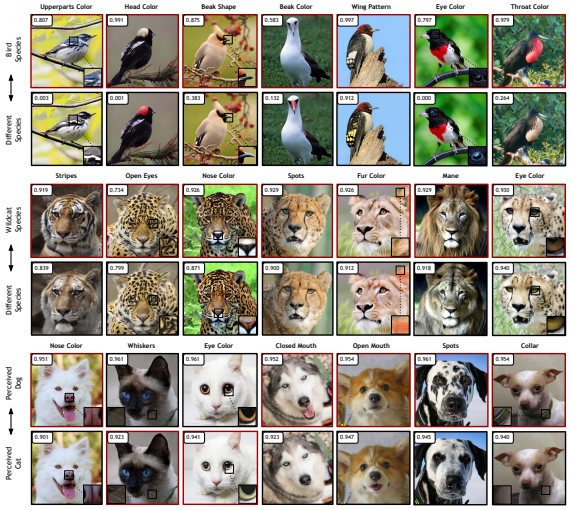

Figure 5. **Top-7 Discovered Attributes Across Different Animal Domains.** Our method successfully identifies key attributes for multiple domains, such as bird, wildcat, and pet species. The original images are depicted with red borders while the edited images are depicted with black borders. For the pet species domain, we used a binary classifier and for the bird and wildcat species domains, we used a multi-classifier.

sifier's response to each semantic feature. Formally, the scoring function f calculates the average classifier C logit scores across N images generated using each semantic attribute s, as defined in Eq. (2), where g represents the generative diffusion model applied to each sample image.

We begin with an initial candidate set S, which includes high-level groups from the semantic hierarchy (e.g., broader categories like 'mouth features' or 'eye features'). Each candidate in S is evaluated by the scoring function $f(X, s, y)$, which quantifies the influence of the candidate on the classifier's output. Only the top B candidates that surpass a predefined score threshold δ are retained, setting a beam width that limits the search to the most impactful candidates. For each candidate in the beam, the algorithm proceeds by expanding to the next hierarchical level, incorporating more specific sub-features (e.g., for "mouth features," sub-features such as "beard" and "mustache" are included). For each expanded candidate b', the scoring function in Eq. (2) is re-evaluated. Only candidates with a score exceeding that of their parent $I(b') > I(b)$ are retained in the next candidate set S_{next}, ensuring that only those additions that yield significant incremental impact are added. This process of expansion, scoring, and filtering continues iteratively, moving from general to more specific attributes at each hierarchical level. By dynamically adjusting the candidate set based on beam width B and scoring threshold δ, our method remains computationally efficient, focusing on high-impact combinations rather than exhaustively evaluating all possible attribute pairings. For generating counterfactual images, we utilize an off-the-shelf diffusion-based editing tool, Ledits++ [8]. However, our approach is versatile enough to accommodate any editing method.

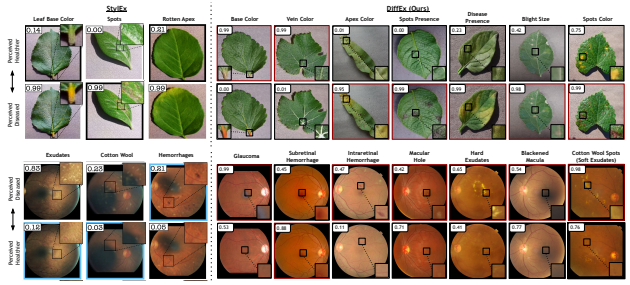

Figure 6. **Visual Comparison of Key Attributes Identified by StylEx and DiffEx in the Plant Health and Retinal Disease Domains.** This figure illustrates the enhanced capability of our method in identifying a broader set of significant attributes compared to StylEx within the plant health and retinal disease domains. DiffEx uncovers more detailed and diagnostically relevant features, such as "leaf vein color" and "macular hole," which provide deeper insights into leaf and retina health.

