ConceptHash: Interpretable Fine-Grained Hashing via Concept Discovery

Kam Woh Ng[1,2] Xiatian Zhu[1,3] Yi-Zhe Song[1,2] Tao Xiang[1,2]

[1]CVSSP, University of Surrey
[2]iFlyTek-Surrey Joint Research Centre on Artificial Intelligence
[3]Surrey Institute for People-Centred Artificial Intelligence

{kamwoh.ng,xiatian.zhu,y.song,t.xiang}@surrey.ac.uk

Abstract

Existing fine-grained hashing methods typically lack code interpretability as they compute hash code bits holistically using both global and local features. To address this limitation, we propose ConceptHash, a novel method that achieves sub-code level interpretability. In ConceptHash, each sub-code corresponds to a human-understandable concept, such as an object part, and these concepts are automatically discovered without human annotations. Specifically, we leverage a Vision Transformer architecture and introduce concept tokens as visual prompts, along with image patch tokens as model inputs. Each concept is then mapped to a specific sub-code at the model output, providing natural sub-code interpretability. To capture subtle visual differences among highly similar sub-categories (e.g., bird species), we incorporate language guidance to ensure that the learned hash codes are distinguishable within fine-grained object classes while maintaining semantic alignment. This approach allows us to develop hash codes that exhibit similarity within families of species while remaining distinct from species in other families. Extensive experiments on four fine-grained image retrieval benchmarks demonstrate that ConceptHash outperforms previous methods by a significant margin, offering unique sub-code interpretability as an additional benefit. Code at: https://github.com/kamwoh/concepthash.

1. Introduction

Learning to hash is an effective approach for constructing large-scale image retrieval systems [49]. Previous methods primarily use pointwise learning algorithms with efficient hash center-based loss functions [19, 23, 67, 90]. However, these methods mainly focus on global image-level information and are best suited for distinguishing broad categories with distinct appearance differences, like apples and buildings. In many real-world applications, it's essential to distinguish highly similar sub-categories with subtle local differences, such as different bird species. In such scenarios, the computation of hash codes that capture these local, class-discriminative visual features, like bird beak color, becomes crucial.

Recent fine-grained hashing methods [14, 65, 80] extract local features and then combine them with global features to compute hash codes. However, this approach lacks interpretability because hash codes are derived from a mix of local and global features. As a result, it becomes challenging to establish the connection between human-understandable concepts (e.g., tail length and beak color of a bird) and individual or blocks of hash code bits (sub-codes). These concepts are typically local, as globally fine-grained classes often share similar overall characteristics (e.g., similar body shapes in all birds).

The importance of model interpretability is growing in practical applications. Interpretable AI models boost user confidence, assist in problem-solving, offer insights, and simplify model debugging [48, 52, 71]. In the context of learning-to-hash, interpretability pertains to the clear connection between semantic concepts and hash codes. For instance, a block of hash code bits or sub-code should convey a specific meaning that can be traced back to a local image region for visual inspection and human comprehension. While the methods introduced in previous works [65, 80] were originally conceived with interpretability in mind, they have made limited progress in this regard. This limitation stems from the fact that their hash codes are computed from aggregated local and global feature representations, making it challenging to establish a direct association between a sub-code and a local semantic concept.

To address the mentioned limitation, we present an innovative concept-based hashing approach named *ConceptHash*, designed for interpretability (see Fig. 1). Our architecture builds upon the Vision Transformer (ViT) [15]. To enable semantic concept learning, we introduce learnable concept

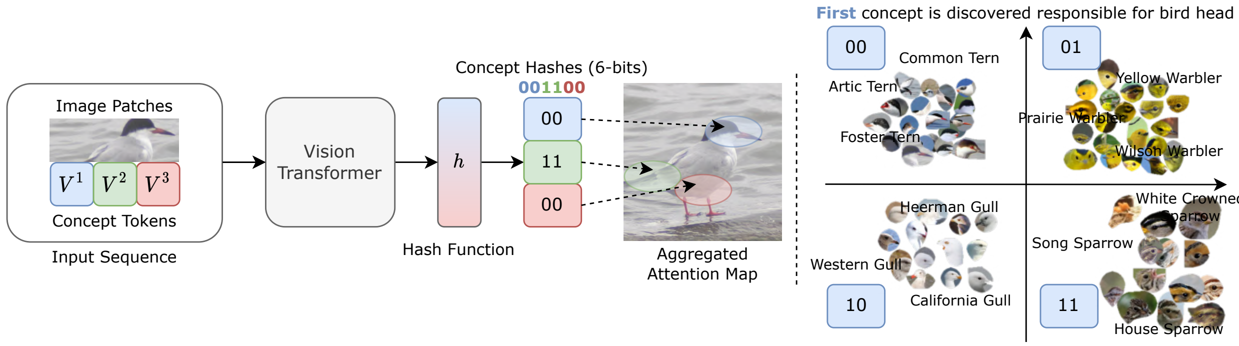


Figure 1. In the proposed ConceptHash, a set of concept tokens (3 tokens in this illustration) are introduced in a vision Transformer to discover automatically human understandable semantics (*e.g.*, bird head by the first concept token for generating the first two-bit sub-code **00**). Further, within the subcode, all similar concepts (*e.g.*, terns, warbler) are semantically grouped.

tokens as *visual prompts*, which are combined with image patch tokens as input to ViT. At the ViT's output, each query token corresponds to a sub-code. Concatenating these sub-codes yields the final hash code. Notably, the visual meaning of each concept token is evident upon inspection. This intrinsic feature makes our model interpretable at the sub-code level since each sub-code directly corresponds to a concept token. Additionally, we harness the rich textual information from a pretrained vision-language model (CLIP [58]) to offer language-based guidance. This ensures that our learned hash codes are not only discriminative within fine-grained object classes but also semantically coherent. By incorporating language guidance, our model learns hash codes that exhibit similarity within species' families while maintaining distinctiveness from species in other families. This approach enhances the expressiveness of the hash codes, capturing nuanced visual details and meaningful semantic distinctions, thereby boosting performance in fine-grained retrieval tasks.

Our **contributions** are as follows. **(1)** We introduce a novel ConceptHash approach for interpretable fine-grained hashing, where each sub-code is associated with a specific visual concept automatically. **(2)** We enhance the semantics of our approach by incorporating a pretrained vision-language model, ensuring that our hash codes semantically distinguish fine-grained classes. **(3)** Extensive experiments across four fine-grained image retrieval benchmarks showcase the superiority of ConceptHash over state-of-the-art methods, achieving significant improvements of 6.82%, 6.85%, 9.67%, and 3.72% on CUB-200-2011, NABirds, FGVC-Aircraft, and Stanford Cars, respectively.

2. Related Work

Learning to hash. Deep learning-based hashing [6, 7, 38, 77, 85] has dominated over conventional counterparts [20, 21, 27, 34, 36, 37, 54, 55, 83]. Recent works focus on a variety of aspects [49], *e.g.*, solving vanishing gradient problems caused by the sign function sign [45, 67], reducing the training complexity from $O(N^2)$ to $O(N)$ with pointwise loss [19, 23, 67, 90] and absorbing the quantization error objective [19, 23] into a single objective. These works usually consider the applications for differentiating coarse classes with clear pattern differences (*e.g.*, houses vs. cars), without taking into account hash code interpretability.

