JUST READ THE CAPTION: A BLACK-BOX PRIVACY ATTACK ON KNOWLEDGE INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

In this work, we demonstrate that vision classifiers leak private and potentially sensitive information about their semantic domain and class labels even under restricted black-box access. We show that softmax confidence acts as a natural in-distribution discriminator that an attacker can exploit to isolate relevant images from noisy web collections, using their captions to infer the meaning of otherwise opaque predictions. We introduce KEEP, a black-box knowledge inference attack that requires only the top-1 class identifier and its confidence, without prior knowledge of the victim's domain, training data, or class names. KEEP probes the victim with general-purpose image–caption pairs and uses a language model to infer its domain from the highest-confidence evidence. This domain then guides the construction of targeted probing pools from which to recover individual class labels. Across fourteen classifiers spanning four architectures and diverse domains, KEEP recovers exact class names with high accuracy, including copyrighted characters, trademarked brands, and personal identities. These results reveal a largely unexplored source of information leakage, with significant implications for the privacy and confidentiality of models trained on sensitive or proprietary data. Our code is available at `https://anonymous.4open.science/r/black-box-knowledge-inference-C195/`.

1 INTRODUCTION

Modern machine learning systems are often trained on private, proprietary, or otherwise sensitive data, such as medical records (Esteva et al., 2017), biometric information (Schroff et al., 2015), or copyrighted material (Carlini et al., 2023). It is often assumed that once a model has been trained, the underlying data becomes effectively concealed within the model parameters and cannot be accessed by external parties. Although a substantial body of work has investigated privacy leakage from trained models, existing attacks mostly either target narrowly defined properties of *individual* training samples (Shokri et al., 2017; Carlini et al., 2022; Choquette-Choo et al., 2021) or rely on strong prior knowledge about the training domain (Fredrikson et al., 2015; Zhang et al., 2020; Kahla et al., 2022), with their effectiveness degrading substantially when such prior domain knowledge or in-distribution data is unavailable (Gu & Chen, 2023; Gu et al., 2025). By contrast, inferring broader properties of a model's training data (e.g., its domain, semantic classes, or label space) under minimal assumptions has received limited attention. This information is valuable not only because it can supply the auxiliary knowledge assumed by existing attacks (Zhang et al., 2020; Chen et al., 2021; Struppek et al., 2022), but also because these properties may *themselves* reveal private or sensitive information about the training data.

For example, in intellectual-property settings, determining that a model was trained on material from a particular film, textbook, or artist may be more consequential than identifying any single frame, passage, or image. Likewise, for a face-recognition model, establishing that a particular person was represented in the training data may constitute a meaningful privacy disclosure even when no specific photograph can be identified. Such source- and identity-level inferences can therefore carry substantial legal, commercial, and privacy implications (Suri & Evans, 2021; Prabhu & Birhane, 2020). These concerns have attracted increasing attention amid multiple lawsuits against major artificial intelligence companies over the alleged use of copyrighted material to train their models (U.S. Copyright Office, 2025).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

In this work, we demonstrate that meaningful domain- and class-level information can be recovered from unknown classifiers. We consider a realistic black-box setting in which an attacker observes only an opaque predicted label and its confidence score, without prior knowledge of the model's training data or application domain. We introduce KEEP (**K**nowledge **E**xtraction via **E**xternal **P**robing), a simple framework that probes the target model with an external collection of captioned images and jointly analyzes its responses and the semantics of the captions to uncover a model's domain and classes. Because it relies only on the top-1 identifier and its confidence, KEEP is independent of the victim's architecture, and operates entirely post hoc without training or optimization. We evaluate KEEP on fourteen classifiers spanning ResNet (He et al., 2016), ViT (Dosovitskiy et al., 2021), ConvNeXt (Liu et al., 2022), and Swin (Liu et al., 2021) architectures, whose class vocabularies include copyrighted characters, trademarked brands, personal identities, and medical diagnoses. KEEP infers each victim's domain and names its labels as free text, outperforming CLIP-based methods (Radford et al., 2021; Oikarinen & Weng, 2023) developed for mechanistic interpretability on ten of the fourteen victims, even when those methods are given the victim's ground-truth class names.

The main contributions of this paper can be summarized as follows:

1. A systematic formulation of *domain- and class-level knowledge inference* as an underexplored form of model leakage.
2. KEEP, a black-box probing framework for recovering domain- and class-level information from unknown image classifiers using general-purpose collections of captioned images, requiring no prior knowledge of the model's training dataset or domain.
3. A comprehensive empirical evaluation on fourteen classifiers spanning four architectures and a wide range of domains, where KEEP recovers both the victim's domain and its individual class names and exposes copyrighted characters, trademarked brands, and personal identities.

Our results reveal that a classifier's output vocabulary constitutes a broader source of information leakage than is captured by the conventional sample-level view of privacy, with potentially serious consequences for models trained on sensitive or proprietary data. By demonstrating that this information can be recovered even under restricted black-box access, we aim to draw attention to an underexplored threat and motivate the development of safer models that protect the information they encode.

2 Related Work

Membership inference. Membership inference attacks ask whether a specific data sample was part of a model's training set, exploiting the fact that models behave differently, typically more confidently, on data they have memorized than on data they see for the first time (Shokri et al., 2017; Yeom et al., 2018). The canonical approach trains *shadow models* that imitate the victim and an attack classifier that reads the victim's output to decide membership (Shokri et al., 2017); later work relaxes the shadow-model and data assumptions (Salem et al., 2018; Choquette-Choo et al., 2021) and reframes the attack as a calibrated per-example likelihood-ratio test (Carlini et al., 2022; Zarifzadeh et al., 2024), but the target remains fundamentally *per-sample*. These attacks answer a binary question about a single record, and presuppose access to data drawn from the training distribution, or close to it. In contrast, we study model-level knowledge inference: recovering the semantic domain and class vocabulary encoded by a classifier without specifying candidate training samples. Our setting requires no in-distribution auxiliary data, relying only on a general-purpose collection of images and black-box access to the victim's predictions.

Model inversion. Model inversion reconstructs a representative input for a *given* output class, recovering, for example, a recognizable face associated with a target identity (Fredrikson et al., 2015). Modern variants often rely on generative priors trained on public data to improve reconstruction fidelity (Zhang et al., 2020; Chen et al., 2021; Struppek et al., 2022), including under black-box and label-only access (Yang et al., 2019; An et al., 2022; Kahla et al., 2022; Han et al., 2023). However, two properties limit their ability to reveal class-level knowledge. First, they assume that the target class is already known: inversion reconstructs an input for a class specified by the attacker rather than discovering the model's label space. Second, they typically require a public prior that is sufficiently

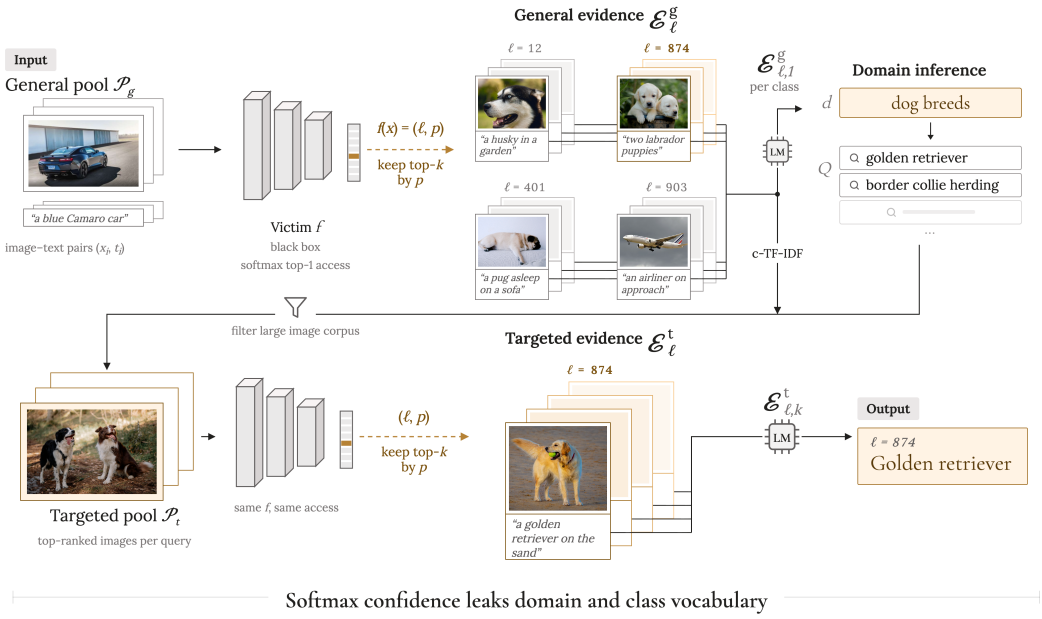


Figure 1: **KEEP recovers class names from a black-box classifier in two probing rounds.** A general pool $\mathcal{P}_g$ of captioned images is fed to the victim f, which returns only an opaque top-1 identifier and its confidence. The captions of each class's most confident responses form its general evidence $\mathcal{E}^g_\ell$. A language model reads the strongest evidence of every class to infer the victim's domain d and a set of search queries Q, which retrieve a targeted pool $\mathcal{P}_t$. Querying f again with $\mathcal{P}_t$ yields targeted evidence $\mathcal{E}^t_\ell$, from which a language model predicts a free-text name for each class.

aligned with the private training distribution, and reconstruction quality degrades as this alignment weakens (Zhang et al., 2020; Struppek et al., 2022; Han et al., 2023; Gu et al., 2025). In contrast, our method infers the semantic identities of a model's classes directly from black-box outputs, without requiring known target labels or auxiliary data aligned with the victim's training distribution. The recovered domain and class information can strengthen existing model inversion attacks by guiding the selection of auxiliary datasets that more closely match the victim's domain (Hayes et al., 2017; Matsumoto et al., 2023; Zhang et al., 2020), the goal pursued by domain inference attacks (Gu & Chen, 2023; Gu et al., 2025), which recover the domain but not the class vocabulary.

3 Method and Threat Model

We consider an attacker with query access to an unknown image classifier f through a prediction API, an access model standard in prior work on attacks against remotely hosted machine learning models (Tramèr et al., 2016; Shokri et al., 2017; Orekondy et al., 2019). We adopt a restrictive setting in which the attacker has access only to the top-1 prediction and its associated softmax confidence score. Formally, for each input image x, the attacker observes only $f(x) = (\ell, p)$, where ℓ is a stable, opaque identifier for the highest-probability class (e.g., an internal class ID, as commonly returned by commercial vision APIs) and p is its confidence; no scores or identifiers for other classes are returned. The attacker has no access to the model parameters, gradients, intermediate activations, and has no prior knowledge of the model's training dataset, application domain, or semantic class names. The attacker's objective is to infer the model's semantic domain and recover the meaning of its output classes.

KEEP (**K**nowledge **E**xtraction via **E**xternal **P**robing) builds on the observation that softmax confidence is a natural in-distribution signal (Hendrycks & Gimpel, 2016). Figure 1 summarizes the procedure. The method probes the victim with a collection of captioned images and treats the captions of its most confident responses as evidence of the semantic concept each output class encodes. Let $\mathcal{P}_g = \{(x_i, t_i)\}_i$ denote a general pool of image-text pairs, where each image x_i is associated

with a textual description t_i, such as a caption or label. The victim is queried with every image in $\mathcal{P}_g$, producing a response $f(x_i) = (\ell_i, p_i)$, where ℓ_i is the opaque top-1 class identifier and p_i its confidence score. The descriptions of the images the victim labels ℓ form its *general evidence set*, $\mathcal{E}^g_\ell = \{t_i : \ell_i = \ell\}$, and we write $\mathcal{E}^g_{\ell,k}$ for its k most confident entries.

General-purpose probing pools have an inherent limitation on specialized victims, whose classes are only sparsely represented in broad collections. To overcome this, we use the general evidence sets themselves to construct a targeted pool. First, the victim's domain is inferred from the strongest evidence of every observed class. A language model reads the most confident description $\mathcal{E}^g_{\ell,1}$ of each ℓ and returns a short domain description d. The inferred domain then guides the construction of a targeted probing pool $\mathcal{P}_t$ of image-text pairs. We propose and study two strategies to build $\mathcal{P}_t$:

1. **Semantic filtering.** From the same general evidence used for domain inference, a language model also returns a small set of search queries Q that describe the domain. We embed d and each query in Q with a text encoder and issue them against a semantic index of the corpus captions, keeping the top-ranked images for every query.
2. **Lexical filtering.** We first enumerate a candidate roster $\mathcal{R}$ of class names. To ground $\mathcal{R}$ in what the victim recognizes rather than in the language model's prior alone, we extract the distinctive terms of each class from its general evidence with class-based TF-IDF (Grootendorst, 2022). Treating the concatenation of $\mathcal{E}^g_\ell$ as one document, a term w is scored as $\mathrm{tf}(w, \ell) \cdot \log\big(1 + A/\mathrm{cf}(w)\big)$, where $\mathrm{tf}(w, \ell)$ is its normalized frequency in class ℓ, $\mathrm{cf}(w)$ its frequency across all classes, and A the average document length. A language model receives d, the number of observed classes, and the top terms of each class, and returns $\mathcal{R}$. Each entry of $\mathcal{R}$ is then issued as a keyword query against a lexical index of the corpus captions, and the top-ranked images for every entry are kept, so that the pool contains a balanced number of probes per candidate name and every caption explicitly mentions the name it was retrieved for.

Finally, the victim is queried with $\mathcal{P}_t$, and targeted evidence sets $\mathcal{E}^t_\ell$ are formed exactly as for $\mathcal{P}_g$. For each observed class, a language model reads its k most confident descriptions $\mathcal{E}^t_{\ell,k}$ and returns a free-text name prediction for the class. The top-1 access model provides KEEP with a natural abstention mechanism when the probing pool is poorly aligned with the victim's learned distribution, as classes that are absent or insufficiently represented in the pool are selected as the highest-probability softmax output with low frequency. Since classes absent from the pool are less likely to win the argmax, KEEP abstains whenever $|\mathcal{E}^t_\ell| < \tau$ rather than predicting from insufficient evidence. All prompts and procedure parameters are reported in Appendix A.4.