4. Experiments

Experimental Setup. Our experiments utilize Ledits++ [8] and Stable Diffusion XL (SDXL) [40] for generating counterfactual images. Twenty-five time steps are omitted to boost computational efficiency while preserving the quality of edits. The edit threshold, a hyperparameter that dictates the global application scope of edits, is adjusted based on each domain. We test our method across diverse classifiers to evaluate its effectiveness in explaining model behavior through semantic influence. Specifically, we utilize classifiers trained on facial attributes data (i.e. age and gender classifiers) [28] as well as plant health [23], retinal disease [13], bird species [64], wildcat/pet species [11], fashion [35], places [79], and food data [6]. The experiments for the face and plant health domains utilize a CLIP-based classifier [41] to evaluate and interpret edits, while experiments within the bird, wildcat, pet, food, fashion, and places domains use CNN-based classifiers. These CNN classifiers were built on the EfficientNet [61] architecture and achieved an accuracy of over 95 percent on their test sets. For the retinal disease domain, we utilized the FLAIR model [51], which is based on a pre-trained vision-language model, to classify various retina scans.

4.1. Qualitative Experiments

Explaining Classifiers with Single Attributes. We used various classifiers to analyze fine-grained semantic features for age categorization, species identification, leaf health assessment, retinal disease detection, etc. In the figures illustrating the visual results, the images with red borders indicate the original, unedited versions, while those with black borders are the edited versions, unless noted otherwise.

Face (Age)		Bird (Species)		Leaves (Health)		Retina Scans (Disease)		Wildcat (Species)		Pet (Cat/Dog)	
StylEx	**Ours**	**StylEx**	**Ours**	**StylEx**	**Ours**	**StylEx**	**Ours**	**StylEx**	**Ours**	**StylEx**	**Ours**
Skin Pigmentation	Eyebrow	Belly Color	Upperparts Color	Base Leaf Color	Base Color	Exudates	Glaucoma	Spots	Stripes	Open Mouth	Nose Color
Eyebrow Thickness	Makeup	Upperparts Color	Head Color	Apex Color	Vein Color	Cotton Wool Spots	Subretinal Hemorrhage	Black Tear Mark	Open Eyes	Closed Mouth	Whiskers
Eyeglasses	Mustache Type	Wing Pattern	Beak Shape	Spots	Apex Color	Hemorrhages	Intraretinal Hemorrhage	Eye Shape + Size	Nose Color	Eye Shape	Eye Color
Hair Color	Teeth	Beak Color	Beak Color	Blight	Spots Presence	Clustered Exudates	Macular Hole	✗	Spots	Dropped Ears	Closed Mouth
Lip Thickness + Position	Lip Volume	Head Color	Wing Pattern	Halos	Disease Presence	✗	Hard Exudates	✗	Fur Color	Pointed Ears	Open Mouth
Bangs	Lip Color	Breast Color	Eye Color	✗	Blight Size	✗	Blackened Macula	✗	Mane	Eye Circumference	Spots
Eye Makeup	Eyelash	✗	Throat Color	✗	Leaf Texture	✗	Soft Exudates	✗	Eye Color	✗	Collar
Facial Hair Color	Beard Type	✗	Wing Color	✗	Spots Color	✗	Retinal Drusen	✗	Tongue	✗	Pointed Ears
✗	Facewear	✗	Crest Presence	✗	Discoloration	✗	Optic Disc Hemorrhage	✗	Pupil Size	✗	Mouth Color
✗	Headwear	✗	Feather Texture	✗	Leaf Orientation	✗	Cataract	✗	Whiskers	✗	Fur Pattern

Table 1. **Comparison of Top Attributes Across Different Domains and Classifiers.** The table above contains a list of the top attributes discovered by DiffEx (Ours) vs. StylEx. The ✗ in the table indicates attributes that were not mentioned in StylEx. It is also important to note that "cotton wool spots" and "soft exudates" refer to the same condition within the retinal disease domain.

Method	Crest Presence	Beak Shape	Throat Color	Feather Texture	Eye Color	Beak Color	Head Color	Upperparts Color	Avg. Correct Response
Grad-CAM	36%	50%	56%	35%	47%	65%	59%	76%	53%
StylEx	68%	85%	79%	82%	74%	68%	91%	65%	76.5%
DiffEx (Ours)	**88%**	**91%**	**88%**	**91%**	**82%**	**82%**	**97%**	**88%**	**88.4%**

Table 2. **Comparison with Other Explainability Methods.** The table above displays the percentage of correct attribute selections for the bird class, as chosen by users when viewing outputs from different explainability methods. It also includes the average percentage of correct responses across all attributes for each method. As shown, for each attribute presented, the majority of users identified the correct attribute when viewing the output generated by DiffEx.