Fine-grained recognition. In many real-world applications, however, fine-grained recognition for similar sub-categories is needed, such as separating different bird species [81]. As the class discriminative parts are typically localized, finding such local regions becomes necessary. Typical approaches include attention mechanisms [9, 28, 57, 78, 87, 94–96], specialized architectures/modules [5, 22, 26, 46, 68, 75, 82, 92, 97], regularization losses [8, 12, 16, 17, 69], and finer-grained data augmentation [16, 40]. They have been recently extended to *fine-grained hashing*, such as attention learning in feature extraction [13, 28, 40, 47, 79, 86] and feature fusion [14, 65, 80]. However, in this study we reveal that these specialized methods are even less performing than recent coarse-grained hashing methods, in addition to lacking of code interpretability. Both limitations can be addressed with the proposed ConceptHash method in a simpler architecture design.

Model interpretability. Seeking model interpretability has been an increasingly important research topic. For interpretable classification, an intuitive approach is to find out the weighted combinations of concepts [30, 33, 64, 66, 84, 89, 91, 98] (a.k.a. prototypes [2, 53, 62]). This is inspired by human's way of learning new concepts via subconsciously discovering more detailed concepts and using them in varying ways for world understanding [39]. The concepts can be learned either through fine-grained supervision (*e.g.*, defining and labeling a handcrafted set of concepts) [33, 60, 89, 93], or weak supervision (*e.g.*, using weak labels such as image-level annotations) [56, 74], or self-supervision (*e.g.*, no any manual labels) [1, 74].

In this study, we delve into the realm of semantic concept learning within the context of learning-to-hash, with a dis-

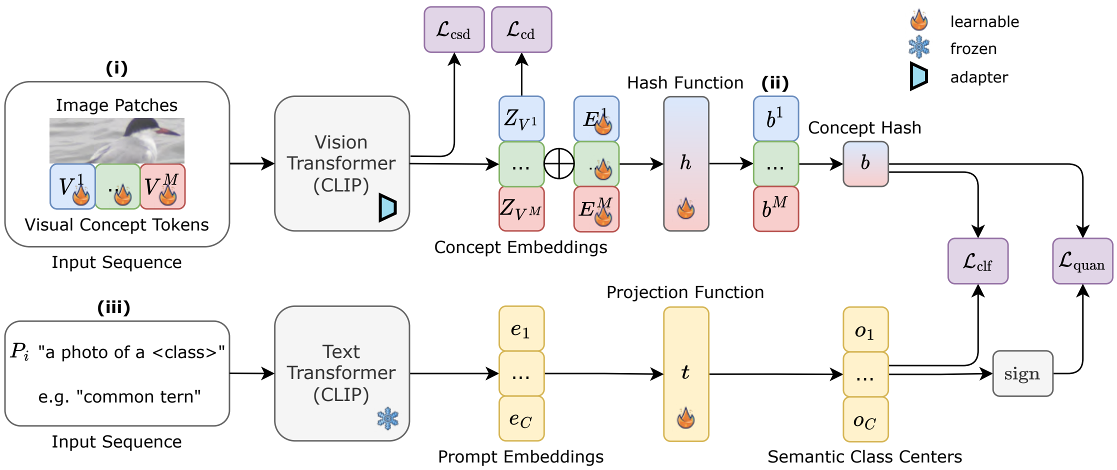


Figure 2. Overview of our ConceptHash model in a Vision Transformer (ViT) framework. To enable sub-code level interpretability, (i) we introduce a set of M concept tokens along with the image patch tokens as the input. After self-attention based representation learning, (ii) each of these concept tokens is then used to compute a sub-code, all of which are then concatenated to form the entire hash code. (iii) To compensate for limited information of visual observation, textual information of class names is further leveraged by learning more semantically meaningful hash class centers. For model training, a combination of classification loss $\mathcal{L}_{\text{clf}}$, quantization error $\mathcal{L}_{\text{quan}}$, concept spatial diversity constraint $\mathcal{L}_{\text{csd}}$, and concept discrimination constraint $\mathcal{L}_{\text{cd}}$ is applied concurrently. To increase training efficiency, Adapter [24] is added to the ViT instead of fine-tuning all parameters.

tinct emphasis on achieving sub-code level interpretability. While A^2-Net [80] has asserted that each bit encodes certain data-derived attributes, the actual computation of each bit involves a projection of both local and global features, making it challenging to comprehend the specific basis for the resulting bit values. In contrast, our approach, ConceptHash, takes a different approach. It begins by identifying common concepts (e.g., head, body) and subsequently learns the corresponding sub-codes within each concept space. Besides, our empirical findings demonstrate that ConceptHash outperforms previous methods in terms of performance.

Vision-language models. Vision-language pretraining at scale [58] has led to a surge of exploiting semantic language information in various problems [11, 31, 42–44, 76, 88]. *Prompting* has emerged as a promising technique for adapting large vision models to perform a variety of downstream tasks, such as classification [18, 29, 99, 100], dense prediction [59], interpretability [51, 89], metric learning [32, 61], self-supervised learning [4], and visual representations [25, 63]. For the first time, we explore the potential of prompting for fine-grained hashing. We introduce visual prompts to capture common concepts such as the head and body of a bird. On the language side, language guidance could complement the subtle visual differences of sub-categories while simultaneously preserving similarity within species belonging to the same family.

3. Methodology

We denote a training dataset with N samples as $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, where x_i is the n-th image with the label $y_i \in \{1, ..., C\}$. Our objective is to learn a hash function $\mathcal{H}(x) = h(f(x))$ that can convert an image x_i into a K-bits interpretable hash code $b \in \{-1, 1\}^K$ in a discriminative manner, where f is an image encoder (*e.g.*, a vision transformer) and h is a hashing function with a linear projection. To that end, we introduce a novel interpretable hashing approach, termed ***ConceptHash***, as illustrated in Fig. 2.

3.1. Concept-based Hashing

Given an image, our ConceptHash aims to generate an interpretable hash code composed by concatenating M sub-codes $\{b^1, ..., b^M\}$. Each sub-code $b^m \in \{-1, 1\}^{K/M}$ expresses a particular visual concept discovered automatically, with K the desired hash code length. To achieve this, we employ a Vision transformer (ViT) architecture denoted as f. At the input, apart from image patch tokens, we introduce a set of M learnable `concept tokens` as visual prompts:

$$Z^{(0)} = \text{concat}(x^1, ..., x^{\text{HW}}, [V^1], ..., [V^M]), \quad (1)$$

where concat denotes the concatenation operation, $[V^m]$ is the m-th concept token, x^i is the i-th image patch token with HW the number of patches per image (commonly, $HW = 7 * 7 = 49$). With this augmented token sequence $Z^{(0)}$, we subsequently leave the ViT model to extract the underlying concepts via the standard self-attention-based representation learning:

$$Z^{(L)} = f(Z^{(0)}) \in \mathbb{R}^{(HW+M)\times D},$$
$$\text{where } Z^{(l)} = \text{MSA}^{(l)}(Z^{(l-1)}), \quad (2)$$

in which $Z^{(l)}$ is the output of the l-th layer in a ViT and $\text{MSA}^{(l)}$ is the self-attention of l-th layer in f (the MLP, Layer Normalization [3], and the residual adding were omitted for simplicity). The last M feature vectors of $Z^{(L)}$ (denoted as Z for simplicity), Z[HW+1:HW+M], is the representation of the concepts discovered in a data-driven fashion, denoted as $Z_{[V^1]}, ..., Z_{[V^M]}$.