4 Experiments

We evaluate KEEP on classifiers spanning several domains and architectures, first with general-purpose web pools of image–caption pairs (Section 4.1) and then with targeted pools built from the victim's inferred domain (Section 4.2). Finally, we evaluate defenses that withhold visibility of the confidence scores (Section 4.3). Throughout, *recovery accuracy* denotes the proportion of the victim's classes for which KEEP correctly recovers the semantic attribute under evaluation, whether the exact class label or a broader attribute in a semantic hierarchy. The hierarchy is used only post hoc to quantify recovery, never by KEEP itself. Unless otherwise noted, all experiments retain the $k = 20$ highest-confidence probes per class and abstain for classes with fewer than $\tau = 3$ probes (Section 3). As a baseline we compare against CLIP-Dissect (Oikarinen & Weng, 2023), which names each output class by matching its top-activating probes against a concept set with CLIP (Radford et al., 2021); we give it the victim's ground-truth class names as the concept set, an oracle access that KEEP never has.

4.1 Victim Confidence Leaks Class Semantics under General Web Probing

Evaluation metrics. To support a controlled evaluation on ImageNet-1K (Deng et al., 2009), we apply KEEP using probing pools with varying degrees of alignment to the victim's training distribution and project the free-text predictions onto closed vocabularies derived from the ImageNet labels and WordNet hierarchy (Miller, 1994) by nearest-neighbor cosine-similarity matching with

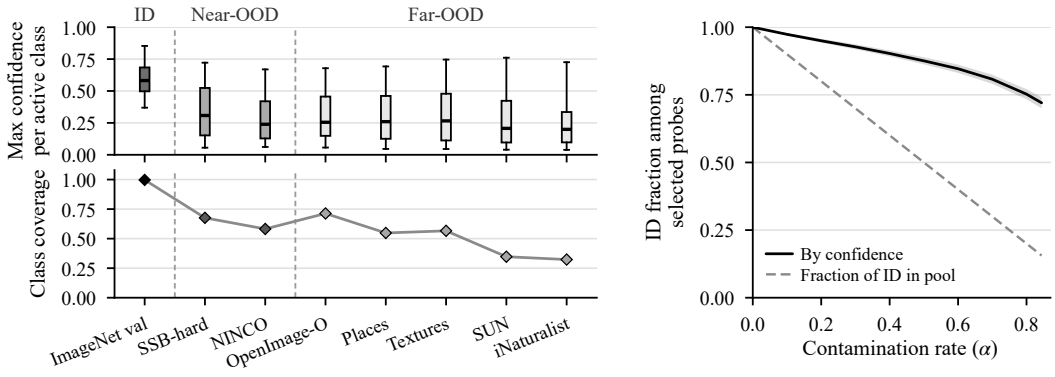


Figure 2: **Relevant probes elicit stronger responses and broader class coverage, preserving the purity of selected evidence under contamination.** We probe an ImageNet-1K (Deng et al., 2009) ResNet-50 (He et al., 2016) with the OpenOOD v1.5 pools (Yang et al., 2022; Zhang et al., 2023): near-OOD SSB-hard (Vaze et al., 2021) and NINCO (Bitterwolf et al., 2023), and far-OOD OpenImage-O (Wang et al., 2022), Places (Zhou et al., 2017), Textures (Cimpoi et al., 2014), SUN (Xiao et al., 2010), and iNaturalist (Van Horn et al., 2018). *Left:* per pool, the maximum confidence per activated class (top) and the fraction of the $1{,}000$ classes predicted as top-1 for at least one probe (bottom). *Right:* mean per-class ID fraction among the top-5 probes selected by confidence, for an ID base of up to 20 validation images per class contaminated with images from the seven OOD pools at rate $\alpha = N_{\mathrm{OOD}}/(N_{\mathrm{ID}} + N_{\mathrm{OOD}})$.

Qwen3-Embedding-0.6B (Zhang et al., 2025). We report cosine similarity between the raw prediction and the true class name, exact accuracy after projection, and two hierarchical metrics: *lexname accuracy*, the fraction of classes for which the prediction and true label share a WordNet lexicographer category (e.g., `noun.animal`) (Miller, 1994); and *IC-recovery*, the information content of their most specific shared concept (Resnik, 1995), normalized by that of the true class. Information content is computed intrinsically from the hierarchy (Seco et al., 2004).

Softmax confidence filters relevant evidence from contaminated pools, leaking class information through general web captions. In-distribution samples elicit stronger responses and activate a wider range of classes than even visually aligned out-of-distribution datasets (Figure 2, left). Softmax confidence thus acts as a natural in-distribution discriminator, as selecting each class's highest-confidence images preferentially retains in-distribution evidence, while OOD images tend to concentrate on fewer, visually similar classes. As we progressively contaminate an in-distribution pool with OOD images, 72% of the selected probes remain in-distribution even when in-distribution images constitute only 16% of the pool ($\alpha = 0.84$; Figure 2, right), so no clean in-distribution set is required. In Table 1 we exploit this to recover class semantics from an ImageNet-1K victim using three general web datasets of image–caption pairs, with LAION (Schuhmann et al., 2022) and CC12M (Changpinyo et al., 2021) slightly ahead of WIT (Srinivasan et al., 2021): exact-match accuracy reaches at most 0.49 against ImageNet's own classes, while lexname accuracy stays close to 80% and IC-recovery at 0.59 ± 0.02 over the whole universe of WordNet nouns, all substantially above chance. Figure 3 repeats this on two additional hierarchical victims, iNaturalist (Van Horn et al., 2018) and FGVC-Aircraft (Maji et al., 2013). CLIP-Dissect outperforms KEEP on ImageNet, where CLIP has demonstrated strong zero-shot performance (Radford et al., 2021), but falls behind on finer, more specialized domains, where captions provide KEEP with explicit semantic cues that are difficult to infer from visual content alone. On all three victims accuracy increases at coarser levels of the hierarchy, so even failures land in a similar semantic grouping, which could provide an attacker with useful information about the domain.

Recovery is data-efficient and improves most with confidence-based probe selection and stronger namers. We ablate the probing-pool size, the number of probes retained per class, the probe-selection strategy, and the namer (Figure 4). Recovery is relatively data-efficient: expanding the LAION pool from 1k to 1M images primarily sharpens fine-grained predictions, with the smallest pool already achieving a lexname accuracy of 0.73 (first panel), and a single caption per class achieves 59% of the accuracy obtained with the default $k = 20$ (second panel). Probe selection

Probe	Free-text	ImageNet vocabulary			WordNet vocabulary		
	Cosine	Accuracy	Lexname	IC	Accuracy	Lexname	IC
ImageNet-1K	0.96 ± 0.01	0.88	0.98	0.94 ± 0.01	0.77	0.89	0.84 ± 0.02
ImageNet-21K	0.91 ± 0.01	0.73	0.95	0.84 ± 0.02	0.57	0.89	0.78 ± 0.02
ImageNet-20K	0.76 ± 0.01	0.28	0.87	0.57 ± 0.02	0.08	0.83	0.56 ± 0.02
LAION	0.82 ± 0.01	0.49	0.92	0.70 ± 0.02	0.32	0.78	0.59 ± 0.02
CC12M	0.82 ± 0.01	0.49	0.90	0.70 ± 0.02	0.32	0.75	0.58 ± 0.03
WIT	0.80 ± 0.01	0.44	0.88	0.64 ± 0.02	0.29	0.70	0.53 ± 0.03
Chance	~ 0.58	10^{-3}	0.43	0.18 ± 0.01	$\sim 10^{-5}$	0.12	0.07 ± 0.01

Table 1: **General web captions recover the semantic meaning of ImageNet's classes and their coarse groupings.** Rows are probing pools: the ImageNet-1K validation set (Deng et al., 2009) as an in-distribution reference, ImageNet-21K (Ridnik et al., 2021), ImageNet-20K (ImageNet-21K without the ImageNet-1K classes), and three web image–caption collections, LAION (Schuhmann et al., 2022), CC12M (Changpinyo et al., 2021), and WIT (Srinivasan et al., 2021). Columns report the free-text cosine similarity and, after projection onto the $1{,}000$ ImageNet classes and $82{,}115$ WordNet noun synsets (Miller, 1994), exact accuracy, lexname accuracy, and IC-recovery (Section 4.1). For the free-text cosine, *Chance* is a random WordNet noun offered as the prediction.

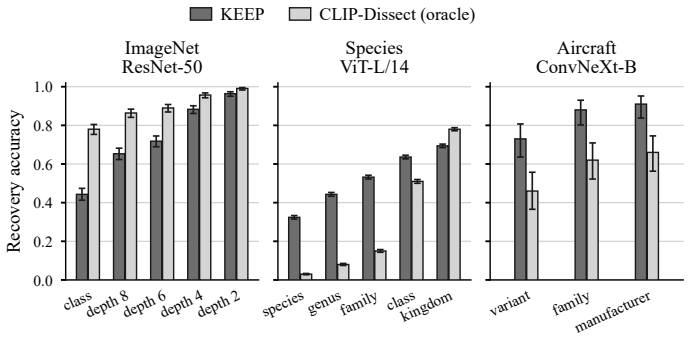


Figure 3: **On hierarchical victims, KEEP's accuracy increases at coarser class levels, and outperforms CLIP-Dissect on fine-grained specialized domains.** Recovery accuracy across class hierarchies using the same WIT-1M probing pool (Srinivasan et al., 2021): ResNet-50 (He et al., 2016) on ImageNet-1K (Deng et al., 2009), ViT-L/14 (Dosovitskiy et al., 2021; Cherti et al., 2023) on iNaturalist (Van Horn et al., 2018), and ConvNeXt-Base (Liu et al., 2022) on FGVC-Aircraft (Maji et al., 2013). CLIP-Dissect (Oikarinen & Weng, 2023) uses the ground-truth vocabulary, also used to project KEEP's ImageNet predictions. ImageNet levels correspond to shared WordNet (Miller, 1994) ancestors at different levels of the hierarchy.

has a substantially larger effect, with exact accuracy increasing from 0.11 to 0.49 when retaining each class's highest-confidence probes rather than random or lowest-confidence probes, a $4.3\times$ improvement (third panel), confirming that confidence identifies the probes carrying the strongest recoverable signal. Finally, accuracy increases monotonically with namer size, reaching 0.49 with Qwen2.5-14B (fourth panel), suggesting that KEEP can directly benefit as stronger language models are used.

4.2 Inferred Domain Knowledge Enables Class Recovery from Specialized Victims

We evaluate KEEP on fourteen classifiers spanning food dishes (Swin-B; Liu et al., 2021) (Bossard et al., 2014), dog breeds (ConvNeXt-T; Liu et al., 2022) (Khosla et al., 2011), flower species (ViT-B/16; Dosovitskiy et al., 2021) (Nilsback & Zisserman, 2008), pet breeds (ViT-B/16) (Parkhi et al., 2012), sneaker models (ViT-B/16) (Iakubovskyi, 2025), aircraft variants (ConvNeXt-B) (Maji et al., 2013), bird species (ViT-B/16) (Wah et al., 2011), scene categories (ResNet-50; He et al., 2016) (Zhou et al., 2017), brand logos (Swin-T) (Suarez Bonilla, 2023), car models

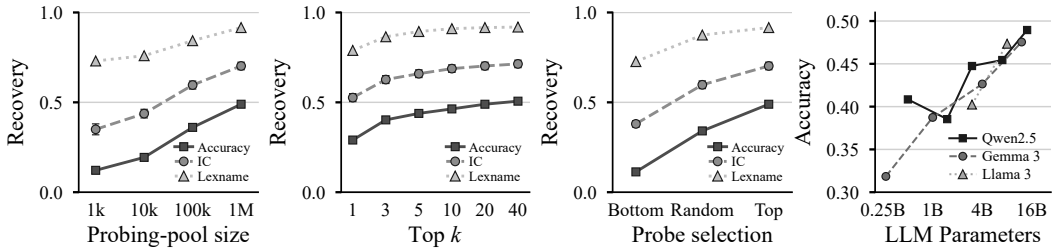


Figure 4: **Confidence-based probe selection drives recovery.** Ablations of the method on the ImageNet-1K victim, varying the LAION (Schuhmann et al., 2022) probing-pool size (first panel, from left to right), the number of retained probes k (second), whether each class is named from its k highest-confidence, random, or lowest-confidence probes among the images that activate it (third), and the language model used for naming (fourth). Unless otherwise varied, experiments use 1M LAION samples, $k = 20$, and Qwen2.5-14B. We report exact accuracy, IC recovery, and lexname accuracy under the ImageNet vocabulary (Section 4.1).

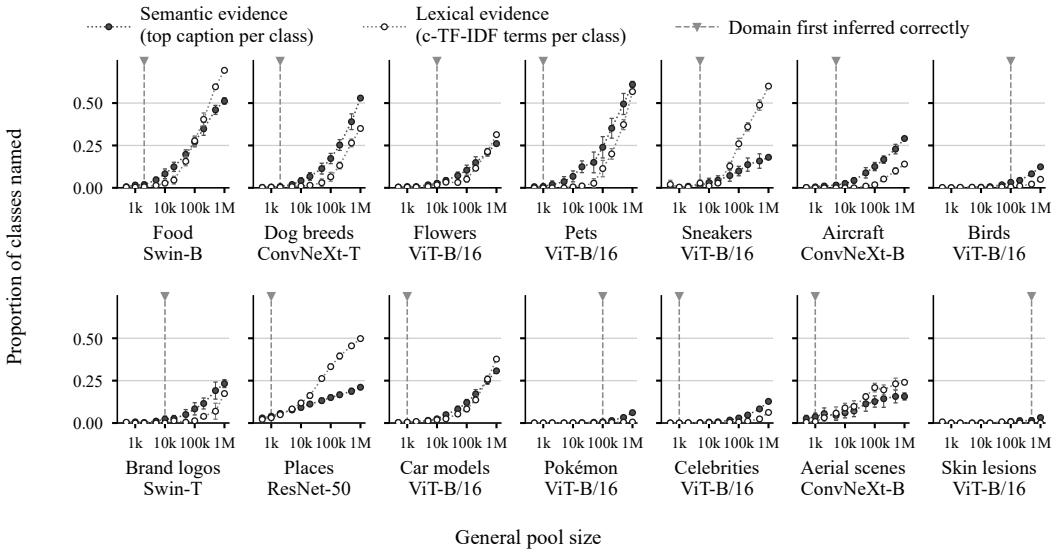


Figure 5: **The victim's domain can be inferred from very few probes, even while the evidence naming its classes is still sparse.** For each pool size, we draw that many probes from the general pool (500k LAION + 500k WIT) and report the proportion of the victim's classes named by their evidence: the most confident caption of each class, $\mathcal{E}^g_{\ell,1}$ (*semantic*), or its three class-based TF-IDF terms (*lexical*), the inputs of the two filtering strategies. A class counts if the caption contains its full name, or if a term contains the name or is contained in it. Dashed rules mark the budget from which the domain is inferred (Table 7); both measures are lower bounds, as evidence can depict a class without naming it.