Face Domain: In the facial domain, as illustrated in Fig. 4, DiffEx identifies and ranks key attributes impacting age classification. For example, features such as "makeup styles," "lip volume," and "accessories" (e.g."hairbands") are associated with perceived youthfulness, whereas attributes like "facial hair" and "eyeglasses" are linked with perceived older age. Notably, the eyeglasses attribute consistently reduces the classifier's score for the "young" label, reflecting its association with older demographics. Additionally, DiffEx uncovers hierarchical attribute structures, demonstrating how subcategories within a feature can have varying effects on classifier outcomes. For instance, as shown in Fig. 4, different beard styles impact the perceived age differently (e.g. a "Balbo beard" significantly increases the age classification score more than a "full beard"). Additional examples of hierarchical explanations of age classifiers are detailed in the appendix (S10).

Animal Domain: Furthermore, DiffEx explains classifiers for a variety of animal types. Fig. 5 highlights the top-7 most influential attributes in the bird, wildcat, and pet domains where our method was able to identify fine-grained semantics to explain classifier behavior. For example, in the wildcat classifier, attributes like *stripes* and *manes* are critical in distinguishing wildcats from other species. Explaining these classifiers also reveals potential biases. For instance, the presence of a *collar* significantly increases the likelihood of an animal being classified as a dog. This bias may stem from the training dataset where images of dogs more frequently featured collars compared to cats.

Medical and Plant Health Domains: Fig. 6 demonstrates the results for retinal disease and plant health domains. For the retinal disease domain, we use the FLAIR model to classify images of retinal fundus scans. This model utilizes detailed domain expert knowledge descriptions, such as "no hemorrhages, microaneurysms, or exudates" compared to general descriptions like "no diabetic retinopathy" to aid in its classification. For the plant health domain, the CLIP classifier we use looks at features such as the presence of spots, fungus, etc. to make its decision.

Multi-Object Domains involving Complex Scenes: Unlike traditional GAN-based methods that primarily focus on single-object scenarios like a cropped face, DiffEx extends its utility by providing a list of relevant features for domains encompassing multiple objects, such as places, food, and fashion. Classifiers within these domains often evaluate multiple elements simultaneously. For example, a fashion classifier might determine the 'formal or casual fit' of an outfit by considering various components such as the hairstyle, top, and bottom. Similarly, an architectural classifier assessing whether a building appears 'urban or rural' may base its evaluation on not just the structure itself but also its exteriors and surroundings. Explaining the decision-making process of such classifiers is crucial, as they analyze multiple objects within a single image, adding complexity to their interpretative frameworks. Fig. 7 illustrates the top-7 attributes identified by DiffEx for the food, place, and fashion domains. Moreover, we demonstrate DiffEx on various diverse domains. For example, for the food domain, DiffEx was able to discover that removing "caviar" or a "plate decoration" from the image made the food appear to be more perceived as fast food compared to fine dining. For the places domain, DiffEx was able to find that adding a "tractor" to an image or a "dirt road" made the place seem more "rural" than "urban." In the fashion domain, the style of a neckline significantly influences whether an outfit was classified as "formal" or "casual." For instance, a V-neckline was often associated with a more casual look, whereas a classic boat neckline was generally perceived as more formal.

Explaining Classifiers with Joint Attributes. Our method goes beyond analyzing individual attributes by identifying attribute combinations that collectively improve classifier