Interpretable hashing. Given each concept representation $Z_{[V^m]}$, we compute a specific sub-code b^m. Formally, we design a concept-generic hashing function h (a linear projection that maps D dimensional vector into K/M bits) as:

$$b^m = \text{h}(Z_{[V^m]} + E_m), \quad b = \text{concat}(b^1, ..., b^M), \tag{3}$$

where $E_m \in \mathbb{R}^{1\times D}$ is the m-th concept specificity embedding that enables a single hashing function to be shared across different concepts. In other words, the concept specificity embedding serves the purpose of shifting the embedding space of each specific concept to a common space, allowing a single hashing function to be applied to all concepts and convert them into hash codes. Note that b (the concatenation of all sub-codes) is a continuous code. To obtain the final hash code, we apply a sign function $\hat{b} = \text{sign}(b)$.

3.2. Language Guidance

Most existing fine-grained hashing methods rely on the information of visual features alone [14, 65, 80]. Due to the subtle visual difference between sub-categories, learning discriminative hashing codes becomes extremely challenging. We thus propose using the readily available semantic information represented as an embedding of the class names as an auxiliary knowledge source (*e.g.*, the semantic relation between different classes).

More specifically, in contrast to using random hash class centers as in previous methods [19, 23, 90], we learn to make them semantically meaningful under language guidance. To that end, we utilize the text embedding function $g(\cdot)$ of a pre-trained CLIP [58] to map a class-specific text prompt ($P \in \{P_c\}_{c=1}^C$ where P_c = `"a photo of a [CLASS]"`) to a pre-trained embedding space, followed by a learnable projection function $t(\cdot)$ to generate the semantic class centers:

$$e_c = g(P_c), \quad o_c = t(e_c). \tag{4}$$

The class centers $o = \{o_c\}_{c=1}^C \in \mathbb{R}^{C\times K}$ then serve as the hash targets for the classification loss in Eq. 6 and 7. This ensures that the learned hash codes are not only discriminative within fine-grained object classes but also semantically aligned. More specifically, the integration of language guidance guides the model to output hash codes that exhibit similarity within families of species while preserving discriminativeness from species belonging to other families (see Sec. 4.3 and Fig. 5).

3.3. Learning Objective

The objective loss function to train our ConceptHash model is formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{clf}} + \mathcal{L}_{\text{quan}} + \mathcal{L}_{\text{csd}} + \mathcal{L}_{\text{cd}}. \tag{5}$$

with each loss term as discussed below.

The first term $\mathcal{L}_{\text{clf}}$ is the classification loss for discriminative learning:

$$\mathcal{L}_{\text{clf}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\cos(o_{y_i}, b_i)/\tau)}{\sum_{c=1}^{C}\exp(\cos(o_c, b_i)/\tau)}, \tag{6}$$

where τ is the temperature ($\tau = 0.125$ by default), C is the number of classes, and cos computes the cosine similarity between two vectors. This is to ensure the hash codes are discriminative.

The second term $\mathcal{L}_{\text{quan}}$ is the quantization error:

$$\mathcal{L}_{\text{quan}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\cos(\hat{o_{y_i}}, b_i)/\tau)}{\sum_{c=1}^{C}\exp(\cos(\hat{o_c}, b_i)/\tau)},$$
$$\text{where} \quad \{\hat{o_c}\}_{c=1}^C = \{\text{sign}(o_c)\}_{c=1}^C. \tag{7}$$

Instead of directly minimizing the quantization error, we use the set of binarized class centers $\hat{o}$ as the classification proxy, which is shown to make optimization more stable [23].

The third term $\mathcal{L}_{\text{csd}}$ is a concept spatial diversity constraint:

$$\mathcal{L}_{\text{csd}} = \frac{1}{NM(M-1)}\sum_{i\neq j}\cos(A_i, A_j), \tag{8}$$

where $A_i \in \mathbb{R}^{N\times HW}$ is the attention map of the i-th concept token in the last layer of the self-attention $\text{MSA}^{(L)}$ of f, obtained by averaging over the multi-head axis, The idea is to enhance attention map diversity [10, 41, 82], thereby discouraging concepts from focusing on the same image region.

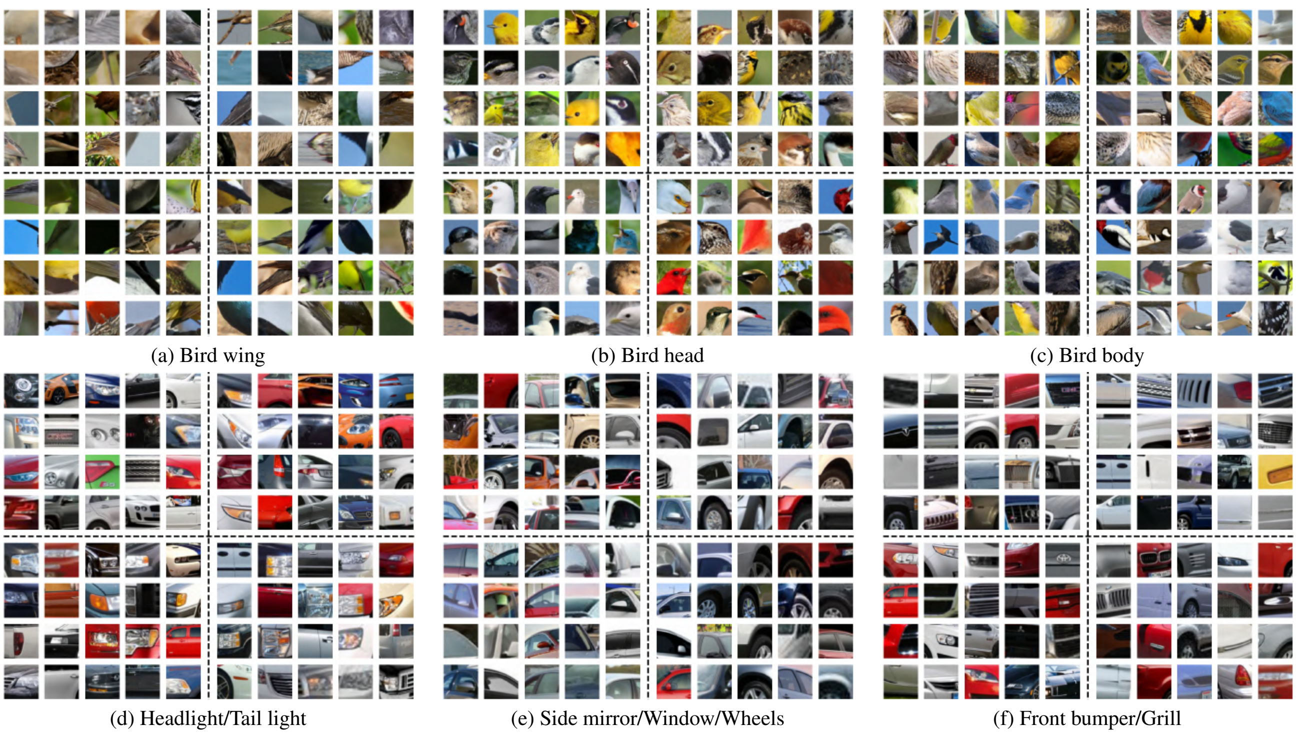
The forth term $\mathcal{L}_{\text{cd}}$ is the concept discrimination constraint:

$$p_{\text{cd}} = \frac{\exp(\cos(\hat{W}_{y_i}, \hat{Z}_{[V^m]_i})/\tau)}{\sum_{c=1}^{C}\exp(\cos(\hat{W}_c, \hat{Z}_{[V^m]_i})/\tau)}$$
$$\mathcal{L}_{\text{cd}} = -\frac{1}{NM}\sum_{i=1}^{N}\sum_{m=1}^{M}\log p_{\text{cd}},$$
$$\text{where} \quad \hat{Z}_{[V^m]_i} = Z_{[V^m]_i} + E^m, \tag{9}$$

where $\{\hat{W}_c\}_{c=1}^C \in \mathbb{R}^{C\times D}$ are learnable weights and $E \in \mathbb{R}^{M\times D}$ is the concept specificity embedding (same as E in Eq. 3). The feature-to-code process incurs substantial information loss (i.e., the projection from $Z_{[V]}$ to b), complicating the optimization. This loss serves a dual purpose: promoting discriminative concept extraction and supplying additional optimization gradients.

Dataset		CUB-200-2011			NABirds			FGVC-Aircraft			Stanford Cars		
Method		16	32	64	16	32	64	16	32	64	16	32	64
ITQ	[21]	7.82	11.53	15.42	3.40	5.50	7.60	8.12	9.78	10.87	7.80	11.41	15.16
HashNet	[7]	14.45	23.64	32.76	6.35	8.93	10.21	20.36	27.13	32.68	18.23	25.54	32.43
DTSH	[77]	25.16	27.18	27.89	3.35	6.00	7.87	21.32	25.65	36.05	20.48	27.40	28.34
GreedyHash	[67]	73.87	81.37	84.43	54.63	74.63	79.61	49.43	75.21	80.81	75.85	90.10	91.98
CSQ	[90]	69.61	75.98	78.19	62.33	71.24	73.61	65.94	72.81	74.05	82.16	87.89	87.71
DPN	[19]	76.63	80.98	81.96	68.82	74.52	76.75	70.86	74.04	74.31	87.67	89.46	89.56
OrthoHash	[23]	75.40	80.23	82.33	69.56	75.32	77.41	73.09	75.95	76.08	87.98	90.42	90.68
ExchNet†	[14]	51.04	65.02	70.03	-	-	-	63.83	76.13	78.69	40.28	69.41	78.69
A^2-Net	[80]	69.03	79.15	80.29	59.60	73.59	77.69	71.48	79.11	80.06	81.04	89.34	90.75
SEMICON	[65]	73.61	81.85	81.84	57.68	71.75	76.07	60.38	73.22	76.56	73.94	85.63	89.08
ConceptHash	(Ours)	**83.45**	**85.27**	**85.50**	**76.41**	**81.28**	**82.16**	**82.76**	**83.54**	**84.05**	**91.70**	**92.60**	**93.01**

Table 1. Comparing with prior art hashing methods. Note, ITQ is an unsupervised hashing method considered as the baseline performance. †: Originally reported results. **Bold**: The best performance.



(a) Bird wing (b) Bird head (c) Bird body

(d) Headlight/Tail light (e) Side mirror/Window/Wheels (f) Front bumper/Grill

Figure 3. We visualize the discovered concepts by our ConceptHash: (a, b, c) The bird body parts discovered on CUB-200-2011. (d, e, f) The car parts discovered on Stanford Cars. Setting: 6-bit hash codes where $M = 3$ concepts are used each for 2-bit sub-code. Bottom-left, top-left, top-right, and bottom-right regions represent the sub-codes 00, 01, 11, and 10 respectively.

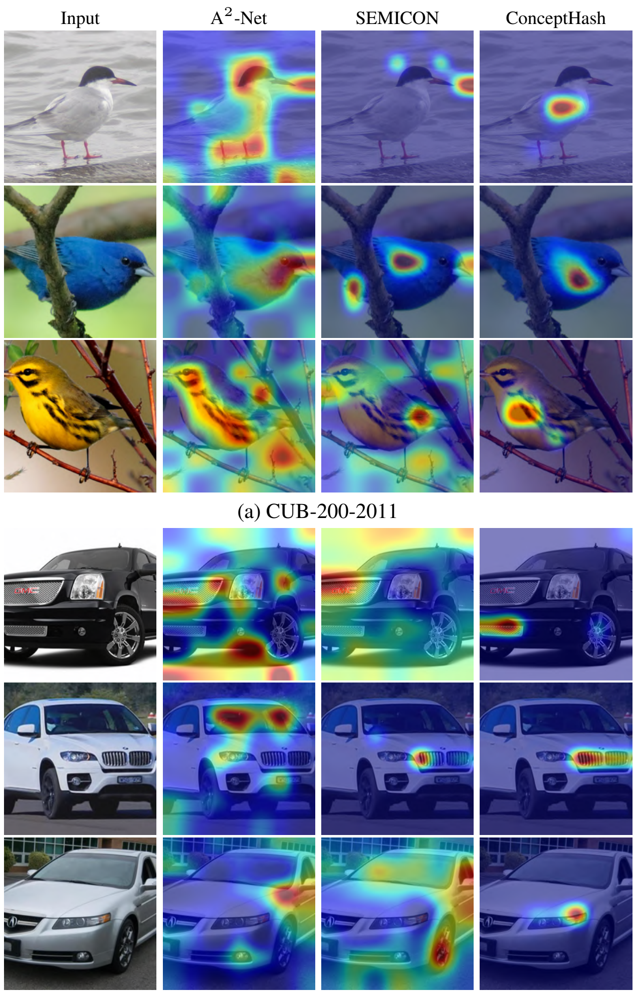
4. Experiments

Datasets We evaluate our method on four fine-grained image retrieval datasets: CUB-200-2011, NABirds, Aircraft, and CARS196. ***CUB-200-2011*** [73] has 200 bird species and 5.9K training/5.7K testing images. ***NABirds*** [72] has 555 bird species and 23K training/24K testing images. ***FGVC-Aircraft*** [50] has 100 aircraft variants and 6.6K training/3.3K testing images. ***Stanford Cars*** [35] has 196 car variants and 8.1K training/8.0K testing images. The experiment setting is exactly the same as those in previous works [14, 65, 80].