(ViT-B/16) (Iakubovskyi, 2024), Pokémon (ViT-B/16) (Imjeffhi, 2022), celebrity identities (ViT-B/16) (Assi, 2024), aerial scenes (ConvNeXt-B) (Cheng et al., 2017), and skin lesions (ViT-B/16) (Shindel, 2026). Note that several of these vocabularies are sensitive in their own right: Pokémon characters, brand logos, and sneaker models are intellectual property, celebrity identities are personal data, and skin conditions are medical diagnoses. Unless otherwise noted, the general pool $\mathcal{P}_g$ is a fixed mix of 500k LAION (Schuhmann et al., 2022) and 500k WIT (Srinivasan et al., 2021) image-text pairs, shared by all victims. GPT-5.6 (OpenAI, 2026) infers the domain d, the queries Q, and the roster $\mathcal{R}$ from the general evidence; semantic filtering retrieves by nearest neighbor from a FAISS (Johnson et al., 2019) inverted-file index of Qwen3-Embedding-0.6B (Zhang et al., 2025) caption embeddings, and lexical filtering from a BM25 (Robertson & Zaragoza, 2009) index, both indexing only the captions of 100M LAION and 6.5M WIT image-text pairs; the images

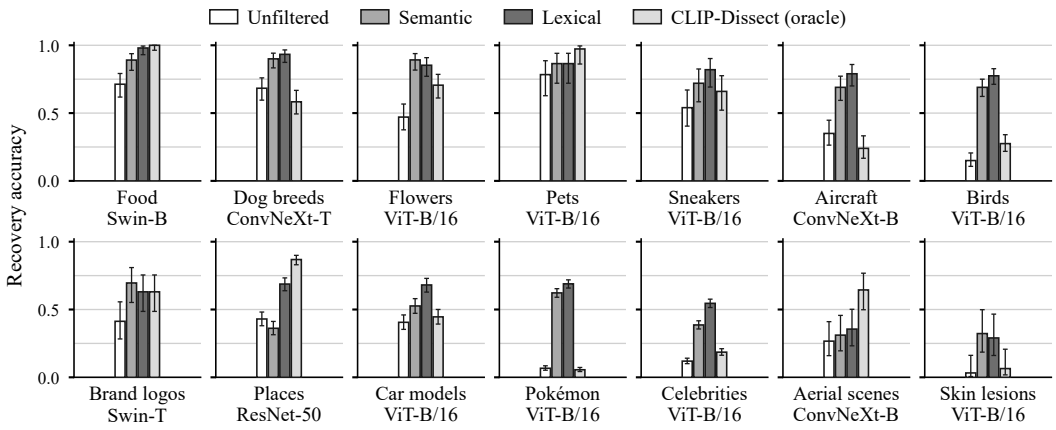


Figure 6: **KEEP recovers the exact class names of fourteen classifiers across architectures and specialized domains, including copyrighted characters, trademarked brands, and personal identities.** Recovery over each victim's complete output space, with abstentions counted as errors. The unfiltered pool is the one million images (500k LAION, 500k WIT) from which the domain is inferred, named directly; the semantic and lexical pools are built from that domain by filtering the same LAION and WIT captions (Section 3, Table 2). CLIP-Dissect (Oikarinen & Weng, 2023) runs as published on the unfiltered pool with the ground-truth class vocabulary as its concept set, since its native concept set cannot name these classes.

behind the retrieved captions are then fetched to form $\mathcal{P}_t$. Qwen2.5-14B (Yang et al., 2025) names each class from its $k = 20$ most confident captions and abstains below $\tau = 3$ probes. Prompts, retrieval depths, and pool sizes are given in Appendix A.4.

Domain knowledge enables targeted pools that increase recovery. While general-purpose pools provide little evidence about specific classes, particularly when only a few probes are available, KEEP's ability to surface in-distribution evidence (Section 4.1) is sufficient to recover the victim's domain: for eleven of our fourteen victims, the domain is inferred correctly from 10k probes or fewer (Figure 5). Targeted pools built from the inferred domain improve recovery over the unfiltered pool on all fourteen victims. With these pools, KEEP also outperforms CLIP-Dissect (Oikarinen & Weng, 2023) on ten victims, despite the baseline's oracle access to the class vocabulary (Figure 6). Lexical filtering is generally more effective than semantic filtering, achieving higher recovery on ten of the fourteen victims and matching it on one. It also achieves higher coverage on eleven victims while probing a smaller pool on thirteen (Table 2). This indicates that the method can operate efficiently with few probes, which is particularly important in API settings where calls may be rate limited. In Appendix A.2 we further show that the full attack, started from only 10k general probes, stays within a few points of its recovery with 1M probes for most victims. While recovery remains limited in domains that are genuinely scarce in web pools, such as medical imaging (Figure 6, Table 2), we emphasize that KEEP correctly recovers classes from intellectual-property and trademarked domains, where leakage has direct consequences and relevant content is readily available in general web pools: 69% of Pokémon characters, 70% of brand logos, and 82% of sneaker models.

The availability of the victim's domain in web pools limits recovery. We analyze our two weakest victims, which illustrate two distinct ways in which targeted web pools fail. Skin lesions are highly specialized medical images that are sparse on the web, whereas aerial scenes are often named on the web, but mostly in captions of ground-level photographs rather than the overhead satellite tiles the classifier was trained on. Both victims show the lowest semantic and lexical coverage and the two smallest lexical pools, with only 5.4k and 3.0k images (Table 2). In both cases, however, the domain is still inferred correctly, allowing an attacker to replace $\mathcal{P}_t$ with a handpicked selection of datasets from that domain. We test this by building $\mathcal{P}_t$ from public datasets of each inferred domain, choosing datasets that are distinct from the victim's training set and overlap minimally with its images. With these pools, recovery rises to 51.6% for skin lesions and 46.7% for aerial scenes (Table 3). A correctly inferred domain therefore lets the attacker compensate with specialized pools where general web pools fall short.

		Food	Dogs	Flowers	Pets	Sneakers	Aircraft	Birds	Places	Logos	Cars	Pokémon	Celebrities	Aerial	Skin
Classes		101	120	102	37	50	100	200	365	46	323	898	997	45	31
Unfiltered	Pool	1M	1M	1M	1M	1M	1M	1M	1M	1M	1M	1M	1M	1M	1M
	Cov. (%)	95.0	90.8	74.5	**100.0**	74.0	**97.0**	62.5	97.8	**100.0**	96.9	42.7	87.9	88.9	48.4
Semantic	Pool	116k	88k	77k	85k	92k	88k	75k	109k	82k	100k	68k	99k	84k	89k
	Cov. (%)	98.0	**100.0**	89.2	**100.0**	90.0	95.0	91.0	85.2	89.1	97.2	90.2	91.7	80.0	**67.7**
Lexical	Pool	50k	16k	25k	17k	13k	41k	60k	36k	92k	70k	11k	88k	3k	5k
	Cov. (%)	**99.0**	99.2	**96.1**	**100.0**	**92.0**	**97.0**	**91.5**	**99.5**	97.8	**98.1**	**95.9**	**96.1**	**91.1**	64.5

Table 2: **Lexical pools that are several times smaller than semantic pools name more of the victim's classes.** For each victim, the unfiltered 1M general pool and the semantic and lexical pools built from its inferred domain: the number of images probed (Pool) and the share of the victim's classes named in at least one caption (Coverage). Bold marks the best pool per victim, ties included.

Victim	$\mathcal{P}_t$	Images	Recovery (%)
Skin lesions	LAION + WIT (semantic filtering)	89.3k	32.3
	LAION + WIT (lexical filtering)	5.4k	29.0
	SkinCAP + Fitzpatrick17k + DermNet + ISIC 2019	52.8k	**51.6**
Aerial scenes	LAION + WIT (semantic filtering)	83.8k	31.1
	LAION + WIT (lexical filtering)	3.0k	35.6
	PatternNet + AID	40.4k	**46.7**

Table 3: **Public datasets chosen from the inferred domain raise recovery for victims underrepresented on the web.** Recovery of the two weakest victims, scored as in Figure 6. The third row takes $\mathcal{P}_t$ to be public datasets of the inferred domain in the victim's own imaging modality: clinical photographs and dermoscopy for skin lesions (SkinCAP (Shen et al., 2024), Fitzpatrick17k (Groh et al., 2021), DermNet (DermNet, 2020), and ISIC 2019 (Hernández-Pérez et al., 2024; Tschandl et al., 2018; Codella et al., 2018)), and Google Earth tiles for aerial scenes (PatternNet (Zhou et al., 2018) and AID (Xia et al., 2017)), with each dataset's caption or label as text. No pool is a victim's training set.

4.3 Defending Against the Attack by Withholding Confidence Scores

Returning only the label degrades class naming and hinders domain inference for most victims. Since KEEP selects its evidence by confidence, a natural defense is an API that returns only the opaque label, which we simulate by describing each class with random probes among those assigned to it instead of its most confident ones. On the ImageNet-1K victim, exact accuracy only decreases from 0.49 to 0.34 and lexname accuracy from 0.92 to 0.88 (Figure 4, third panel), as the label alone still identifies the probes of each class. Domain inference is affected far more: without the confidence score, the domain is inferred reliably for only four of the fourteen victims. As targeted pools are built from the inferred domain, withholding the score disrupts the attack on specialized victims, while milder restrictions, such as rounding the score or adding noise to it, leave recovery unchanged. Appendix A.3 explores this and other defense mechanisms in detail.

5 Conclusion

We have shown that a classifier's class vocabulary is a considerable source of leakage. Probing a black-box API with captioned web images, KEEP recovers the domain and a large share of the class names of classifiers across architectures and domains, with strong recovery on victims trained on copyrighted characters and trademarked brands, for which evidence is readily available in web pools. Since KEEP relies only on the top-1 label and its softmax confidence, the approach extends in principle to any classifier whose probes can be described, such as text or speech models, and automated captioning could supply descriptions for images that have none. We hope these results draw attention to a form of leakage that sample-level privacy does not capture and that carries direct consequences for models trained on sensitive data.

AI USE STATEMENT

Large language models were used as components of the proposed methodology to generate free-text class labels, infer classifier domains, and produce semantic and lexical search queries, as described in the paper. Generative AI tools were also used to assist with writing and debugging research code and to improve the manuscript's grammar, clarity, and structure. All AI-assisted outputs were reviewed and verified by the authors, including testing AI-assisted code for correctness. The authors take full responsibility for the paper's content, claims, and results.

ETHICS STATEMENT

While inferring information from black-box image classifiers may support model auditing and reveal information unintentionally exposed by prediction APIs, it could also be misused to recover proprietary class information or sensitive identities. Our experiments use research models and publicly available datasets; we do not query production systems or third-party commercial APIs. We present these findings to highlight this potential attack surface and motivate stronger protections for deployed classifiers, particularly those operating on sensitive data.

REPRODUCIBILITY STATEMENT

The supplementary material contains the full pipeline, the pool and index builders, and the saved results behind every table and figure, together with scripts that recompute the reported numbers and check them against this paper, which requires no GPU, no API access, and no downloads. Section 3 describes KEEP, Section 4 the settings shared by all experiments, and Appendix A.4 every prompt and retrieval setting; the inferred domains and rosters that drive the targeted pools are listed in Appendix A.2. The probe pools are built from public caption corpora and public datasets whose images we do not redistribute, so the supplement ships the builders that reconstruct each pool, the size of every pool used, and the captions each class was named from. All victims are public checkpoints, identified in the supplement, and per-victim rules and per-class recovery verdicts are included.

REFERENCES

Shengwei An, Guanhong Tao, Qiuling Xu, Yingqi Liu, Guangyu Shen, Yuan Yao, Jingwei Xu, and X. Zhang. Mirror: Model inversion for deep learningnetwork with high fidelity. *Proceedings 2022 Network and Distributed System Security Symposium*, 2022. URL https://api.semanticscholar.org/CorpusID:247313542.

Tony Assi. tonyassi/celebrity-classifier: A fine-tuned ViT-base celebrity classifier. https://huggingface.co/tonyassi/celebrity-classifier, 2024. Accessed July 22, 2026.

Julian Bitterwolf, Maximilian Müller, and Matthias Hein. In or out? Fixing ImageNet out-of-distribution detection evaluation. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2471–2506. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/bitterwolf23a.html.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1897–1914, 2022. doi: 10.1109/SP46214.2022.9833649.

Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 5253–5270, Anaheim, CA, August 2023. USENIX Association. ISBN 978-1-939133-37-3. URL https://www.usenix.org/conference/usenixsecurity23/presentation/carlini.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3558–3568, June 2021.

Si Chen, Mostafa Kahla, Ruoxi Jia, and Guo-Jun Qi. Knowledge-enriched distributional model inversion attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16178–16187, October 2021.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017. doi: 10.1109/JPROC.2017.2675998.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2829, 2023.

Christopher A. Choquette-Choo, Florian Tramer, Nicholas Carlini, and Nicolas Papernot. Label-only membership inference attacks. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 1964–1974. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/choquette-choo21a.html.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

Noel C. F. Codella, David Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). In *IEEE 15th International Symposium on Biomedical Imaging (ISBI)*, pp. 168–172, 2018.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

DermNet. Dermnet: Image data for 23 categories of skin diseases. https://www.kaggle.com/datasets/shubhamgoel27/dermnet, 2020. Images from the DermNet dermatology atlas, https://dermnet.com. Accessed September 2026.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, Feb 2017. ISSN 1476-4687. doi: 10.1038/nature21056. URL https://doi.org/10.1038/nature21056.

Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, CCS '15, pp. 1322–1333, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450338325. doi: 10.1145/2810103.2813677. URL https://doi.org/10.1145/2810103.2813677.

Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. Evaluating deep neural networks trained on clinical images in dermatology with the Fitzpatrick 17k dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1820–1828, 2021.

Maarten R. Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *ArXiv*, abs/2203.05794, 2022. URL https://api.semanticscholar.org/CorpusID:247411231.

Yuechun Gu and Keke Chen. Gan-based domain inference attack. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i12.26663. URL https://doi.org/10.1609/aaai.v37i12.26663.