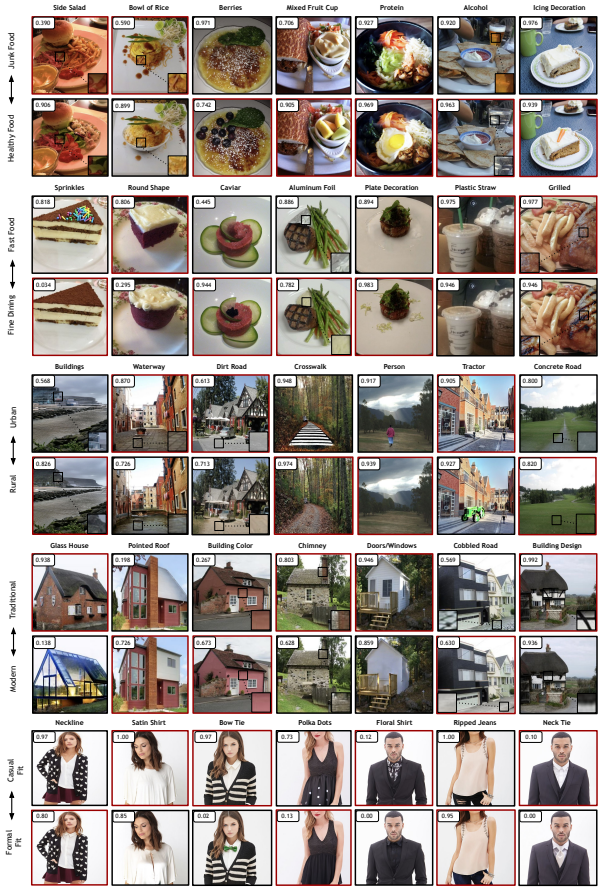

Figure 7. **Top-7 Attributes Discovered by DiffEx for Multi-Object Domains involving Complex Scenes.** Compared to existing methods, DiffEx was used to identify meaningful attributes for multi-object domains, such as "food," "places," and "fashion." By identifying semantic features that make food appear "healthier" or part of "fine dining," as well as attributes that give a place a more "rural" or "modern" feel, DiffEx can serve as a powerful tool for understanding perceptions through targeted edits.

Rating	Bird Domain	Face Domain
Edit Quality	3.386 ± 0.223	3.659 ± 0.248
Disentanglement	3.163 ± 0.197	3.204 ± 0.213

Table 3. **Edit Quality and Disentanglement Ratings.** These scores across various domains from User Study 1 are shown above.

interpretation. While single attributes may have a minimal impact on logits, their combined effect can substantially influence the classifier's output, uncovering subtle interactions that shape decision-making. For example, as shown in Fig. 8, Section C, individual changes in "lip color" or "eye makeup" result in minimal score changes, but when modified together, they make the subject appear significantly younger. This analysis shows how classifiers may respond more robustly to specific attribute combinations. Additional examples are provided in the appendix (S7, S8, S9).

Extending DiffEx for Multi-Classifier Analysis. We adapt DiffEx for multi-classifier applications to uncover semantic attributes essential for tasks like retinal disease classification (Fig. 6) and bird and wildcat species identification (Fig. 5). This approach highlights key features such as "upperparts color" and "beak shape" for bird species, "stripes" and "nose color" for wildcat species, and types of hemorrhages and exudates for retinal conditions. The results in Fig. 6 for the retinal diseases domain and Fig. 5 for the bird/wildcat species domain demonstrate that modifying these attributes significantly alters classifier scores, impacting assessments for species identification and disease likelihood.

Qualitative Comparisons. In Fig. 6, we present a visual comparison of the top attributes and classification scores identified by DiffEx against those identified by StylEx for the plant health and retinal disease domains. As illustrated, DiffEx successfully uncovers a broader set of semantically meaningful attributes compared to StylEx. For instance, DiffEx identifies detailed attributes such as "leaf vein color" and "spots color," in addition to the more general attributes found by StylEx, like "leaf base color" and "apex color." Table 1 provides a comprehensive comparison of the top features identified by StylEx and our method, highlighting DiffEx's superior ability to uncover semantics that are crucial for the classifier's decision-making process. In particular, the table presents an extended list of features across domains such as faces, bird species, plant health, retina scans, wildcat species, and pet types, all of which influence a classifier's score—further emphasizing DiffEx's ability to uncover fine-grained and contextually rich features. Additionally, in Figure 8 in Sections A and B, we visually compare DiffEx with other diffusion-based, counterfactual generation methods, such as FastDiME [68], DiME [26], and ACE [27] in the face and medical domains. As seen in Figure 8, Section B, DiffEx is able to preserve important features, such as a person's eyes, glasses, etc., in various cases, such as removing a smile or making an older person look younger. Furthermore, Figure 8, Section A shows DiffEx's ability to generate counterfactuals in various medical subdomains, such as skin lesions and chest x-rays. Specifically, for the skin lesion domain, DiffEx was able to remove ruler markings in images from the ISIC dataset [12, 63], while in the chest x-rays domain, DiffEx was able to remove chest drains and pacemakers from images from the NIH [67] dataset and CheXpert [25] dataset. As seen in Figure 8, Section A, DiffEx achieved similar results to FastDiME (Med) [68], a diffusion-based counterfactual model that was trained on medical images.