Implementation details We use the pre-trained CLIP [58] for our image encoder (a ViT/B-32 [15]) and text encoder (a 12-stacked layers Transformer). The SGD optimizer is

Dataset	CUB-200-2011			NABirds			FGVC-Aircraft			Stanford Cars		
Hash centers	16	32	64	16	32	64	16	32	64	16	32	64
Random vectors	77.00	79.61	82.12	71.93	73.66	76.20	71.22	77.28	79.19	88.57	87.17	88.46
Learnable vectors	81.55	82.39	83.86	75.80	79.76	81.66	82.08	83.43	83.62	91.03	91.92	92.84
Language (Ours)	**83.45**	**85.27**	**85.50**	**76.41**	**81.28**	**82.16**	**82.76**	**83.54**	**84.05**	**91.70**	**92.60**	**93.01**

Table 2. Effect of language guidance in forming the hash class centers.



(a) CUB-200-2011

(b) Stanford Cars

Figure 4. The regions where the hash function will focus on while computing a hash code.

Method	CUB-001	CUB-002	CUB-003
A^2-Net	34.0	37.7	34.4
SEMICON	43.7	60.5	34.4
Ours	**25.2**	**27.3**	**19.3**

Table 3. Results of landmark localization errors on CUB-200-2011. Normalized L2 distance (%) is reported.

$\mathcal{L}_{\text{quan}}$	$\mathcal{L}_{\text{csd}}$	$\mathcal{L}_{\text{cs}}$	CUB-200-2011		Stanford Cars	
			16	64	16	64
✗	✗	✗	68.65	82.00	81.85	91.20
✓	✗	✗	81.12	84.14	90.03	92.72
✓	✓	✗	81.63	84.79	90.63	92.82
✓	✗	✓	83.02	85.10	91.57	92.75
✓	✓	✓	**83.45**	**85.50**	**91.70**	**93.01**

Table 4. Loss ablation: The effect of adding gradually different loss terms of ConceptHash.

adopted with a momentum of 0.9 and a weight decay of 0.0001. The training epoch is 100 and the batch size is 32. We use a cosine decay learning rate scheduler with an initial learning rate of 0.001 and 10 epochs of linear warm-up. We adopt standard data augmentation strategies in training (*i.e.*, random resized crop and random horizontal flip only). For training efficiency, we insert learnable Adapters [24] to the frozen image encoder (see supplementary material for details). We use the same pretrained backbone, adapters, and training setting to fairly compare all the methods.

Performance metrics We adopt *mean average precision* which is the mean of average precision scores of the top R retrieved items, denoted as mAP@R. We set R = *full retrieval size* following previous works [14, 65, 80].

4.1. Comparison with the State-of-the-Art Methods

For comparative evaluation of ConceptHash, we consider both state-of-the-art coarse-grained (HashNet [7], DTSH [77], GreedyHash [67], CSQ [90], DPN [19], and OrthoHash [23]) and fine-grained (A^2-Net [80] and SEMICON [65]) methods. For fair comparisons, the same CLIP pre-trained image encoder (ViT/B-32) is used in all methods for feature extraction. Our implementation is based on the original source code.

Results. The fine-grained image retrieval results are reported in Table 1. We make several observations. **(1)** Among all the competitors, our ConceptHash achieves the best accuracy consistently across all the hash code lengths and datasets, particularly in the case of low bits (*e.g.*, 16 bits).

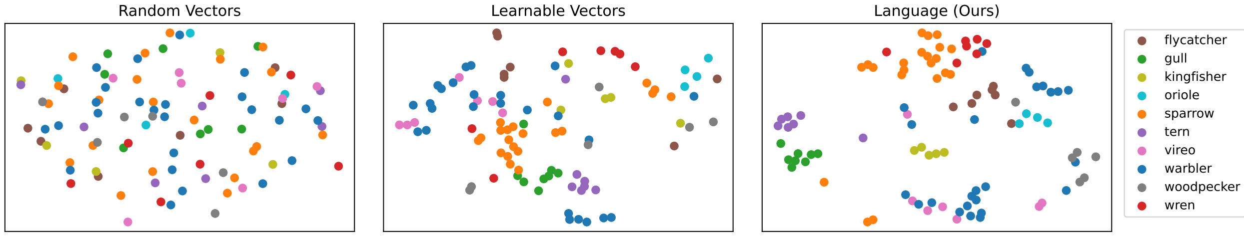
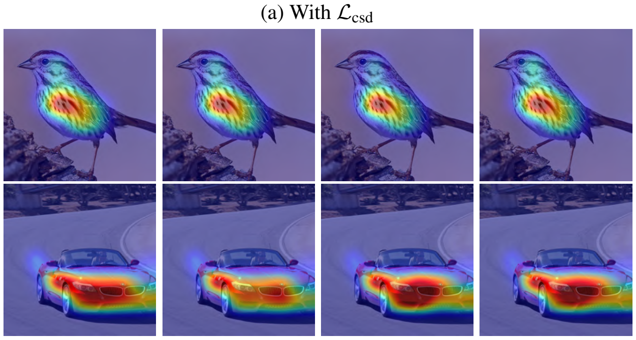


Figure 5. tSNE of the hash centers. The top 10 families of fine-grained classes in CUB-200-2011 are plotted for clarity.

(a) With $\mathcal{L}_{\mathrm{csd}}$



(b) Without $\mathcal{L}_{\mathrm{csd}}$

Figure 6. Effect of concept spatial diversity $\mathcal{L}_{\mathrm{csd}}$: The attention maps at the last layer of the Vision Transformer. Setting: $M = 4$.

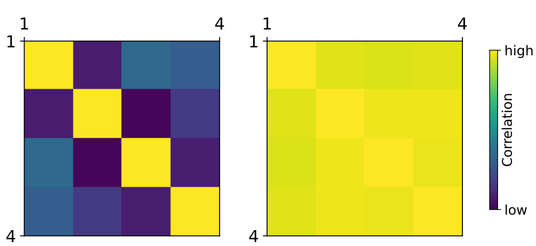


Figure 7. The correlation matrix between attention maps at the last layer of the vision transformer when training with (left) and without (right) the proposed concept spatial diversity constraint $\mathcal{L}_{\mathrm{csd}}$. This is averaged over all training images of CUB-200-2011 with $M = 4$.

This suggests the performance advantage of our method in addition to the code interpretability merit. In particular, it exceeds over the best alternative by a margin of up to 6.82%, 6.85%, 9.67%, and 3.72% on CUB-200-2011, NABirds, Aircraft, and CARS196, respectively. **(2)** Interestingly, previous coarse-grained hashing methods (*e.g.*, OrthoHash) even outperform the latest fine-grained hashing counterparts (*e.g.*, SEMICON). This suggests that their extracted local features are either not discriminative or uncomplimentary to the global features.

4.2. Interpretability Analysis

To examine *what concepts our ConceptHash can discover*, we start with a simple setting with 3 concepts each for a 2-bit sub-code (*i.e.*, a 6-bit hash code). We train the model on CUB-200-2011 and Standford Cars, respectively. For each concept, we find its attention in the attention map and crop the corresponding heat regions for visual inspection. As shown in Fig. 3, our model can automatically discover the body parts of a bird (*e.g.*, head, body, and wings) and car parts (*e.g.*, headlight, window, wheels, grill) from the images without detailed part-level annotation. This validates the code interpretability of our method.