Yuechun Gu, Jiajie He, and Keke Chen. Adaptive domain inference attack with concept hierarchy. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, KDD '25, pp. 413–424, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400712456. doi: 10.1145/3690624.3709332. URL https://doi.org/10.1145/3690624.3709332.

Gyojin Han, Jaehyun Choi, Haeil Lee, and Junmo Kim. Reinforcement learning-based black-box model inversion attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20504–20513, June 2023.

Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. Logan: Membership inference attacks against generative models. *Proceedings on Privacy Enhancing Technologies*, 2019: 133 – 152, 2017. URL https://api.semanticscholar.org/CorpusID:52211986.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *ArXiv*, abs/1610.02136, 2016. URL https://api.semanticscholar.org/CorpusID:13046179.

Carlos Hernández-Pérez, Marc Combalia, Sebastian Podlipnik, Noel C. F. Codella, Veronica Rotemberg, Allan C. Halpern, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Brian Helba, Susana Puig, Veronica Vilaplana, and Josep Malvehy. BCN20000: Dermoscopic lesions in the wild. *Scientific Data*, 11:641, 2024. doi: 10.1038/s41597-024-03387-w.

Dmytro Iakubovskyi. dima806/car_models_image_detection: A fine-tuned ViT-base car-model classifier. https://huggingface.co/dima806/car_models_image_detection, 2024. Accessed September 25, 2026.

Dmytro Iakubovskyi. dima806/popular_sneakers_detection: A fine-tuned ViT-base sneaker classifier. https://huggingface.co/dima806/popular_sneakers_detection, 2025. Accessed September 18, 2026.

Imjeffhi. imjeffhi/pokemon_classifier: A fine-tuned ViT-base Pokémon classifier. https://huggingface.co/imjeffhi/pokemon_classifier, 2022. Accessed July 22, 2026.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

Mostafa Kahla, Si Chen, Hoang Anh Just, and Ruoxi Jia. Label-only model inversion attacks via boundary repulsion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15045–15053, June 2022.

Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, June 2011.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows . In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9992–10002, Los Alamitos, CA, USA, October 2021. IEEE Computer Society. doi: 10.1109/ICCV48922.2021.00986. URL https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.00986.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11976–11986, June 2022.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew B. Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *CoRR*, abs/1306.5151, 2013. URL http://arxiv.org/abs/1306.5151.

Tomoya Matsumoto, Takayuki Miura, and Naoto Yanai. Membership inference attacks against diffusion models. In *2023 IEEE Security and Privacy Workshops (SPW)*, pp. 77–83, 05 2023. doi: 10.1109/SPW59333.2023.00013.

George A. Miller. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*, 1994. URL https://aclanthology.org/H94-1111/.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008.

Tuomas Oikarinen and Tsui-Wei Weng. CLIP-dissect: Automatic description of neuron representations in deep vision networks. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=iPWiwWHc1V.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition. https://openai.com/index/gpt-5-6/, 2026. Model gpt-5.6-sol. Accessed September 2026.

Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. Knockoff nets: Stealing functionality of black-box models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

Vinay Uday Prabhu and Abeba Birhane. Large image datasets: A pyrrhic win for computer vision? *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1536–1546, 2020. URL https://api.semanticscholar.org/CorpusID:220265500.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'95, pp. 448–453, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. ISBN 1558603638.

Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. URL https://openreview.net/forum?id=Zkj_VcZ6ol.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389, April 2009. ISSN 1554-0669. doi: 10.1561/1500000019. URL https://doi.org/10.1561/1500000019.

A. Salem, Yang Zhang, Mathias Humbert, Mario Fritz, and Michael Backes. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. *ArXiv*, abs/1806.01246, 2018. URL https://api.semanticscholar.org/CorpusID:46933970.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: an open large-scale dataset for training next generation image-text models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Nuno Seco, Tony Veale, and Jer Hayes. An intrinsic information content metric for semantic similarity in wordnet. In *European Conference on Artificial Intelligence*, 2004. URL https://api.semanticscholar.org/CorpusID:12036832.

Yuhao Shen, Liyuan Sun, Yan Xu, Wenbin Liu, Shuping Zhang, Shawn Afvari, Zhongyi Han, Jiaoyan Song, Yongzhi Ji, Tao Lu, Xiaonan He, Xin Gao, and Juexiao Zhou. SkinCaRe: A multimodal dermatology dataset annotated with medical caption and chain-of-thought reasoning. *arXiv preprint arXiv:2405.18004*, 2024. SkinCAP is the captioning subset, released at https://huggingface.co/datasets/joshuachou/SkinCAP.

Aiden Shindel. aidenshindel/vit.vit-base-patch16-224-finetuned-SkinDisease: A fine-tuned ViT-base skin-disease classifier. https://huggingface.co/aidenshindel/vit.vit-base-patch16-224-finetuned-SkinDisease, 2026. Accessed September 18, 2026.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 3–18, 2017. doi: 10.1109/SP.2017.41.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, pp. 2443–2449, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380379. doi: 10.1145/3404835.3463257. URL https://doi.org/10.1145/3404835.3463257.

Lukas Struppek, Dominik Hintersdorf, Antonio De Almeida Correira, Antonia Adler, and Kristian Kersting. Plug & play attacks: Towards robust and flexible model inversion attacks. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 20522–20545. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/struppek22a.html.

Felix Suarez Bonilla. fsuarez/autotrain-logo-identifier-90194144191: A fine-tuned Swin-tiny brand-logo classifier. https://huggingface.co/fsuarez/autotrain-logo-identifier-90194144191, 2023. Accessed September 18, 2026.

Anshuman Suri and David Evans. Formalizing and estimating distribution inference risks. *ArXiv*, abs/2109.06024, 2021. URL https://api.semanticscholar.org/CorpusID:237491948.

Florian Tramèr, Fan Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction apis. In *Proceedings of the 25th USENIX Conference on Security Symposium*, SEC'16, pp. 601–618, USA, 2016. USENIX Association. ISBN 9781931971324.

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: 180161, 2018. doi: 10.1038/sdata.2018.161.

U.S. Copyright Office. Copyright and artificial intelligence, part 3: Generative AI training. Report of the register of copyrights, U.S. Copyright Office, May 2025. Pre-publication version.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. *ArXiv*, abs/2110.06207, 2021. URL https://api.semanticscholar.org/CorpusID:238634102.

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4921–4930, June 2022.

Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. AID: A Benchmark Data Set for Performance Evaluation of Aerial Scene Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981, July 2017. doi: 10.1109/TGRS.2017.2685945.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3485–3492, 2010. doi: 10.1109/CVPR.2010.5539970.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Dan Hendrycks, Yixuan Li, and Ziwei Liu. Openood: Benchmarking generalized out-of-distribution detection. 2022.

Ziqi Yang, Jiyi Zhang, Ee-Chien Chang, and Zhenkai Liang. Neural network inversion in adversarial setting via background knowledge alignment. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, CCS '19, pp. 225–240, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450367479. doi: 10.1145/3319535.3354261. URL https://doi.org/10.1145/3319535.3354261.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pp. 268–282, 2018. doi: 10.1109/CSF.2018.00027.

Sajjad Zarifzadeh, Philippe Liu, and Reza Shokri. Low-cost high-power membership inference attacks. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 58244–58282. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/zarifzadeh24a.html.

Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Yixuan Li, Ziwei Liu, Yiran Chen, and Hai Li. Openood v1.5: Enhanced benchmark for out-of-distribution detection. *arXiv preprint arXiv:2306.09301*, 2023.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *ArXiv*, abs/2506.05176, 2025. URL https://api.semanticscholar.org/CorpusID:279243736.

Yuheng Zhang, Ruoxi Jia, Hengzhi Pei, Wenxiao Wang, Bo Li, and Dawn Song. The secret revealer: Generative model-inversion attacks against deep neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

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.

Weixun Zhou, Shawn Newsam, Congmin Li, and Zhenfeng Shao. PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval. *ISPRS Journal of Photogrammetry and Remote Sensing*, 145:197–209, 2018. ISSN 0924-2716. doi: https://doi.org/10.1016/j.isprsjprs.2018.01.004. URL https://www.sciencedirect.com/science/article/pii/S0924271618300042. Deep Learning RS Data.

A Appendix

The appendix follows the order of the main text. Appendix A.1 breaks down the ImageNet results of Section 4.1 by WordNet category. Appendix A.2 reports the inferred domains, the targeted pools, and the number of probes that the domain inference and the full attack need (Section 4.2). Appendix A.3 gives the complete results of the defenses of Section 4.3. Appendix A.4 lists every prompt and the retrieval settings, and Appendix A.5 collects qualitative examples: the queries and rosters derived for every victim, the captions from which the domain is inferred, and the probes and captions behind recovered classes.

A.1 General web pools

Lexname accuracy per category. Table 4 decomposes the lexname accuracy reported in Section 4.1, Table 1 by the ground-truth lexname of each ImageNet-1K class. For every probing pool and both closed vocabularies, we report the fraction of classes within each WordNet lexical category whose free-text prediction snaps to a synset of the same category.

A.2 Domain inference and targeted pools

Inferred domains and targeted pools. Table 5 lists, for every victim of Figure 6, the domain description and the unit noun that GPT-5.6 infers from the caption of each output class's most confident probe. The description seeds the semantic and lexical pool construction of Section 4.2, and the unit noun is the answer type requested from the naming model. Table 6 gives, for the same victims, the size and coverage of the roster and of every pool, and the recovery of each pool and of their union.

Probes needed by the domain inference. Table 7 gives, for every victim, the budget of general probes from which its domain is inferred, which is the rule behind the dashed lines of Figure 5.

The full attack needs a small fraction of the queries of blind probing. The targeted pools of Section 4.2 are built from a first stage that probes the victim with the whole general pool of 1M images, so the total number of queries is dominated by domain inference. We therefore repeat the lexical attack when its first stage sees only 10k or 1k general probes, and take everything the attacker derives from that draw alone: the domain description, the unit noun, the number of active classes, and the caption terms from which the planner generates the roster (Table 8). When the domain is inferred from 10k probes, recovery stays within four points of the 1M first stage for nine of the twelve victims, with 15k to 90k queries in total, targeted pool included. It decreases for sneakers, brand logos, and celebrities, and for the latter two the roster covers fewer classes, as the caption

Vocabulary	Lexname	n	Probing pool					
			ImageNet-1K	ImageNet-21K	ImageNet-20K	LAION	CC12M	WIT
ImageNet	artifact	518	0.98	0.95	0.86	0.94	0.91	0.86
	animal	398	0.98	0.97	0.92	0.95	0.94	0.95
	food	48	1.00	0.85	0.69	0.81	0.79	0.75
	plant	15	0.73	0.73	0.47	0.33	0.27	0.47
	object	11	1.00	0.73	0.45	0.91	0.91	0.55
	communication	6	1.00	0.67	0.67	0.67	0.50	0.17
	person	3	0.67	0.33	0.33	0.33	0.33	0.00
	substance	1	1.00	1.00	0.00	1.00	1.00	1.00
	All classes	1000	0.98	0.95	0.87	0.92	0.90	0.88
WordNet	artifact	518	0.88	0.87	0.80	0.73	0.69	0.58
	animal	398	0.93	0.94	0.91	0.86	0.86	0.88
	food	48	0.79	0.73	0.71	0.73	0.73	0.65
	plant	15	0.67	0.80	0.60	0.53	0.47	0.67
	object	11	0.82	0.64	0.45	0.64	0.55	0.64
	communication	6	0.83	1.00	0.50	0.83	0.67	0.33
	person	3	0.33	0.00	0.00	0.00	0.00	0.00
	substance	1	1.00	1.00	0.00	1.00	1.00	1.00
	All classes	1000	0.89	0.89	0.83	0.78	0.75	0.70

Table 4: **Lexname accuracy per WordNet lexical category.** Breakdown of the lexname accuracy of Table 1 by the ground-truth lexname of each ImageNet-1K class. n is the number of ImageNet-1K classes in each category.

Victim classifier	Inferred domain	Unit noun
Food	prepared foods and named dishes (Food-101-style classes)	food dish
Dog breeds	Domestic dog breeds and closely related wild canids	dog breed
Flowers	flower species and horticultural flower types	flower species
Pets	domestic cat and dog breeds	pet breed
Sneakers	sneaker and athletic-footwear models	shoe model
Aircraft	fixed-wing aircraft models and variants	aircraft model or variant
Birds	wild bird species, especially North American songbirds and waterbirds	bird species
Places	indoor and outdoor scene/place types, very likely the Places365 scene-recognition taxonomy	scene category
Brand logos	brand and organization logos	logo
Car models	road vehicles distinguished by make and model	car model
Pokémon	Pokémon species and forms	Pokémon species
Celebrities	named public-figure face identities	person identity
Aerial scenes	NWPU-RESISC45 aerial and satellite land-use scene categories	remote-sensing scene class
Skin lesions	dermatologic and mucocutaneous conditions	skin condition

Table 5: **Domains inferred by GPT-5.6 under top-1 access.** Every victim's broad domain is recovered from a single general-purpose probing pool. The actual training dataset is detected for only three of the fourteen victims (Food-101, Places365, and NWPU-RESISC45).

		Roster		Pool size (images)			Recovery (%)				CLIP-D
Victim	Classes	Names	Cov. (%)	Unf.	Sem.	Lex.	Unf.	Sem.	Lex.	Union	Unf.
Food	101	1461	100	1M	116k	49.8k	71	89	98	100	100
Dog breeds	120	702	92	1M	87.6k	16.2k	68	90	93	95	58
Flowers	102	651	100	1M	76.8k	24.7k	47	89	85	94	71
Pets	37	565	100	1M	84.7k	17.2k	78	86	86	92	97
Sneakers	50	1176	94	1M	91.8k	13.5k	54	72	82	88	66
Aircraft	100	1482	98	1M	88.1k	41.1k	35	69	79	89	24
Birds	200	4057	86	1M	75.2k	60.1k	15	69	78	86	28
Places	365	585	99	1M	109k	35.9k	43	36	69	76	87
Brand logos	46	1458	89	1M	82.4k	91.9k	41	70	63	85	63
Car models	323	1616	98	1M	100k	70.1k	41	53	68	80	45
Pokémon	898	1457	100	1M	68.4k	10.7k	7	62	69	75	6
Celebrities	997	2090	80	1M	98.7k	88.1k	12	39	55	65	19
Aerial scenes	45	45	100	1M	83.8k	3.0k	27	31	36	51	64
Skin lesions	31	775	81	1M	89.3k	5.4k	3	32	29	35	6

Table 6: **The two targeted pools recover complementary classes, and the roster determines what lexical filtering targets.** Roster is the candidate vocabulary $\mathcal{R}$ generated for lexical filtering: the number of names it contains and the percentage of the victim's classes it covers; a class counts as covered if its name equals, contains, or is contained in a roster entry. Pool size is the number of images probed; the unfiltered pool is the 1M general pool shared by all victims. Recovery is scored as in Figure 6, with abstentions counted as errors, and Union counts a class as recovered when either targeted pool recovers it. CLIP-D is CLIP-Dissect (Oikarinen & Weng, 2023) with the ground-truth class vocabulary on the unfiltered pool.