4.2. Quantitative Experiments

Baselines. To quantitatively evaluate the effectiveness of DiffEx compared to other explainability methods, such as StylEx [31] and Grad-CAM [49], we conducted a series of

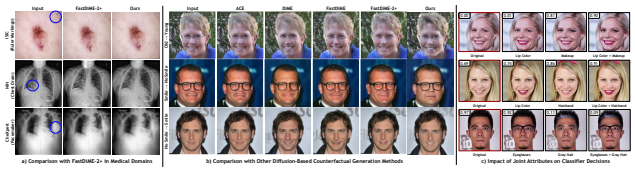

Figure 8. **More Comparisons and Exploring Joint Attributes.** (a) We demonstrate DiffEx's ability to produce counterfactuals comparable to FastDiME-2+ in skin lesion and chest X-ray domains. (b) compares DiffEx with ACE, DiME, FastDiME, and FastDiME-2+, showing its capacity to add or remove attributes in a disentangled manner. (c) highlights DiffEx's effectiveness in explaining classifier decisions by analyzing the influence of joint attributes, with perceived age scores (e.g., "younger" label) displayed to show how attribute pairings affect classifier outputs.

comprehensive user studies to assess how easily participants could identify the relevant features extracted by our method.

User Study 1: Visual Quality and Disentanglement. To evaluate the visual quality and disentanglement of edited images for the face and bird domains, we conducted a user study with 50 participants on Prolific.com. For each domain, we showed pairs of unedited and edited images for ten attributes and asked users to assess whether the edited image contained the desired attribute and was disentangled. For each pair of images, we asked users to rate the edit and disentanglement from one to five, with five representing the highest score. Our results (see Table 3) indicate that our edits reflected the intended attributes while minimizing unrelated changes (see S6.1 in the appendix for more details).

User Study 2: Comparison with Grad-CAM and StylEx. To evaluate how effectively our method, DiffEx, explains various semantics within a specific domain compared to other explainability methods (Grad-CAM and StylEx), we conducted a user study with 35 participants on Prolific. We focused on images from the bird domain, generating three sets of three images per attribute: one original image, one edited image, and one image illustrating the explainability method. For Grad-CAM, participants were shown a heatmap overlay on the edited image. For StylEx, we displayed the edited image alone, as this method requires users to manually label the edits. For DiffEx, participants viewed the edited image and the attribute automatically assigned by the VLM. We then asked users to select the attribute (e.g., "beak color," "crest presence") that best explained the edited image from four choices. Results indicate that DiffEx significantly outperforms Grad-CAM and StylEx in explaining edited attributes in images (see Table 2). Please refer to S6.2 in the appendix for details.

User Study 3: Disentanglement in Images with Multiple Attributes. We evaluated DiffEx's disentanglement ability when editing multiple attributes simultaneously via a user study. Users rated images on a scale of 1 to 5 (5 = very disentangled, 1 = entangled). DiffEx achieved an average score of 4.6 for single-attribute edits and 4.3 for edits involving four or more attributes, demonstrating effective handling of multiple edits (see S6.3 in the appendix for more details).

Ablation Studies. To assess the robustness of our method, we tested various Vision-Language Models (VLMs) and Diffusion Models. For VLMs, we probed Llama 3.2 [20], Qwen2-VL [66], and Claude 3.5 Haiku [1] within the face domain, measuring semantic consistency via IoU scores against GPT-4. The results were 0.94, 0.87, and 0.93, respectively, which indicate substantial semantic overlap. For diffusion models, we compared top attributes from Stable Diffusion (SD) 1.5 and SD 2.0 [46] against SDXL [40] using Normalized Discounted Cumulative Gain (nDCG) to measure the attributes' overlap, achieving scores of 0.75 and 0.76, respectively. These results confirm the stability of our method across different models.