Attention quality. Although fine-grained hashing methods (*e.g.*, A^2-Net [80] and SEMICON [65]) lack the code interpretability, local attention has been also adopted. We further evaluate the quality of attention with our ConceptHash and these methods. As shown in Fig. 4, we observe that while A^2-Net and SEMICON both can identify some discriminative parts of the target object along with background clutters, our model tends to give more accurate and more clean focus. This is consistent with the numerical comparison in Table 1, qualitatively verifying our method in the ability to identify the class discriminative parts of visually similar object classes.

Localization ability. We follow [26] to evaluate the quality of alignment between attention and ground-truth human-understandable concepts (*i.e.*, bird parts for CUB-200-2011 dataset). We compare our method to A^2-Net [80] and SEMICON [80]. As shown in Table 3, our method is significantly superior in the alignment error. These results suggest our model achieves higher interpretability in localizing parts.

Figure 8. Example attention maps. Setting: $M = 8$.

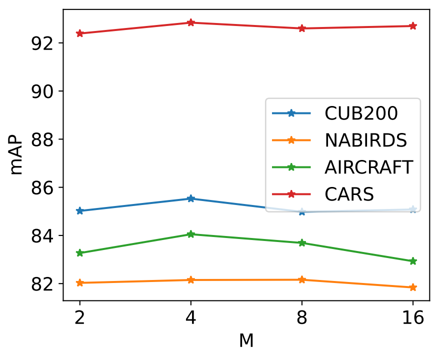


Figure 9. Impact of the concept number M. Setting: 64 bits.

4.3. Further Analysis

Impact of language guidance. We evaluate the effect of language guidance (Eq. 4) by comparing with two alternative designs without using the language information: (i) `Random vectors`: Using random orthogonal centers [23] without using visual information; (ii) `Learnable vectors`: Learning the class centers with discrete labels thus using the visual information. We observe from Table 2 that: **(1)** Our language-guided hash centers yield the best performance, validating our consideration that using extra textual information is useful for visually challenging fine-grained recognition. **(2)** Among the two compared designs, `Learnable vectors` is significantly superior, suggesting that hash class centers are a critical component in learning to hash and also that imposing the language information into class centers is a good design choice. **(3)** It is worth noting that, even without any language guidance (the `Learnable vectors` row of Table 2), our results are clearly superior to the compared alternatives (see Table 1).

For visual understanding, we plot the distribution of class centers on CUB-200-2011 using the tSNE [70]. For clarity, we select the top-10 families of bird species. As shown in Fig. 5, the class centers do present different structures. Using the visual information by `Learnable vectors`, it is seen that some classes under the same family are still farther apart (*e.g.*, the *kingfisher* family, gold colored). This limitation can be mitigated by our language guidance-based design. Furthermore, the class centers present more consistency with the family structures. For example, *tern* and *gull* are both seabirds, staying away from the other non-seabird families. This further validates that the semantic structural information captured by our ConceptHash could be beneficial for object recognition.

Loss design. We examine the effect of key loss terms. We begin with the baseline loss $\mathcal{L}_{\text{clf}}$ (Eq. 6). We observe in Table 4 that: (1) Without the quantization loss $\mathcal{L}_{\text{quan}}$ (Eq. 7), a significant performance drop occurs, consistent with the conventional findings of learning to hash. (2) Adding the concept spatial diversity constraint $\mathcal{L}_{\text{csd}}$ (Eq. 8) is helpful, confirming our consideration on the scattering property of underlying concepts. We find that this term helps to reduce the redundancy of attention (see Fig. 6 and Fig. 7). (3) Using the concept discrimination loss $\mathcal{L}_{\text{cd}}$ (Eq. 9) further improves the performance, as it can increase the discriminativeness of the extracted concepts.

5. Conclusion

In this work, we have introduced a novel concept-based fine-grained hashing method called ConceptHash. This method is characterized by learning to hash with sub-code level interpretability, along with leveraging language as extra knowledge source for compensating the limited visual information. Without manual part labels, it is shown that our method can identify meaningful object parts, such as head/body/wing for birds and headlight/wheel/bumper for cars. Extensive experiments show that our ConceptHash achieves superior retrieval performance compared to existing art methods, in addition to the unique code interpretability.

Limitations It is noted that increase in the number of concepts M can lead to overfitting and negatively impact interpretability, resulting in attention maps being scattered randomly around (see Fig. 9 and Fig. 8). The discovered concepts require manual inspection as the general clustering methods. Addressing these limitations will be the focus of our future work.

References

[1] David Alvarez Melis and Tommi Jaakkola. Towards robust interpretability with self-explaining neural networks. In *Advances in neural information processing systems*, 2018. 2

[2] Sercan Ö Arik and Tomas Pfister. Protoattend: Attention-based prototypical learning. *Journal of Machine Learning Research*, 2020. 2

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 4

[4] Mohamed El Banani, Karan Desai, and Justin Johnson. Learning visual representations via language-guided sampling. *arXiv preprint arXiv:2302.12248*, 2023. 3

[5] Ardhendu Behera, Zachary Wharton, Pradeep RPG Hewage, and Asish Bera. Context-aware attentional pooling (cap) for fine-grained visual classification. In *AAAI Conference on Artificial Intelligence*, 2021. 2

[6] Yue Cao, Mingsheng Long, Bin Liu, and Jianmin Wang. Deep cauchy hashing for hamming space retrieval. In *Computer Vision and Pattern Recognition*, 2018. doi: 10.1109/CVPR.2018.00134. 2

[7] Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Philip S. Yu. Hashnet: Deep learning to hash by continuation. In *International Conference on Computer Vision*, 2017. 2, 5, 6

[8] Dongliang Chang, Yifeng Ding, Jiyang Xie, Ayan Kumar Bhunia, Xiaoxu Li, Zhanyu Ma, Ming Wu, Jun Guo, and Yi-Zhe Song. The devil is in the channels: Mutual-channel loss for fine-grained image classification. *IEEE Transactions on Image Processing*, 2020. 2

[9] Dongliang Chang, Kaiyue Pang, Yixiao Zheng, Zhanyu Ma, Yi-Zhe Song, and Jun Guo. Your" flamingo" is my" bird": Fine-grained, or not. In *Computer Vision and Pattern Recognition*, 2021. 2

[10] Tianlong Chen, Zhenyu Zhang, Yu Cheng, Ahmed Awadallah, and Zhangyang Wang. The principle of diversity: Training stronger vision transformers calls for reducing all levels of redundancy. In *Computer Vision and Pattern Recognition*, 2022. 4

[11] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*, 2020. 3

[12] Yue Chen, Yalong Bai, Wei Zhang, and Tao Mei. Destruction and construction learning for fine-grained image recognition. In *Computer Vision and Pattern Recognition*, 2019. 2

[13] Zhen-Duo Chen, Xin Luo, Yongxin Wang, Shanqing Guo, and Xin-Shun Xu. Fine-grained hashing with double filtering. *IEEE Transactions on Image Processing*, 2022. 2