	General probes								
Victim	1k	2k	5k	10k	20k	50k	100k	200k	500k
Car models	✓								
Celebrities	✓								
Places	✓								
Pets	✓								
Aerial scenes	✓								
Dog breeds	×	✓							
Food	×	✓							
Aircraft	×	×	✓						
Sneakers	×	×	✓						
Flowers	×	×	×	✓					
Brand logos	×	×	×	✓					
Pokémon	×	×	×	×	×	×	✓		
Birds	×	×	×	×	×	×	✓		
Skin lesions	×	×	×	×	×	×	×	×	✓

Table 7: **How many general probes the domain inference needs.** For each budget N, we draw N probes at random from the general pool (500k LAION + 500k WIT), take every class's most confident probe, and give the captions of these probes to the planner exactly as in Section 3. A check marks the budget from which the domain is inferred correctly, where Figure 5 draws its dashed rule; a cross marks the smaller budgets at which it is not.

terms of a small draw carry little information about an open vocabulary of names. At 1k probes the attack is unchanged for places and car models. When the domain is not inferred, the roster covers few or none of the classes and recovery falls to two percent or less, the one exception being pets, whose roster still covers two thirds of the breeds; domain inference is thus the step on which the rest of the attack depends. Two caveats apply. The targeted pools of this experiment were built from the images already downloaded for Section 4.2, so rosters with many new names find fewer of their images, and the 10k and 1k recoveries are lower bounds; this accounts for part of the drop on brand logos, celebrities, and sneakers. All three budgets are also scored by the automatic procedure alone, without the manual verification behind Figure 6, so the 1M column differs from the lexical pool of Table 6 by up to ten points.

	1M probes		10k general probes				1k general probes				Blind 1M
Victim	Cov.	Rec.	Domain	Cov.	Queries	Rec.	Domain	Cov.	Queries	Rec.	Rec.
Food	100	97	✓	100	56.1k	98	×	6	7.2k	2	71
Dog breeds	92	96	✓	96	25.7k	97	×	26	40.3k	2	68
Flowers	100	86	✓	100	14.5k	82	×	5	8.7k	2	47
Pets	100	86	✓	100	27.0k	89	×	68	98.1k	65	78
Sneakers	94	86	✓	98	22.8k	78	×	0	23.4k	0	54
Aircraft	98	78	✓	99	49.5k	82	×	0	12.9k	1	35
Birds	86	78	×	4	20.9k	0	×	2	52.6k	0	15
Places	99	79	✓	98	74.8k	82	✓	99	23.2k	83	43
Brand logos	89	61	✓	78	106k	46	✓	89	105k	57	41
Car models	98	68	✓	98	87.1k	67	✓	98	77.9k	68	41
Pokémon	100	69	✓	100	20.4k	67	×	1	59.8k	0	7
Celebrities	80	55	✓	56	80.7k	38	✓	51	82.1k	34	12
Aerial scenes	100	40	✓	100	19.4k	42	✓	29	87.1k	16	27
Skin lesions	81	29	×	0	54.1k	0	×	0	77.0k	0	3

Table 8: **Domain-aware probing recovers the classes with a small fraction of the queries of blind probing.** The lexical attack of Section 4.2 with a first stage of 1M, 10k, or 1k general probes, one draw per budget. *Domain* marks whether the domain is inferred correctly from the draw; *Cov.* is the percentage of the victim's classes covered by the roster, where a class counts as covered if its name equals, contains, or is contained in a roster entry; *Queries* is the total number of images sent to the victim, general probes plus targeted pool; *Rec.* is top-1 recovery (%) with abstentions counted as errors. *Blind 1M* is the unfiltered pool of Table 6.

Specialized pools. Table 9 completes Table 3 with every public dataset of the inferred domain alone and their successive unions.

Victim	$\mathcal{P}_t$	Images	Recovery (%)
Skin lesions	SkinCAP	4.0k	29.0
	ISIC 2019	25.3k	16.1
	DermNet atlas	19.6k	22.6
	SkinCAP + Fitzpatrick17k	7.9k	32.3
	SkinCAP + Fitzpatrick17k + ISIC 2019	33.2k	41.9
	SkinCAP + Fitzpatrick17k + ISIC 2019 + DermNet	52.8k	**51.6**
Aerial scenes	PatternNet	30.4k	37.8
	AID	10.0k	33.3
	PatternNet + AID	40.4k	**46.7**

Table 9: **Recovery rises with every public dataset added to the specialized pool.** Each public dataset alone and the successive unions that form the specialized rows of Table 3, scored identically. For skin lesions, the union recovers more classes than any single dataset: ISIC 2019 contributes the dermoscopy-only classes and the DermNet atlas the rare clinical presentations.

A.3 RESTRICTING THE CONFIDENCE RETURNED BY THE API

KEEP uses the confidence score at two points: domain inference reads, for every output class, the caption of its most confident probe, and class naming reads the captions of its $k = 20$ most confident probes (Section 3). A natural defense is therefore to withhold or degrade the confidence that the API returns. We simulate such APIs from the victim responses already collected, without issuing new queries: the opaque label of every probe is left unchanged, and only the rule that selects each class's probes changes to what the restricted API still allows.

Withholding the confidence score prevents domain inference for most victims. Under a label-only API the attacker cannot rank the probes assigned to a class, so every class contributes the caption of one probe drawn at random among them instead of that of its most confident probe. Everything else follows Section 4.2: the same general pool of 1M probes, planner, and prompt, and a description counts as the victim's domain if it names it, not if it names an adjacent or unrelated one. We repeat the random draw three times per victim (Table 10). With the confidence score, the domain of all fourteen victims is inferred from this pool (Table 5); without it, the domain is inferred in every draw for four victims, in some draws for four more, and in no draw for the remaining six, which include the Pokémon, celebrity, and skin-lesion classifiers. A failed draw does not return a vaguer description but a different one, typically a public benchmark that could have produced the captions. The victims that resist the defense are those whose classes are common in web images, so that a random probe assigned to a class still depicts it; for the others, most probes assigned to a class are unrelated to it, and the confidence score is what separates the few relevant probes from the rest (Figure 5).

Victim	Draws	Description returned under label-only access
Aircraft	✓ ✓ ✓	aircraft models and variants
Places	✓ ✓ ✓	Places365-style visual scene and place types
Car models	✓ ✓ ✓	fine-grained car models and variants
Aerial scenes	✓ ✓ ✓	named geographic landmarks and places
Flowers	× ✓ ✓	countries and national geographic identity
Food	✓ × ×	Caltech-101 general object categories
Dog breeds	✓ × ×	countries and territories depicted in geolocated images
Pets	× ✓ ×	brand and organization logos
Sneakers	× × ×	world cities and municipalities
Birds	× × ×	countries and territories in geographic image recognition
Brand logos	× × ×	describable visual textures and surface patterns
Pokémon	× × ×	ImageNet-style general visual object categories
Celebrities	× × ×	ImageNet-1K-style object, organism, and scene classes
Skin lesions	× × ×	ImageNet-style everyday objects, animals, vehicles, and scenes

Table 10: **Domain inference under a label-only API.** Each class contributes the caption of one random probe among those assigned to it, drawn three times per victim from the general pool of 1M probes; a check marks a draw whose description names the victim's domain. The description shown is that of the first draw that misses the domain or, when none misses, of the first draw. With the confidence score, the domain of every victim is inferred from the same pool (Table 5).

Class naming survives every restriction short of withholding the score, and degrades gracefully even then. We repeat the probe-selection ablation of Figure 4 on the ImageNet-1K victim (10^6 LAION probes, $k = 20$, Qwen2.5-14B) under four kinds of API (Table 11): one that rounds the confidence, one that adds Gaussian noise to it, one that returns only the label and rejects inputs below a confidence threshold, and one that returns only the label. Rounding the confidence to two decimals or to one leaves recovery unchanged, and so does noise with a standard deviation of up to 0.1, since the attacker only needs a coarse ordering of the probes. Recovery decreases only under noise that makes the score uninformative to a legitimate user: with $\sigma = 0.4$, above the confidence of almost every probe (the median over the pool is 0.18), exact accuracy is still 0.41. Returning only the label reduces exact accuracy from 0.49 to 0.34 while lexname accuracy remains at 0.88, as the label alone already selects the probes that the victim assigns to the class. Rejecting low-confidence inputs does not strengthen a label-only API but weakens it: the decision to answer is a one-bit confidence that removes unrelated probes on the attacker's behalf. Rejecting inputs below 0.3 raises

exact accuracy from 0.34 to 0.46 while 988 classes are still named, which recovers most of what withholding the score had removed; at 0.7 only 1.5% of the probes are answered and 643 classes are named, so that exact accuracy over all $1{,}000$ classes falls to 0.28.

The API returns	Probes selected	Named	Exact	Exact (all)	IC	Lexname
Label and confidence	most confident	999	0.489	0.489	0.702	0.916
Confidence rounded to 2 decimals	most confident	999	0.484 ± 0.003	0.484 ± 0.003	0.701 ± 0.002	0.918 ± 0.002
Confidence rounded to 1 decimal	most confident	999	0.494 ± 0.004	0.494 ± 0.004	0.708 ± 0.004	0.923 ± 0.005
Noisy confidence, $\sigma = 0.05$	most confident	999	0.494 ± 0.010	0.494 ± 0.010	0.707 ± 0.009	0.923 ± 0.006
Noisy confidence, $\sigma = 0.1$	most confident	999	0.489 ± 0.000	0.488 ± 0.000	0.704 ± 0.004	0.921 ± 0.006
Noisy confidence, $\sigma = 0.2$	most confident	999	0.459 ± 0.002	0.459 ± 0.002	0.684 ± 0.001	0.910 ± 0.002
Noisy confidence, $\sigma = 0.4$	most confident	999	0.405 ± 0.007	0.404 ± 0.007	0.644 ± 0.006	0.898 ± 0.003
Label only, rejects below 0.7	random among accepted	643	0.430 ± 0.003	0.277 ± 0.002	0.622 ± 0.004	0.855 ± 0.004
Label only, rejects below 0.5	random among accepted	891	0.471 ± 0.004	0.420 ± 0.004	0.673 ± 0.003	0.902 ± 0.002
Label only, rejects below 0.3	random among accepted	988	0.461 ± 0.007	0.455 ± 0.006	0.676 ± 0.004	0.898 ± 0.002
Label only, rejects below 0.1	random among accepted	998	0.396 ± 0.007	0.395 ± 0.007	0.633 ± 0.005	0.886 ± 0.010
Label only	random	999	0.341 ± 0.008	0.341 ± 0.008	0.598 ± 0.005	0.875 ± 0.006
Label and confidence	least confident	999	0.113	0.113	0.380	0.727

Table 11: **Class naming on ImageNet-1K under APIs that restrict the returned confidence.** Every row names each class from $k = 20$ probes among those assigned to it, selected by the returned confidence when one is returned and at random otherwise. *Named* is the number of classes with at least one usable probe, which decreases when the API rejects low-confidence inputs; *Exact (all)* counts the classes that could not be named as errors over all $1{,}000$ classes, and the remaining columns are computed over the named classes under the ImageNet vocabulary, as in Figure 4. Values are means over three random seeds with their standard deviation. The last row selects the least confident probes, the lower bound of Figure 4.

A.4 Implementation details

We report verbatim every prompt used in the paper. Section A.4.1 covers the naming step shared by all experiments (Section 3); Section A.4.2 covers the two additional language-model calls and the retrieval instruction used to build the targeted pools of Section 4.2. No prompt reveals the victim's class vocabulary, taxonomy, or number of classes to the language model.

A.4.1 Naming prompts

All experiments share a single query template: a short instruction followed by the k retained textual descriptions, one per line, each prefixed with "– ". The language model is queried once per class at temperature 0, and its reply is taken directly as the free-text prediction:

```
<instruction>
Captions:
- <description 1>
- <description 2>
  ...
- <description k>
```

The instruction is the only part that varies across experiments. For the ImageNet experiments of Section 4.1 it is fully generic. For the fine-grained victims probed with a general web pool, the only victim-specific information it contains is the coarse domain of the classifier (aircraft or organism), and for the species victim it additionally fixes the answer format to a scientific binomial. For the targeted-pool experiments of Section 4.2 the instruction is a fixed template whose two slots are filled automatically from the inferred domain, so no wording is chosen per victim.

ImageNet, label pools (ImageNet-1K/21K/20K). The evidence texts are the ground-truth class labels of the retained pool images, and the list header reads `Labels:` instead of `Captions:`.

```
These category labels are from images that each most
strongly activate the SAME neuron of an image classifier.
```

```
Name the single most specific concept these images show;
ignore the minority of unrelated labels.  Reply with ONLY
the name, as a short noun phrase.
```

ImageNet, caption pools (LAION, CC12M, WIT).

```
These image captions are from images that each most strongly
activate the SAME neuron of an image classifier.  Name the
single most specific concept these images show; ignore a
minority of unrelated or noisy entries.  Reply with ONLY the
name, as a short noun phrase.
```

Aircraft.

```
These image captions are from images that each most strongly
activate the SAME neuron of an aircraft classifier.  Name
the single most specific aircraft these images show; ignore
a minority of unrelated entries.  Reply with ONLY the name.
```

Species.

```
These image captions are from images that each most strongly
activate the SAME neuron of an organism classifier.  Name
the single most specific organism these images show, as a
scientific binomial 'Genus species'; ignore a minority of
unrelated entries.  Reply with ONLY the binomial name.
```

Targeted pools (all fourteen victims of Figure 6). One template for every victim and both pool constructions. The two slots <category> and <unit noun> are the *Inferred domain* and *Unit noun* columns of Table 5, produced by the domain-inference call below; the attacker never edits them. The model is Qwen2.5-14B at temperature 0 with $k = 20$ captions per class.

```
These image captions are from images that each most strongly
activate the SAME neuron of a <category> classifier.  Name
the single <unit noun> these images show; ignore a minority
of unrelated entries.  Reply with ONLY the <unit noun> name.
```

For example, the celebrity victim is queried with "… of a named public-figure face identities classifier. Name the single person identity these images show …".