5. Discussion

Limitations. While our approach provides valuable insights into classifier behavior through semantic edits, it has limitations. Relying on VLM-curated semantics restricts us to the quality and scope of the initial corpus, which may miss relevant features. Additionally, using off-the-shelf editing models like Ledits++ can result in entangled edits (a common issue in image editing algorithms [71]) or fail to capture fine-grained details like earlobes or nose rings (see Appendix S1). Nonetheless, our framework is flexible and can be enhanced with domain-specific adjustments, such as task-specific RAGs [70] or editing methods [7, 14]. RAGs can automatically retrieve semantics in specialized domains (e.g. medical) to improve generalization.

Conclusion. In this work, we introduce DiffEx, a novel approach for explaining classifier decisions by utilizing semantic edits within diffusion models. By harnessing the power of vision language models, we curate a comprehensive, hierarchical semantic corpus across various domains and propose a novel algorithm inspired by beam search to filter and rank the most impactful features. DiffEx ranks these semantic features based on their influence on classifier logits, capturing both individual and joint attribute effects, which are crucial for understanding complex classifier behaviors. Through experiments conducted on a wide range of domains—including face, bird, and medical classifiers—we showcase the robustness and adaptability of our approach. Our work provides a powerful tool for interpreting model decisions across diverse applications and promoting transparency in AI-driven classification systems.

References

[1] The claude 3 model family: Opus, sonnet, haiku. 8

[2] Mostofa Ahsan, Rahul Gomes, Md. Minhaz Chowdhury, and Kendall E. Nygard. Enhancing machine learning prediction in cybersecurity using dynamic feature selector. *Journal of Cybersecurity and Privacy*, 1(1):199–218, 2021. 1

[3] Leila Amgoud. Explaining black-box classifiers: Properties and functions. *International Journal of Approximate Reasoning*, 155:40–65, 2023. 1

[4] Maximilian Augustin, Valentyn Boreiko, Francesco Croce, and Matthias Hein. Diffusion visual counterfactual explanations. *Advances in Neural Information Processing Systems*, 35:364–377, 2022. 3

[5] Ouahiba Azouaoui and Amine Chohra. Soft computing based pattern classifiers for the obstacle avoidance behavior of intelligent autonomous vehicles (iav). *Applied intelligence*, 16:249–272, 2002. 1

[6] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part VI 13*, pages 446–461. Springer, 2014. 5

[7] Manuel Brack, Felix Friedrich, Dominik Hintersdorf, Lukas Struppek, Patrick Schramowski, and Kristian Kersting. SEGA: Instructing text-to-image models using semantic guidance. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2, 8

[8] Manuel Brack, Felix Friedrich, Katharia Kornmeier, Linoy Tsaban, Patrick Schramowski, Kristian Kersting, and Apolinário Passos. Ledits++: Limitless image editing using text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8861–8870, 2024. 2, 5

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 4

[10] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019. 2

[11] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. *CoRR*, abs/1912.01865, 2019. 5

[12] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019. 7

[13] Jorge Cuadros and George Bresnick. Eyepacs: An adaptable telemedicine system for diabetic retinopathy screening. *Journal of Diabetes Science and Technology*, 3(3):509–516, 2009. 5

[14] Yusuf Dalva and Pinar Yanardag. Noiseclr: A contrastive learning approach for unsupervised discovery of interpretable directions in diffusion models. *arXiv preprint arXiv:2312.05390*, 2023. 8

[15] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[16] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. *Advances in neural information processing systems*, 32, 2019. 3

[17] Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. Ganalyze: Toward visual definitions of cognitive image properties. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5744–5753, 2019. 3

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[19] Yash Goyal, Amir Feder, Uri Shalit, and Been Kim. Explaining classifiers with causal concept effect (cace). *arXiv preprint arXiv:1907.07165*, 2019. 3