[14] Quan Cui, Qing-Yuan Jiang, Xiu-Shen Wei, Wu-Jun Li, and Osamu Yoshie. Exchnet: A unified hashing network for large-scale fine-grained image retrieval. In *European Conference on Computer Vision*. Springer, 2020. 1, 2, 4, 5, 6

[15] 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 *International Conference on Learning Representations*, 2021. 1, 5

[16] Ruoyi Du, Dongliang Chang, Ayan Kumar Bhunia, Jiyang Xie, Zhanyu Ma, Yi-Zhe Song, and Jun Guo. Fine-grained visual classification via progressive multi-granularity training of jigsaw patches. In *European Conference on Computer Vision*. Springer, 2020. 2

[17] Abhimanyu Dubey, Otkrist Gupta, Ramesh Raskar, and Nikhil Naik. Maximum-entropy fine grained classification. In *Advances in Neural Information Processing Systems*, 2018. 2

[18] Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. *arXiv preprint arXiv:1909.00512*, 2019. 3

[19] Lixin Fan, Kam Woh Ng, Ce Ju, Tianyu Zhang, and Chee Seng Chan. Deep polarized network for supervised learning of accurate binary hashing codes. In *International Joint Conference on Artificial Intelligence*, 2020. 1, 2, 4, 5, 6

[20] Aristides Gionis, Piotr Indyk, Rajeev Motwani, et al. Similarity search in high dimensions via hashing. In *International Conference on Very Large Data Bases*, 1999. 2

[21] Yunchao Gong, Svetlana Lazebnik, Albert Gordo, and Florent Perronnin. Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012. 2, 5

[22] Ju He, Jie-Neng Chen, Shuai Liu, Adam Kortylewski, Cheng Yang, Yutong Bai, and Changhu Wang. Transfg: A transformer architecture for fine-grained recognition. In *AAAI Conference on Artificial Intelligence*, 2022. 2

[23] Jiun Tian Hoe, Kam Woh Ng, Tianyu Zhang, Chee Seng Chan, Yi-Zhe Song, and Tao Xiang. One loss for all: Deep hashing with a single cosine similarity based learning objective. In *Advances in Neural Information Processing Systems*, 2021. 1, 2, 4, 5, 6, 8

[24] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, 2019. 3, 6

[25] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *Computer Vision and Pattern Recognition*, 2021. 3

[26] Zixuan Huang and Yin Li. Interpretable and accurate fine-grained recognition via region grouping. In *Computer Vision and Pattern Recognition*, 2020. 2, 7

[27] Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Annual ACM Symposium on Theory of Computing*, 1998. 2

[28] Sheng Jin, Hongxun Yao, Xiaoshuai Sun, Shangchen Zhou, Lei Zhang, and Xiansheng Hua. Deep saliency hashing for fine-grained retrieval. *IEEE Transactions on Image Processing*, 2020. 2

[29] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Computer Vision and Pattern Recognition*, 2023. 3

[30] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, 2018. 2

[31] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, 2021. 3

[32] Konstantin Kobs, Michael Steininger, and Andreas Hotho. Indirect: Language-guided zero-shot deep metric learning for images. In *Winter Conference on Applications of Computer Vision*, 2023. 3

[33] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning*, 2020. 2

[34] Weihao Kong and Wu-jun Li. Isotropic hashing. In *Advances in Neural Information Processing Systems*, 2012. 2

[35] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *International Conference on Computer Vision Workshops*, 2013. 5

[36] Brian Kulis and Trevor Darrell. Learning to hash with binary reconstructive embeddings. In *Advances in Neural Information Processing Systems*, 2009. 2

[37] Brian Kulis and Kristen Grauman. Kernelized locality-sensitive hashing for scalable image search. In *International Conference on Computer Vision*, 2009. 2

[38] Hanjiang Lai, Yan Pan, Ye Liu, and Shuicheng Yan. Simultaneous feature learning and hash coding with deep neural networks. In *Computer Vision and Pattern Recognition*, 2015. 2

[39] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 2015. 2

[40] Wenxi Lang, Han Sun, Can Xu, Ningzhong Liu, and Huiyu Zhou. Discriminative feature mining hashing for fine-grained image retrieval. *Journal of Visual Communication and Image Representation*, 2022. 2

[41] Jian Li, Xing Wang, Zhaopeng Tu, and Michael R Lyu. On the diversity of multi-head attention. *Neurocomputing*, 2021. 4

[42] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in neural information processing systems*, 2021. 3

[43] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[44] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, 2020. 3

[45] Yunqiang Li and Jan van Gemert. Deep unsupervised image hashing by maximizing bit entropy. In *AAAI Conference on Artificial Intelligence*, 2021. 2

[46] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear cnn models for fine-grained visual recognition. In *International Conference on Computer Vision*, 2015. 2

[47] Di Lu, Jinpeng Wang, Ziyun Zeng, Bin Chen, Shudeng Wu, and Shu-Tao Xia. Swinfghash: Fine-grained image retrieval via transformer-based hashing network. In *British Machine Vision Conference*, 2021. 2

[48] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, 2017. 1

[49] Xiao Luo, Daqing Wu, Chong Chen, Minghua Deng, Jianqiang Huang, and Xian-Sheng Hua. A survey on deep hashing methods. *arXiv preprint arXiv:2003.03369*, 2020. 1, 2

[50] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 5

[51] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. In *International Conference on Learning Representations*, 2023. 3

[52] Christoph Molnar. *Interpretable machine learning*. Lulu.com, 2020. 1

[53] Meike Nauta, Ron van Bree, and Christin Seifert. Neural prototype trees for interpretable fine-grained image recognition. In *Computer Vision and Pattern Recognition*, 2021. 2

[54] Mohammad Norouzi and David J. Fleet. Minimal loss hashing for compact binary codes. In *International Conference on Machine Learning*, 2011. 2

[55] Mohammad Norouzi, David J Fleet, and Russ R Salakhutdinov. Hamming distance metric learning. In *Advances in Neural Information Processing Systems*, 2012. 2

[56] Tuomas Oikarinen, Subhro Das, Lam M. Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. In *International Conference on Learning Representations*, 2023. 2

[57] Yuxin Peng, Xiangteng He, and Junjie Zhao. Object-part attention model for fine-grained image classification. *IEEE Transactions on Image Processing*, 2017. 2

[58] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. 2, 3, 4, 5

[59] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Computer Vision and Pattern Recognition*, 2022. 3

[60] Mattia Rigotti, Christoph Miksovic, Ioana Giurgiu, Thomas Gschwind, and Paolo Scotton. Attention-based interpretability with concept transformers. In *International Conference on Learning Representations*, 2022. 2

[61] Karsten Roth, Oriol Vinyals, and Zeynep Akata. Integrating language guidance into vision-based deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16177–16189, 2022. 3

[62] Dawid Rymarczyk, Łukasz Struski, Jacek Tabor, and Bartosz Zieliński. Protopshare: Prototypical parts sharing for similarity discovery in interpretable image classification. In *ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021. 2