A.4.2 Domain inference, roster generation, and retrieval

The targeted pools of Section 4.2 require two further language-model calls per victim, both issued once to GPT-5.6, and one fixed retrieval instruction. Their inputs are entirely attacker-derived: the captions and confidence scores returned by probing the victim with the general-purpose pool.

Domain inference. One structured call whose output is constrained to a JSON object with the five fields named in the instruction. The user message lists, for every output class observed under top-1 access, the caption of the single image that activated it most strongly; captions are truncated to 220 characters. System message:

```
An unknown image classifier has one output neuron per class.
For each neuron you are given the caption of the single web
image that MOST strongly activates it -- i.e.  one
representative example image per class.  From these
per-class example captions, infer what the classifier
recognizes and how an auditor with NO access to its class
list could build a better, more targeted pool of probe
images by searching a large web-image corpus by caption.

Return:  (1) category -- a short phrase for the kind of
object/subject it recognizes; (2) semantic_queries -- 3-6
```

```
short caption-search phrases (broad, for similarity
retrieval); (3) candidate_names -- specific proper names of
individual members of this category that likely appear in
web captions (the lexical roster; as many as you are
confident of, most prominent first); (4) unit_noun -- the
singular noun for ONE member of this category, at the
granularity the classifier distinguishes, as an auditor
would ask for it (e.g. 'car model', 'bird species', 'skin
condition', 'landmark', 'food dish'); this is the answer
TYPE, not the category phrase; (5) notes -- one sentence on
your confidence and any ambiguity.
```

User message:

```
Captions (one per neuron):
- <caption of the top image of class 1>
- <caption of the top image of class 2>
  ...
```

Roster generation. One call that turns the inferred domain into the candidate vocabulary used by lexical filtering. Besides the domain, the user message carries the number of distinct class identifiers observed under top-1 access and, for each of them, the three most distinctive caption terms (c-TF-IDF) among its top-activating probes. The reply is parsed one name per line. System message:

```
You build an EXHAUSTIVE candidate list to audit a black-box
image classifier whose class list is unknown.  You are given
its inferred domain, an estimate of how many classes it has,
and for each output neuron the most distinctive caption
terms (c-TF-IDF) from the web images that most strongly
activate it.  The terms are noisy (some are misfires) but
they reveal the vocabulary and the NAMING FORMAT the classes
are written in.  Learn that format, then enumerate as many
specific named members of the domain as you can and
OVER-GENERATE past the estimate -- extra or wrong names are
filtered later by probing the victim, so coverage matters
far more than precision.  Include members the terms do not
mention.  Output ONLY member names, one per line, each in
the canonical form implied by the domain.  No numbering, no
commentary.
```

User message:

```
Inferred domain:  <category>
Estimated number of classes:  ~<number of observed class
identifiers>
Per-neuron distinctive terms (top-3 c-TF-IDF, <N> neurons):
n0:  <term> | <term> | <term>
n1:  <term> | <term> | <term>
     ...
Enumerate the complete roster of members (aim to exceed the
estimate -- list 1000+ if the universe is that large).  One
name per line.
```

Semantic retrieval instruction. Semantic filtering embeds the inferred `category` string and each of the `semantic_queries` with Qwen3-Embedding-0.6B (Zhang et al., 2025) and retrieves their nearest caption neighbors. Every query, and every caption at index-build time, is prefixed with the same instruction:

```
Instruct:  Given a category or concept, retrieve image
captions that depict it.
Query:  <text>
```

Semantic index retrieval. The category string d is searched at top-100k and each query in Q at top-20k against FAISS indexes (inverted file with product quantisation, inner-product metric, 256 probes) of Qwen3-Embedding-0.6B embeddings of 100M LAION and 6.5M WIT captions, merged by score. The resulting pool sizes per victim are the semantic pools of Table 6.

Lexical index retrieval. Each roster entry is searched as a BM25 phrase query in a SQLite FTS5 index over the same 100M LAION and 6.5M WIT captions and its top-50 matches kept. The resulting pool sizes per victim are the lexical pools of Table 6.

A.5 QUALITATIVE EXAMPLES

Table 12 shows what the attacker derives from the general evidence of every victim, the semantic queries Q and the lexical roster $\mathcal{R}$ of Section 3, and Table 13 the captions from which the domain itself is inferred. Figures 7–13 then show, for every victim classifier, five recovered classes together with the evidence KEEP used to name them. For each class, the five numbered images are the highest-confidence probes the victim assigned to that class under top-1 access, drawn from the lexically filtered pool, and the numbered lines beneath them are their captions as retrieved (truncated where marked), the only input the naming model receives. The leftmost tile in each row (ID) is a class exemplar from the same pool: an image the victim assigned to the class whose caption names it. Classes are sampled evenly across the victim's confidence ranking among the classes recovered exactly; the celebrity panel instead shows five hand-picked public figures. Figure 14 shows ten classes of the ImageNet-1K victim, probed with the general LAION pool of Section 4.1, as it has no targeted pool. Captions are shown verbatim, including misattributed or off-target entries, which illustrate the noise the naming model must ignore.

Table 12: **Semantic queries and lexical rosters derived for every victim.** The queries Q are returned by the planner with the domain of Table 5 and retrieve the semantic pool; every name of the roster $\mathcal{R}$ is issued as a phrase query to retrieve the lexical pool. Each roster is shown by its size and five names at even spacing through the alphabetical list; its coverage is in Table 6.

Victim	Semantic queries Q	Lexical roster $\mathcal{R}$
Food	plated restaurant dish close-up; homemade recipe finished dish; traditional cuisine food plate; overhead photo of meal; dessert pastry served on plate; popular appetizer or main course	1,461 names: apple pie, crab sandwich, rice and beans, chicken tagine, zucchini noodles, …
Dog breeds	purebred dog breed portrait; dog show breed standing; dog puppy breed; terrier hound retriever shepherd dog; wolf coyote dingo dhole wild canid; zoo African wild dog	702 names: Affenpinscher, Estonian Hound, Poodle, Goldador, Atelocynus microtis, …
Flowers	close-up photograph of a flower in bloom; single flower botanical specimen; colorful garden flower macro; ornamental flowering plant; wildflower species in bloom; botanical garden flower identification	651 names: pink primrose, blue lotus, hollyhock, red camellia, zinnia, …
Pets	dog breed portrait; cat breed portrait; purebred dog standing; pedigree cat close-up; dog show breed; pet breed identification	565 names: affenpinscher, cocker spaniel, manchester terrier, thai ridgeback, tortoiseshell cat, …
Sneakers	named sneaker model product photo; Nike Adidas Converse shoe model; retro basketball sneaker colorway; running shoe model trainers; skate shoe model low top high top; athletic footwear release information	1,176 names: Adidas 4D Fusio, Asics Tiger Corsair, Nike Air Flight 89, Nike Zoom Fly 3, ZoomX Vaporfly Next% 3, …
Aircraft	aircraft spotting photo model registration; commercial airliner at airport; military aircraft in flight; private business jet aircraft; general aviation light airplane; aircraft type airline aviation photo	1,482 names: MAX 7, C-145, IAI Westwind, Curtiss P-40 Warhawk, Supermarine Spitfire, …
Birds	identified wild bird species; North American bird wildlife photo; songbird perched on branch; warbler sparrow flycatcher identification; seabird waterfowl wildlife photograph; bird common and scientific name	4,057 names: black-bellied whistling-duck, long-tailed tyrant, ruddy kingfisher, southern red bishop, black-headed duck, …

Victim	Semantic queries Q	Lexical roster $\mathcal{R}$
Places	interior of a place or room; exterior view of a building; people inside a public venue; natural landscape stock photo; urban place travel photo; aerial view of a location	585 names: airfield, florist shop/indoor, sauna, covered bridge, zoo, …
Brand logos	company logo and brand mark; corporate wordmark or emblem; product packaging with visible logo; brand logo on storefront or signage; organization and media logos; automotive badge and emblem	1,458 names: Adidas, Exxon, Charles Schwab, Volcom, Zynga, …
Car models	car make model exterior photos; used car for sale make model; new vehicle review and specifications; SUV sedan coupe model photos; car front side rear view; motor show production and concept cars	1,616 names: Abarth 124 Spider, Dodge Hornet, Lada Samara, Pontiac Grand Am, Zeekr X, …
Pokémon	Pokémon character official artwork; Pokémon Pokédex sprite icon; Pokémon card plush figure; Pokémon game screenshot character; shiny legendary Pokémon; Pokémon fan art evolution	1,457 names: Bulbasaur, Walrein, Brionne, Alolan Marowak, Totem Alolan Sandshrew, …
Celebrities	celebrity portrait headshot; actor red carpet close-up; singer press photo portrait; politician official portrait; professional athlete headshot; public figure event photo	2,090 names: Aaron Eckhart, Gavin Rossdale, Nargis Fakhri, Kate Gosselin, Zachary Knighton, …
Aerial scenes	aerial view land use; satellite image landscape; top-down drone photography; USGS orthophoto; remote sensing scene; NASA Earth observation image	45 names: airplane, dense residential, meadow, roundabout, wetland, …
Skin lesions	dermatology clinical photograph; skin lesion rash plaque nodule; pigmented skin and nail lesion; oral mucosal ulcer clinical photo; skin cancer pathology image; inflammatory skin disease close-up	775 names: Acanthosis nigricans, Contact urticaria, Hurler syndrome, Perforating granuloma annulare, Zygomycosis, …

Table 13: **What the domain inference reads.** For every victim, three of the captions given to the planner from the general pool of 1M probes, each the caption of a class's most confident probe, and the domain it returned (Table 5): the two most confident captions that name one of the victim's classes and the most confident that does not, marked *. Only part of the captions name a class, from one for skin lesions to about half for dog breeds and car models; the domain follows from their agreement rather than from any single caption.

Victim	Captions read at $N = 1$M	Inferred domain
Food	Gourmet French Macarons: 24 or 72 DessertHub Gourmet… Risi e Bisi (Risotto of peas, pancetta, and onion) The dark chocolate isn't as smooth as I'd like but the…*	prepared foods and named dishes (Food-101-style classes)
Dog breeds	Mexican Hairless The Chow Is A Well Mannered Dog Quite Good With Children… African wild dog chasing an elk in the Okavango Delta.*	Domestic dog breeds and closely related wild canids
Flowers	A bee pollinating an unidentified species of purple… Marigold African 'Calando Orange' (Tagetes Erecta L.)… Protea Cynaroides, taken at Kirstenbosch, South Africa*	flower species and horticultural flower types
Pets	Nowofundland na Krajowej Wystawie Psów Rasowych w… abyssinian cat breed sketch of a pug 17 best ideas about drawings of cats on…*	domestic cat and dog breeds
Sneakers	Shop Genuine Adidas yeezy boost 350 v2 by1604 retailers… adidas Samba OG Chaussures pour Femme: Amazon.fr:… Vans White & Burgundy Authentic Mens Trainers*	sneaker and athletic-footwear models
Aircraft	Aer Lingus Airbus A320-200 (Irish registration EI-CVB)… Eurofighter Typhoon of the Italian Air Force, Taxies to… 1314 - Lockheed C-130H Hercules - Taiwan - Air Force*	fixed-wing aircraft models and variants
Birds	Amphispiza bilineata English: A Black-throated Sparrow in… Least Auklet (Aethia pusilla) Ogrodowa Street, Lodz, Poland. Main entrance to the…*	wild bird species, especially North American songbirds and waterbirds

Victim	Captions read at $N = 1$M	Inferred domain
Places	reminds me of when my dad made an igloo for me and my… Balearic islands windmill wind mills Spain Stock Images Good Morning America Live Tickets : Teen with cancer…*	indoor and outdoor scene/place types, very likely the Places365 scene-recognition taxonomy
Brand logos	1995: James Gosling launched Java Korea Times: Samsung Galaxy Camera 2 To Be Unveiled At… Download SEOPressor WordPress Plugin For Free*	brand and organization logos
Car models	New 2019 Ferrari GTC4LUSSO For Sale In Greenwich, CT… 2016 Alfa Romeo 4C Spider Photo Pre-Owned 2006 HUMMER H1 4DR WGN OPEN TOP Sport Utility…*	road vehicles distinguished by make and model
Pokémon	Wurmple by DiegoAGM Normal Gengar sprite Stripes Toggle Clasp, Sterling Silver, 13mm - 1 piece*	Pokémon species and forms
Celebrities	Whoopi_Goldberg_at_a_NYC_No_on_Proposition_8_Rall… HOLLYWOOD, CA - MAY 23: Actor Jesse Eisenberg attends the… Paulo Uccello, Painter & Mathmetician, Florence,…*	named public-figure face identities
Aerial scenes	Beach of Sospetto in San Menaio, Italy Italiano: Vista… Take Off Newark Airport Boeing 747-400 Aerofly FS 2 Čeština: Letecký pohled na budovu národní knihovny od…*	NWPU-RESISC45 aerial and satellite land-use scene categories
Skin lesions	basal cell carcinoma Psoriasisplaque kleinflaechig* epigynum of female Rugathodes sexpunctatus*	dermatologic and mucocutaneous conditions