[20] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 8

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. (arXiv:1512.03385), 2015. arXiv:1512.03385. 1

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[23] David P. Hughes and Marcel Salathe. An open access repository of images on plant health to enable the development of mobile disease diagnostics. (arXiv:1511.08060), 2016. arXiv:1511.08060 [cs]. 5

[24] Indu Ilanchezian, Valentyn Boreiko, Laura Kühlewein, Ziwei Huang, Murat Seçkin Ayhan, Matthias Hein, Lisa Koch, and Philipp Berens. Generating realistic counterfactuals for retinal fundus and oct images using diffusion models. *arXiv preprint arXiv:2311.11629*, 2023. 3

[25] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, pages 590–597, 2019. 7

[26] Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie. Diffusion models for counterfactual explanations. In *Proceedings of the Asian Conference on Computer Vision*, pages 858–876, 2022. 3, 7

[27] Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie. Adversarial counterfactual visual explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16425–16435, 2023. 7

[28] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 4396–4405, Long Beach, CA, USA, 2019. IEEE. 5

[29] Ziv Katzir and Yuval Elovici. Quantifying the resilience of machine learning classifiers used for cyber security. *Expert Systems with Applications*, 92:419–429, 2018. 1

[30] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 1

[31] Oran Lang, Yossi Gandelsman, Michal Yarom, Yoav Wald, Gal Elidan, Avinatan Hassidim, William T Freeman, Phillip Isola, Amir Globerson, Michal Irani, et al. Explaining in style: Training a gan to explain a classifier in stylespace. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 693–702, 2021. 2, 3, 7

[32] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 1

[33] Joffrey L. Leevy, John Hancock, Richard Zuech, and Taghi M. Khoshgoftaar. Detecting cybersecurity attacks across different network features and learners. *Journal of Big Data*, 8(1):38, 2021. 1

[34] Xuhong Li, Haoyi Xiong, Xingjian Li, Xuanyu Wu, Xiao Zhang, Ji Liu, Jiang Bian, and Dejing Dou. Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond. *Knowledge and Information Systems*, 64 (12):3197–3234, 2022. 2

[35] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016. 5

[36] OpenAI. Gpt-4 technical report, 2024. 2, 4

[37] Akin Ozcift and Arif Gulten. Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms. *Computer Methods and Programs in Biomedicine*, 104(3):443–451, 2011. 1

[38] Yi Peng, Guoxun Wang, Gang Kou, and Yong Shi. An empirical study of classification algorithm evaluation for financial risk prediction. *Applied Soft Computing*, 11(2):2906–2915, 2011. 1

[39] Emmanuel Pintelas, Meletis Liaskos, Ioannis E Livieris, Sotiris Kotsiantis, and Panagiotis Pintelas. Explainable machine learning framework for image classification problems: case study on glioma cancer prediction. *Journal of imaging*, 6(6):37, 2020. 1

[40] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 5, 8

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 5

[42] R. Joshua Samuel Raj, S. Jeya Shobana, Irina Valeryevna Pustokhina, Denis Alexandrovich Pustokhin, Deepak Gupta, and K. Shankar. Optimal feature selection-based medical image classification using deep learning model in internet of medical things. *IEEE Access*, 8:58006–58017, 2020. 1

[43] Sylvestre-Alvise Rebuffi, Ruth Fong, Xu Ji, and Andrea Vedaldi. There and back again: Revisiting backpropagation saliency methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8839–8848, 2020. 2

[44] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery. 1

[45] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 2

[46] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3, 8

[47] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surveys*, 16:1–85, 2022. 2

[48] Axel Sauer and Andreas Geiger. Counterfactual generative networks. (arXiv:2101.06046), 2021. arXiv:2101.06046 [cs]. 3

[49] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 2, 7

[50] Andy Shih, Arthur Choi, and Adnan Darwiche. A symbolic approach to explaining bayesian network classifiers. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, page 5103–5111. AAAI Press, 2018. 1

[51] Julio Silva-Rodriguez, Hadi Chakor, Riadh Kobbi, Jose Dolz, and Ismail Ben Ayed. A foundation language-image model of the retina (flair): Encoding expert knowledge in text supervision. *Medical Image Analysis*, 99:103357, 2025. 5