[63] Mert Bulent Sariyildiz, Julien Perez, and Diane Larlus. Learning visual representations with caption annotations. In *European Conference on Computer Vision*, 2020. 3

[64] Yoshihide Sawada and Keigo Nakamura. Concept bottleneck model with additional unsupervised concepts. *IEEE Access*, 2022. 2

[65] Yang Shen, Xuhao Sun, Xiu-Shen Wei, Qing-Yuan Jiang, and Jian Yang. Semicon: A learning-to-hash solution for large-scale fine-grained image retrieval. In *European Conference on Computer Vision*. Springer, 2022. 1, 2, 4, 5, 6, 7

[66] Wolfgang Stammer, Marius Memmel, Patrick Schramowski, and Kristian Kersting. Interactive disentanglement: Learning concepts by interacting with their prototype representations. In *Computer Vision and Pattern Recognition*, 2022. 2

[67] Shupeng Su, Chao Zhang, Kai Han, and Yonghong Tian. Greedy hash: Towards fast optimization for accurate hash coding in cnn. In *Advances in Neural Information Processing Systems*, 2018. 1, 2, 5, 6

[68] Guolei Sun, Hisham Cholakkal, Salman Khan, Fahad Khan, and Ling Shao. Fine-grained recognition: Accounting for subtle differences between similar classes. In *AAAI conference on Artificial Intelligence*, 2020. 2

[69] Ming Sun, Yuchen Yuan, Feng Zhou, and Errui Ding. Multi-attention multi-class constraint for fine-grained image recognition. In *European Conference on Computer Vision*, 2018. 2

[70] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 2008. 8

[71] Bas HM Van der Velden, Hugo J Kuijf, Kenneth GA Gilhuijs, and Max A Viergever. Explainable artificial intelligence (xai) in deep learning-based medical image analysis. *Medical Image Analysis*, 2022. 1

[72] Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Computer Vision and Pattern Recognition*, 2015. 5

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

[74] Bowen Wang, Liangzhi Li, Yuta Nakashima, and Hajime Nagahara. Learning bottleneck concepts in image classification. In *Computer Vision and Pattern Recognition*, 2023. 2

[75] Jun Wang, Xiaohan Yu, and Yongsheng Gao. Feature fusion vision transformer for fine-grained visual categorization. In *British Machine Vision Conference*, 2021. 2

[76] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *Computer Vision and Pattern Recognition*, 2016. 3

[77] Xiaofang Wang, Yi Shi, and Kris M Kitani. Deep supervised hashing with triplet labels. In *Asian Conference on Computer Vision*, 2016. 2, 5, 6

[78] Yaming Wang, Vlad I Morariu, and Larry S Davis. Learning a discriminative filter bank within a cnn for fine-grained recognition. In *Computer Vision and Pattern Recognition*, 2018. 2

[79] Yimu Wang, Xiu-Shen Wei, Bo Xue, and Lijun Zhang. Piecewise hashing: A deep hashing method for large-scale fine-grained search. In *Pattern Recognition and Computer Vision*. Springer-Verlag, 2020. ISBN 978-3-030-60638-1. doi: 10.1007/978-3-030-60639-8_36. URL https://doi.org/10.1007/978-3-030-60639-8_36. 2

[80] Xiu-Shen Wei, Yang Shen, Xuhao Sun, Han-Jia Ye, and Jian Yang. A^2-net: Learning attribute-aware hash codes for large-scale fine-grained image retrieval. In *Advances in Neural Information Processing Systems*, 2021. 1, 2, 3, 4, 5, 6, 7

[81] Xiu-Shen Wei, Yi-Zhe Song, Oisin Mac Aodha, Jianxin Wu, Yuxin Peng, Jinhui Tang, Jian Yang, and Serge Belongie. Fine-grained image analysis with deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2

[82] Philippe Weinzaepfel, Thomas Lucas, Diane Larlus, and Yannis Kalantidis. Learning super-features for image retrieval. In *International Conference on Learning Representations*, 2021. 2, 4

[83] Yair Weiss, Antonio Torralba, and Rob Fergus. Spectral hashing. In *Advances in Neural Information Processing Systems*, 2009. 2

[84] Shirley Wu, Mert Yuksekgonul, Linjun Zhang, and James Zou. Discover and cure: Concept-aware mitigation of spurious correlation. *arXiv preprint arXiv:2305.00650*, 2023. 2

[85] Rongkai Xia, Yan Pan, Hanjiang Lai, Cong Liu, and Shuicheng Yan. Supervised hashing for image retrieval via image representation learning. In *AAAI Conference on Artificial Intelligence*, 2014. 2

[86] Xinguang Xiang, Yajie Zhang, Lu Jin, Zechao Li, and Jinhui Tang. Sub-region localized hashing for fine-grained image retrieval. *IEEE Transactions on Image Processing*, 2021. 2

[87] Xuhui Yang, Yaowei Wang, Ke Chen, Yong Xu, and Yonghong Tian. Fine-grained object classification via self-supervised pose alignment. In *Computer Vision and Pattern Recognition*, 2022. 2

[88] Yang Yang, Yadan Luo, Weilun Chen, Fumin Shen, Jie Shao, and Heng Tao Shen. Zero-shot hashing via transferring supervised knowledge. In *ACM International Conference on Multimedia*, 2016. 3

[89] Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. *arXiv preprint arXiv:2211.11158*, 2022. 2, 3

[90] Li Yuan, Tao Wang, Xiaopeng Zhang, Francis EH Tay, Zequn Jie, Wei Liu, and Jiashi Feng. Central similarity quantization for efficient image and video retrieval. In *Computer Vision and Pattern Recognition*, 2020. 1, 2, 4, 5, 6

[91] Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. In *International Conference on Learning Representations Workshops*, 2022. 2

[92] Ziyun Zeng, Jinpeng Wang, Bin Chen, Tao Dai, and Shu-Tao Xia. Pyramid hybrid pooling quantization for efficient fine-grained image retrieval. *arXiv preprint arXiv:2109.05206*, 2021. 2

[93] Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu. Interpretable convolutional neural networks. In *Computer Vision and Pattern Recognition*, 2018. 2

[94] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *International Conference on Computer Vision*, 2017. 2

[95] Heliang Zheng, Jianlong Fu, Zheng-Jun Zha, and Jiebo Luo. Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. In *Computer Vision and Pattern Recognition*, 2019.

[96] Heliang Zheng, Jianlong Fu, Zheng-Jun Zha, Jiebo Luo, and Tao Mei. Learning rich part hierarchies with progressive attention networks for fine-grained image recognition. *IEEE Transactions on Image Processing*, 2019. 2

[97] Xiawu Zheng, Rongrong Ji, Xiaoshuai Sun, Yongjian Wu, Feiyue Huang, and Yanhua Yang. Centralized ranking loss with weakly supervised localization for fine-grained object retrieval. In *International Joint Conference on Artificial Intelligence*, 2018. 2

[98] Bolei Zhou, Yiyou Sun, David Bau, and Antonio Torralba. Interpretable basis decomposition for visual explanation. In *European Conference on Computer Vision*, 2018. 2

[99] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-

language models. In *Computer Vision and Pattern Recognition*, 2022. 3

[100] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 2022. 3