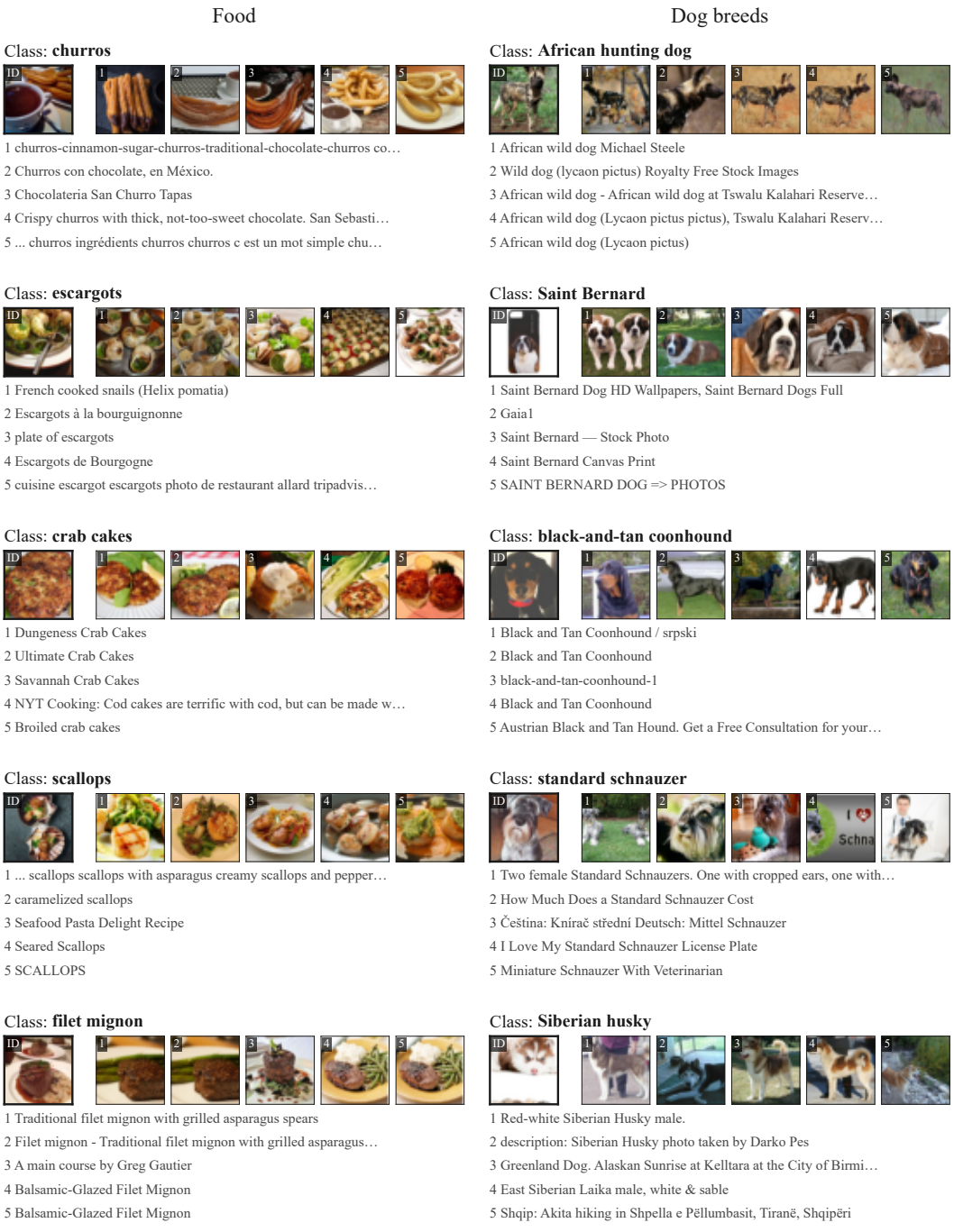


Figure 7: **Top-activating probes and captions for recovered classes: food dishes (left) and dog breeds (right).** Layout as described in Appendix A.5; every class shown was recovered exactly by KEEP.

Flowers

Class: **king protea**



1 Photo of Protea cynaroides at the San Francisco Botanical Garden
2 Photo of Protea cynaroides at the San Francisco Botanical Garden
3 King protea, Upcountry Maui
4 Protea Cynaroides, taken at Kirstenbosch, South Africa
5 Königsprotea (Protea cynaroides).

Class: **stemless gentian**



1 stemless gentian Kochscher Enzian cupă
2 gentian flower: Blue Gentian flowers Stock Photo
3 Stemless gentian (Gentiana acaulis)
4 Flowers of beautiful gentian (Gentiana boissieri), endemic gent…
5 Macro photography of Gentiana acaulis (stemless gentian) in Lju…

Class: **passion flower**



1 Passiflora caerulea. Ljubljana Botanical Garden, Slovenia. Size…
2 Picture of a Passion flower taken at Wave Hill Gardens, Riverda…
3 Passion-Flower
4 Passion flower
5 Blue passion flower-macro photo

Class: **peruvian lily**



1 Lirio del Perú (Alstroemeria), jardín del molino, Sierra de San…
2 Alstroemeria Select Everest
3 Easter Lily Photograph - Easter Lily Fine Art Print
4 The FTD® Pink Peruvian Lily Boutonniere
5 Peruvian Lily flowers Shower Curtain

Class: **pink-yellow dahlia**



1 Pink & Yellow Dahlia Flowers
2 Photo sur Toile Dahlia Pale pink dahlia
3 Dahlia 'Graceland'. Esperanto: Floro de Dalia. Français : Fleur…
4 Unique Zinnia
5 Yellow flower

Pets

Class: **Sphynx**



1 Sphynx cat
2 Sphinx cat posing for camera.
3 Cat Sphynx. Kittens.
4 Don Sphynx (DON SPHYNX) cat. Isolated on black background
5 Don Sphynx cat (DSP) cattery Pyrel *Sk

Class: **saint bernard**



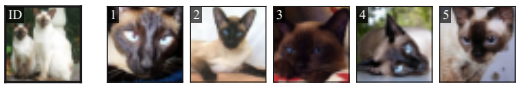
1 Saint Bernard Canvas Print
2 Saint Bernard Dog HD Wallpapers, Saint Bernard Dogs Full
3 SAINT BERNARD DOG => PHOTOS
4 Saint Bernard Phone Case
5 Baby on Saint Bernard

Class: **miniature pinscher**



1 Miniature Pinscher.
2 Miniature Pinscher.
3 Gotti, a 3 year old Miniature Pinscher with cropped ears.
4 a Miniature Pinscher
5 Doberman dog

Class: **Siamese**



1 A Siamese Cat displaying the typical blue, cross-eyed eyes typi…
2 cat siamese
3 Lando 8 Month Old Tonkinese Cute Cats And Dogs Tonkinese Cat Si…
4 images of siamese cats
5 devon rex cat

Class: **Bombay**

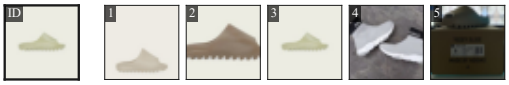


1 BOMBAY CAT
2 Black tortoiseshell cat "liliana
3 Beautiful Black CATS ( ) Rare cat breeds and Breed information…
4 Close up of a black domestic cat
5 A black cat, feral, female, in Springfield, Missouri, USADeutsc…

Figure 8: **Top-activating probes and captions for recovered classes: flowers (left) and pet breeds (right).** Layout as described in Appendix A.5; every class shown was recovered exactly by KEEP.

Sneakers

Class: **yeezy slide**



1 adidas YEEZY SLIDE KIDS "PURE" PURE/PURE/PURE 21SS-S_photo_large
2 adidas Yeezy Slide Earth Brown FV8425 Release Date Info
3 adidas - Yeezy Slide - Resin/Resin - Up There
4 Balenciaga Speed Trainer Sneakers Grey
5 Adidas Yeezy Slide ""Pure"" Size 9 (GZ5554) *IN HAND* READY TO…

Class: **nike air max plus (tn)**



1 nike air max plus zwart
2 Nike Air Max Plus Kinderschoen
3 Mizuno Wave Prophecy 7
4 mizuno wave prophecy 7 sale
5 nike air max plus blue

Class: **adidas nmd r1**



1 adidas - NMD_R1 Shoes
2 Adidas NMD_R2 - Red (CQ2404)
3 Adidas NMD R1 Primeknit 'Japan Triple Black'
4 adidas Prophere Trace Olive
5 2bf40119afdad adidas NMD R1 Primeknit Shoes - Green

Class: **nike air jordan 4**



1 air jordan 4 retro royalty
2 Air Jordan 4 Knicks
3 AIR JORDAN 4 'ROYALTY'
4 air jordan 4 royalty
5 Air Jordan 4 Retro Bred

Class: **new balance 992**



1 New Balance 992 Grey m992gr Men's Brand New Fast Ship In Hand D…
2 New Balance 992 Grey Blue Red M992AG Size 7 VNDS 100% Authentic
3 New Balance 991 - Made in England
4 Saucony Excursion Lace in Citron/Navy/Grey. #boysrunningshoes #…
5 New Balance 991 -MADE IN ENGLAND-

Aircraft

Class: **DC-6**



1 DC-7 of Servairwise
2 Japan Air Lines DC-6A "City of Nara" at the San Francisco Airpo…
3 This was one of those "collector's items" at the time and I nev…
4 Eine Douglas DC-7 der TASSA
5 Douglas DC-6B AN-BFN of LANICA at Miami Airport in 1970

Class: **DC-10**



1 N238NW, McDonnell Douglas DC-10-30ER, Northwest Airlines
2 Ariana Afghan Airlines McDonnell Douglas DC-10-30 YA-LAS
3 img082 McDonnell Douglas DC-10-10 N66NA National Airlines © Ala…
4 VARIG MD-11; PP-VPL@FRA;19.10.1994
5 A McDonnell Douglas DC-10 of Balair at Basle Airport

Class: **747-200**



1 Boeing 747-200 British Airways
2 Boeing 747-400F for sale
3 A Boeing 747 of Pan Am at Zurich Airport Deutsch: Eine Boeing 7…
4 Virgin Atlantic Airways (G-VMIA) Boeing 747-100 aircraft, Dubli…
5 El Al Boeing 747-200

Class: **A380**



1 Airbus A380
2 Airbus A380
3 Airbus A380
4 Airbus A380-841 F-WWSC, MSN124, at Airbus Finkenwerder (EDHI).…
5 Emirates A380, Airbus A380

Class: **767-200**



1 ABX Air Boeing 767-200 at Cincinnati/Northern Kentucky Internat…
2 Thai airways Boeing 747
3 We were on holding point for runway 20L in SDU airport when the…
4 Airbus A320neo Las Vegas - McCarran International (LAS / KLAS)…
5 Tampa Cargo -Boeing 767-200 Decal

Figure 9: **Top-activating probes and captions for recovered classes: sneaker models (left) and aircraft variants (right).** Layout as described in Appendix A.5; every class shown was recovered exactly by KEEP.

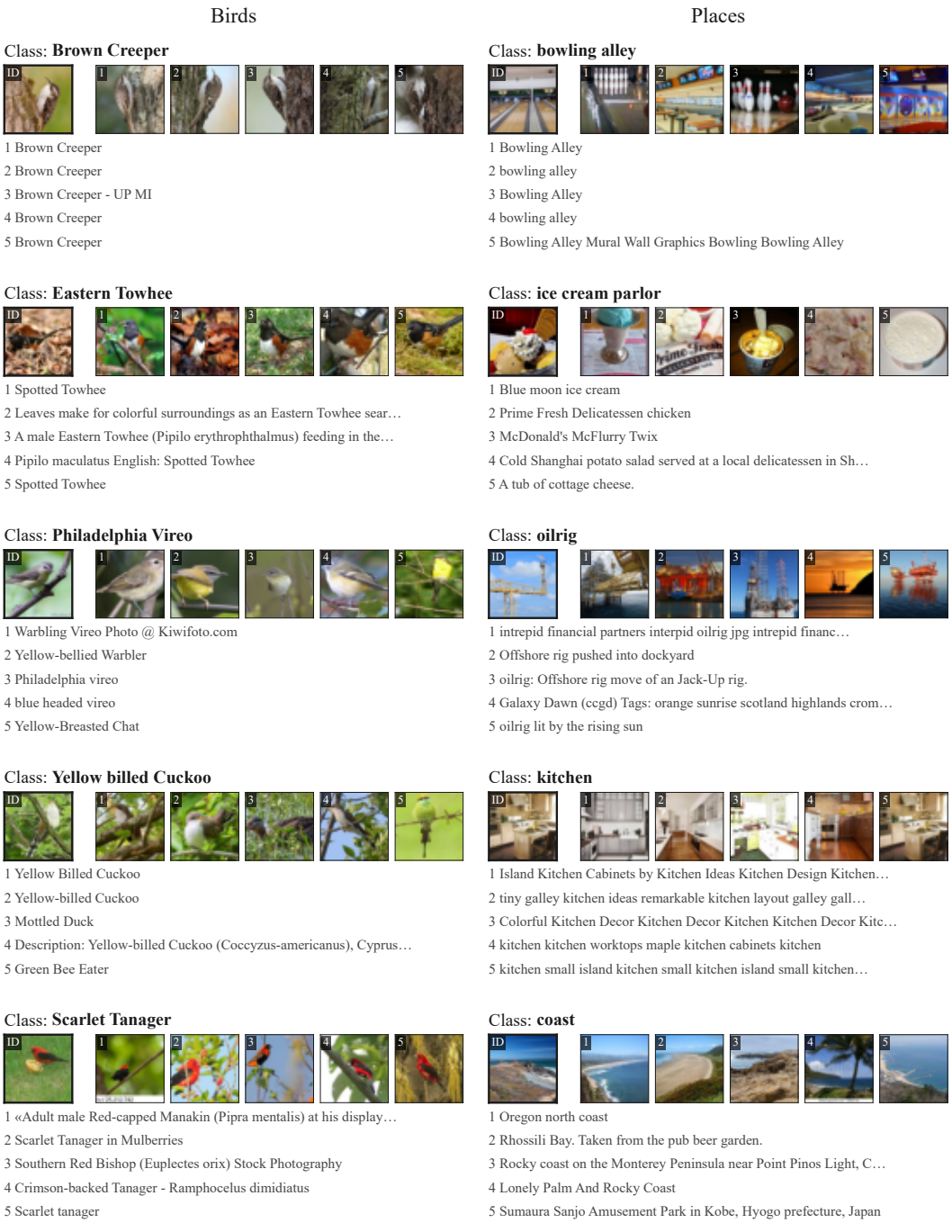
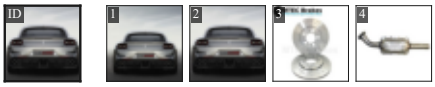


Figure 10: **Top-activating probes and captions for recovered classes: bird species (left) and scene categories (right).** Layout as described in Appendix A.5; every class shown was recovered exactly by KEEP.