[52] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. 2014. 1

[53] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 2

[54] Sumedha Singla, Brian Pollack, Junxiang Chen, and Kayhan Batmanghelich. Explanation by progressive exaggeration. (arXiv:1911.00483), 2020. arXiv:1911.00483 [cs]. 3

[55] Sumedha Singla, Motahhare Eslami, Brian Pollack, Stephen Wallace, and Kayhan Batmanghelich. Explaining the black-box smoothly- a counterfactual approach. (arXiv:2101.04230), 2022. arXiv:2101.04230 [cs, eess]. 3

[56] Suraj Srinivas and François Fleuret. Full-gradient representation for neural network visualization. *Advances in neural information processing systems*, 32, 2019. 2

[57] Gregor Stiglic, Primoz Kocbek, Nino Fijacko, Marinka Zitnik, Katrien Verbert, and Leona Cilar. Interpretability of machine learning-based prediction models in healthcare. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(5):e1379, 2020. 2

[58] Hongjin Su, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, et al. Selective annotation makes language models better few-shot learners. *arXiv preprint arXiv:2209.01975*, 2022. 4

[59] Prabu Subramani, Khalid Sattar, Rocío De Prado, Balasubramanian Girirajan, and Marcin Wozniak. Multi-classifier feature fusion-based road detection for connected autonomous vehicles. *Applied Sciences*, 11(17):7984, 2021. 1

[60] Jie Sun and Hui Li. Financial distress prediction based on serial combination of multiple classifiers. *Expert Systems with Applications*, 36(4):8659–8666, 2009. 1

[61] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2020. 5

[62] Manoj Thakur and Deepak Kumar. A hybrid financial trading support system using multi-category classifiers and random forest. *Applied Soft Computing*, 67:337–349, 2018. 1

[63] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018. 7

[64] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge J. Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 5

[65] Wahyono, Laksono Kurnianggoro, Joko Hariyono, and Kang-Hyun Jo. Traffic sign recognition system for autonomous vehicle using cascade svm classifier. In *IECON 2014 - 40th Annual Conference of the IEEE Industrial Electronics Society*, pages 4081–4086, 2014. 1

[66] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 8

[67] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017. 7

[68] Nina Weng, Paraskevas Pegios, Eike Petersen, Aasa Feragen, and Siavash Bigdeli. Fast diffusion-based counterfactuals for shortcut removal and generation. In *European Conference on Computer Vision*, pages 338–357. Springer, 2025. 3, 7

[69] M. Wiggins, A. Saad, B. Litt, and G. Vachtsevanos. Evolving a bayesian classifier for ecg-based age classification in medical applications. *Applied Soft Computing*, 8(1):599–608, 2008. 1

[70] Junde Wu, Jiayuan Zhu, and Yunli Qi. Medical graph rag: Towards safe medical large language model via graph retrieval-augmented generation. *arXiv preprint arXiv:2408.04187*, 2024. 8

[71] Qiucheng Wu, Yujian Liu, Handong Zhao, Ajinkya Kale, Trung Bui, Tong Yu, Zhe Lin, Yang Zhang, and Shiyu Chang. Uncovering the disentanglement capability in text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1900–1910, 2023. 8

[72] Shawn Xu, Subhashini Venugopalan, and Mukund Sundararajan. Attribution in scale and space. (arXiv:2004.03383), 2020. arXiv:2004.03383 [cs]. 2

[73] Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. On completeness-aware concept-based explanations in deep neural networks. *Advances in neural information processing systems*, 33: 20554–20565, 2020. 3

[74] Muhammad Bilal Zafar, Michele Donini, Dylan Slack, Cédric Archambeau, Sanjiv Das, and Krishnaram Kenthapadi. On the lack of robust interpretability of neural text classifiers. *arXiv preprint arXiv:2106.04631*, 2021. 1

[75] MD Zeiler. Visualizing and understanding convolutional networks. In *European conference on computer vision/arXiv*, 2014. 2

[76] Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu. Interpretable convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[77] Yu Zhang, Peter Tiňo, Aleš Leonardis, and Ke Tang. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5):726–742, 2021. 2

[78] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 2

[79] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 5