Car models

Class: **Ferrari GTC4Lusso**



1 The Ferrari GTC4Lusso Is Officially Dead
2 New 2017 Ferrari GTC4LUSSO
3 Audi 90 B3 2.3 Front Drilled Grooved brake Discs 87-91
4 Mercury Topaz Catalytic ConverterCatalytic Converter

Class: **Genesis G90**



1 Genesis G90 Prices, Reviews And Pictures Us News
2 2005 Mercury Montego Premier - sister-car of the Ford Five Hund…
3 Genesis G90 2018 Review, Specs, Price
4 2017 genesis g90 priced from 68 100
5 Bentley Brooklands

Class: **Audi TT**



1 2009 Audi TT RS Coupe
2 Audi TTS service in Orange County
3 Audi Tt Rs Rear wallpaper
4 BMW X6 customised by Hamann Motorsport, pictured in London
5 Latest Audi TTS Roadster

Class: **Honda Pilot**



1 Honda Pilot Reviews
2 Citroen C4 Cactus sunroof
3 lexus vs honda pilot 2017 honda pilot pricing for sale edmunds
4 2003 Honda Pilot Review
5 Honda Passport

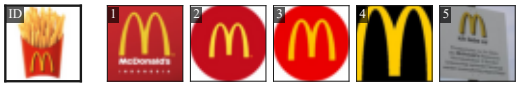
Class: **Honda Civic**



1 Honda Civic
2 Honda Civic Touring 2016
3 Honda Civic Touring 2016
4 2017 Honda Civic VTi-S sedan. Photographed in Fremantle, Wester…
5 2016 Honda Civic Touring

Brand logos

Class: **McDonalds**



1 McDonald's Indonesia
2 McDonald's background
3 McDonald's background
4 McDonald's logo
5 McDonald's Kempten

Class: **Addidas**



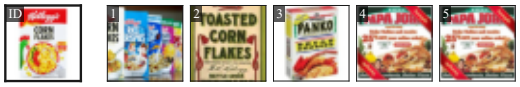
1 Adidas Performance Logo
2 ADIDAS PERFORMANCE logo
3 adidas wallpaper marble adidas wallpaper photo adidas pinterest…
4 WallpaperPhotoChaussures WallpaperPhotoChaussures Pinterest Adi…
5 Logo der Allianz-Partei

Class: **Nestle**



1 nestle-recrutement
2 Nestle.png
3 Nestle waters
4 nestle.png
5 Nestle Waters

Class: **CornFlakes**



1 Kellogg-s-cereal-products
2 Very first corn flakes package from 1906
3 kikkoman-international-inc-b26023-kikkoman-panko-bread-crumbs-j…
4 Coupon codes for papa johns
5 Papa johns discount coupons

Class: **Converse**



1 logo converse
2 Texaco Pet Sign With Images Texaco Sign Art Vintage
3 Früheres Logo von Alba Volán Székesfehérvár
4 Citizens Bank
5 IMF – International Monetary Fund Logo [imf.org]

Figure 11: **Top-activating probes and captions for recovered classes: car models (left) and brand logos (right).** Layout as described in Appendix A.5; every class shown was recovered exactly by KEEP.

Pokémon

Class: **Mareep**



1 Monster Shiny-Mareep
2 Pokemon Mareep
3 Mareep HOME.png
4 Mareep pokemon
5 Pokemon Cord Keeper! Tsunagete Pokemon - Mareep (Merriep)

Class: **Weedle**



1 Pokemon GO Weedle
2 Pokemon Weedle EMBROIDERY MACHINE FILES pattern design hus jef…
3 013: Weedle by Mabelma
4 Weedle (Grandes Encuentros TCG).png
5 Weedle (with and without outline) - Flea Circus Designs

Class: **Chimecho**



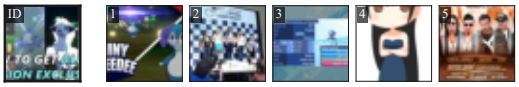
1 Pokemon Chimecho Papercraft
2 Imagen titulada Catch Uxie, Mesprit and Azelf in Pokemon Diamon…
3 Imagen titulada Catch Uxie, Mesprit and Azelf in Pokemon Diamon…
4 Plush Poke Yako Froslass japan plush

Class: **Litten**



1 product image for Pokemem Litten 24/149 Incineroar GX Cosmos Ho…
2 Peluche Pokemon Litten 20cm Pokemon Center 2016
3 Pokemon: Litten Plush
4 My version of Litten (miniature) by november-ludgate

Class: **Indeedee**



1 Indeedee - (Pokémon) - [LIVE!] SHINY 5% INDEEDEE in the Glimwoo…
2 Katy Applin on the Podium for WFU, 2014 National Championships
3 Starters Package - Popplio
4 ID: Chibi Azelf in Evening Dress by Azelf101
5 Shahram_Solati-Kamran-Hooman-Sami_Beigi-Sandy-Jamshid-Duesseldo…

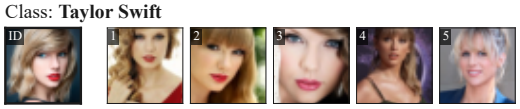
Celebrities

Class: **Barack Obama**



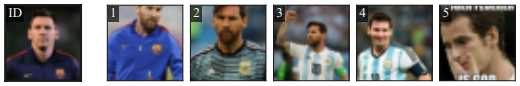
1 barack obama
2 barack-obama
3 Official Senate photo of Barack Obama (D-IL)
4 - .. English: Senator Barack Obama campaigning in New Hampshire…
5 Barack Obama Painting - Barack Obama Fine Art Print

Class: **Taylor Swift**



1 Taylor Swift taylor swift hot
2 I Taylor Swift. She is my all-time favorite singer. Go swifties…
3 Taylor Swift #CosmeticsByCortney.com #Taylor Swift #CortneyLove…
4 Taylor Swift at the 2019 American Music Awards
5 Lucy Punch

Class: **Lionel Messi**



1 Lionel Messi is the all-time record winner of the prize having…
2 Argentine footballer Lionel Messi on 26 June 2018, ahead of the…
3 [non-Latin caption]
4 lionel messi lionel messi biography childhood achievements
5 Andy Murray Meme - andy murray memes quickmeme

Class: **Oprah Winfrey**



1 oprah winfrey biography
2 Up Close: Oprah Winfrey
3 brandy norwood 1994 - photo #48
4 Oprah Winfrey
5 Up Close: Oprah Winfrey

Class: **Jackie Chan**



1 Jackie Chan Wallpapers
2 President Rodrigo Duterte Rodrigo Duterte, Historical Pictures,…
3 Jackie Chan Collection
4 Jackie Chan Picture
5 Jackie Chan Smiling

Figure 12: **Top-activating probes and captions for recovered classes: Pokémon (left) and celebrities, hand-picked (right).** Layout as described in Appendix A.5; every class shown was recovered exactly by KEEP.

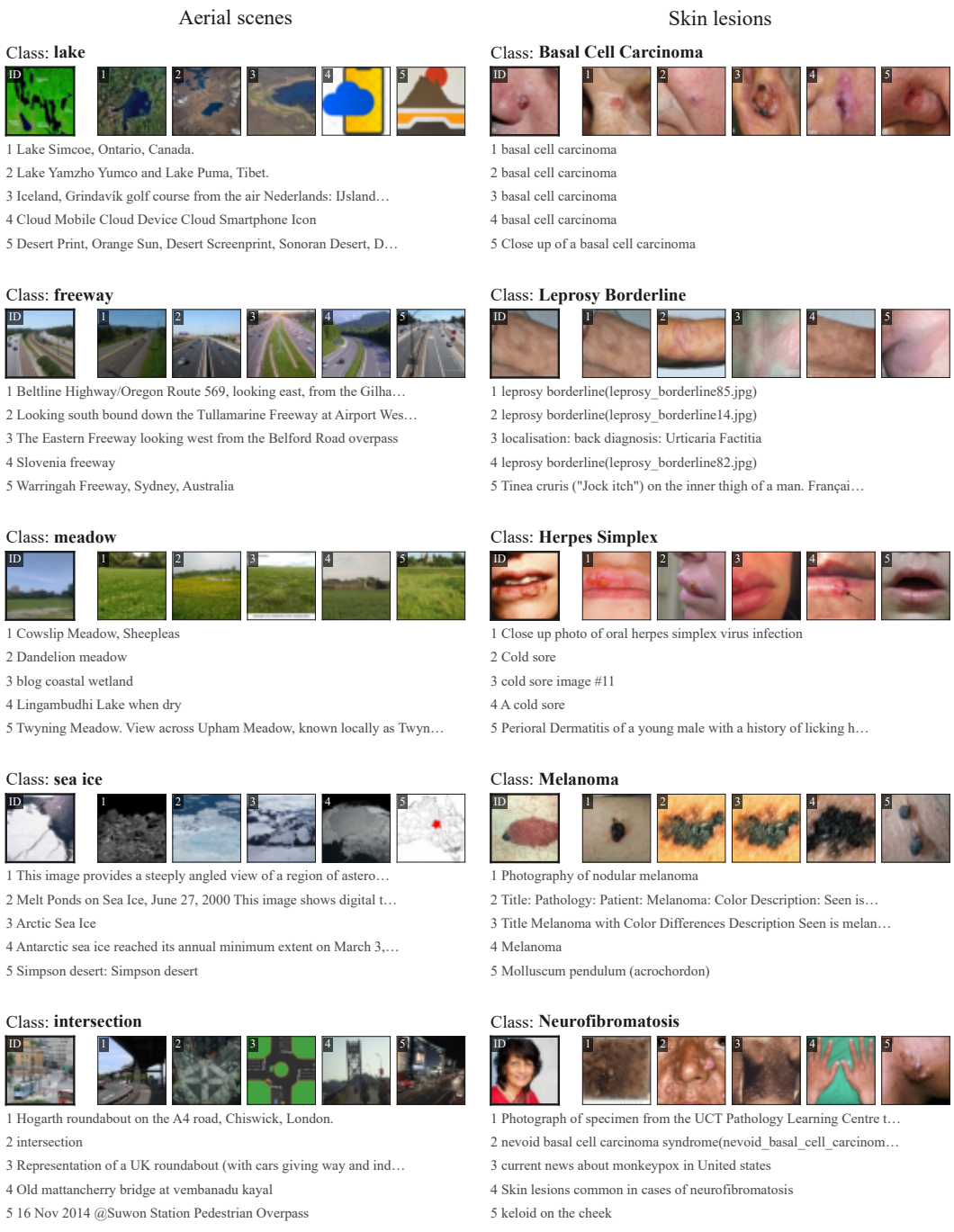


Figure 13: **Top-activating probes and captions for recovered classes: aerial scenes (left) and skin conditions (right).** Layout as described in Appendix A.5; every class shown was recovered exactly by KEEP.

ImageNet-1K, general LAION pool

Class: **jean**



1 light blue badge roll-up hem loose ankle length trousers jeans…
2 Trends to love: Overalls
3 Pinterest pin graphic that reads 'Is Madewell Ethical or Sustai…
4 2019 New Style Europe And America Men Hip Hop Washing Casual At…
5 Winter limited edition ITEM Super silky touch back brushed colo…

Class: **bath towel**



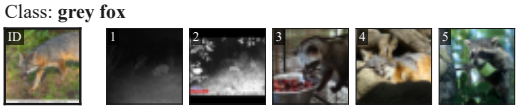
1 Cotton gauze towel wash towel adults and children daily quick-d…
2 Professional Management Of Housekeeping Operations
3 RTS Elemental Pentacle Embroidered Bath Towel Wicca witchcraft…
4 towels washcloth yellow
5 Monogram Turkish Cotton Towel, Alternate thumbnail 2, color, 454

Class: **mountain bike**



1 Battery Operated Bicycle LED Strip Lights Waterproof Wheel Fram…
2 Alberto convinces Calla to fill her water bottle from a town fo…
3 Cycling silhouettes, set of vector cyclist
4 Term 4- Expo Skills, Bridge Building, Bikes
5 Specialized S-Works SJ 2010 TEST

Class: **grey fox**



1 A lone Feral Hog photographed inside the boundaries of the VCDB.
2 Fox near Hawick - Night Vision Infrared HD Wildlife Camera clip…
3 The story behind the most expensive gourmet coffee in the world…
4 Channel Island Fox at the Los Angeles Zoo
5 Wildlife Photograph - Raccoon--up We Go by Larry Allan

Class: **snail**



1 Snail Engraving Illustration Royalty Free Stock Photos
2 texture background snail
3 Giant african land snail eating - photo#4
4 When the snail smiles have you seen before ? Snails Snail Colle…
5 Snail on a wet leaf wallpaper

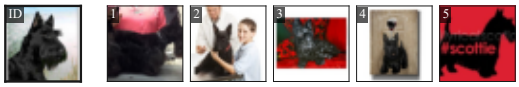
ImageNet-1K, general LAION pool

Class: **mailbox**



1 Vintage hunting with The Style Contingent founder in New Mexico…
2 Matheson Ancient Scotland Mailbox Clan A91
3 Posting A Letter To Red British Postbox On Street
4 Ireland biggest selection ROOF tiles reclaimed
5 Image for Victorian Wall Post Box - Handcross Road - Balcombe -…

Class: **scotch terrier**



1 Scottish Terrier Nachher 1
2 Veterinarian giving a shot to a little boy's Scotty dog. White…
3 Haggis the Scottish Terrier Card
4 'Dog Au Vin, Scottish Terrier' by Fab Funky, Giclee Canvas Wall…
5 #scottie postcard (HA6)

Class: **lynx**



1 Lynx in snow — Stock Photo
2 Bobcat Stalking North America Framed Print By Tim Fitzharris
3 Young Bobcats by Michael S. Quinton
4 Owls For Kids: A Children's Book About Owls, Owl Facts, Life, a…
5 Lynx ou chat sauvage rouge Photos stock

Class: **snow leopard**



1 Snow leopard Stock Images
2 Mp Photos - Snow Leopard Uncia Uncia Portrait by Gerry Ellis
3 Side Swipe Snow Leopard Cub Canvas Print #STG93} by David Strib…
4 Snow leopard with funny and angry expression
5 Help out endangered snow leopards by watching Born in China ope…

Class: **black-footed ferret**



1 I once had the best ferret in all the world. Her name was Kalei…
2 So Cute 3 I Have Mice But Rats And Mice Are Some Of My Favorite…
3 Ferret sitting, facing, isolated on white
4 Ferret Family Adopts A Kitten
5 Young & Heart Racoon Tote 13""x13

Figure 14: **Top-activating probes and captions for recovered classes: ImageNet-1K from the general LAION pool of Section 4.1, ten classes.** Layout as described in Appendix A.5; every class shown was recovered exactly by KEEP.