Prompt-Guided Attention Head Selection for Focus-Oriented Image Retrieval

Yuji Nozawa Yu-Chieh Lin Kazumoto Nakamura Youyang Ng
Kioxia Corporation
{yuji1.nozawa,yuchieh.lin,kazumoto1.nakamura,youyang.ng}@kioxia.com

Abstract

The goal of this paper is to enhance pretrained Vision Transformer (ViT) models for focus-oriented image retrieval with visual prompting. In real-world image retrieval scenarios, both query and database images often exhibit complexity, with multiple objects and intricate backgrounds. Users often want to retrieve images with specific object, which we define as the Focus-Oriented Image Retrieval (FOIR) task. While a standard image encoder can be employed to extract image features for similarity matching, it may not perform optimally in the multi-object-based FOIR task. This is because each image is represented by a single global feature vector. To overcome this, a prompt-based image retrieval solution is required. We propose an approach called Prompt-guided attention Head Selection (PHS) to leverage the head-wise potential of the multi-head attention mechanism in ViT in a promptable manner. PHS selects specific attention heads by matching their attention maps with user's visual prompts, such as a point, box, or segmentation. This empowers the model to focus on specific object of interest while preserving the surrounding visual context. Notably, PHS does not necessitate model retraining and avoids any image alteration. Experimental results show that PHS substantially improves performance on multiple datasets, offering a practical and training-free solution to enhance model performance in the FOIR task.

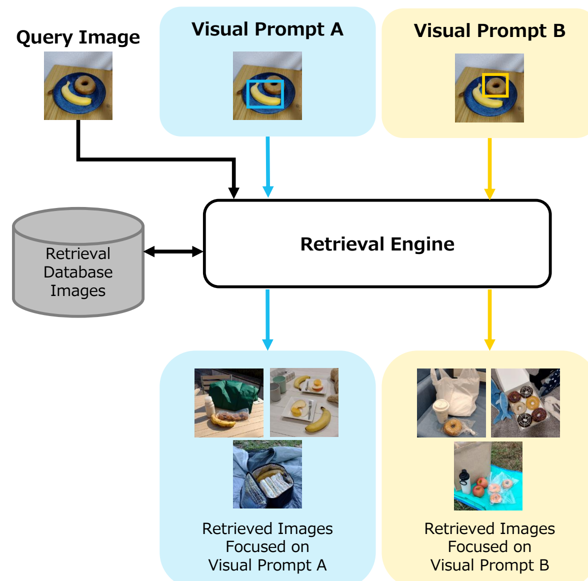

Figure 1. Overview of Focus-Oriented Image Retrieval (FOIR) task with illustration of visual prompting approach. FOIR task simulates real-world scenarios wherein (1) both query and retrieval database images often exhibit complexity, with multiple objects and intricate backgrounds; (2) users are visually interested in retrieving images containing specific object.

1. Introduction

Image retrieval (IR) encompasses a wide range of applications, including face recognition [48], landmark retrieval [43], and online shopping [27]. A common approach for retrieving images involves extracting image features and determining their similarity. This process is referred to as Content-Based Image Retrieval (CBIR) [22, 43, 51]. Standard CBIR approach demonstrates good performance when dealing with less complicated images, especially those that contain a single object. To improve the accuracy, researchers have employed deep neural network models such as Convolutional Neural Networks [30] and Vision Transformers (ViTs) [18] to extract image features, and have localized the retrieval problem to specific objects, a task setting known as Localized CBIR [47] or Object-Based Image Retrieval (OBIR) [25]. However, existing studies in Localized CBIR often employ image alteration preprocessing techniques such as masking and cropping, optimize the retrieval of images with less complexity, and tend to overlook the consideration of user intention [4, 25, 47, 65]. In reality, users may have the desire to retrieve a specific object while considering certain visual contextual constraints in complex images. The use of image alteration techniques can introduce errors during preprocessing or result in a loss of visual context [70], leading to retrieval failures. These inherent limitations potentially hinder its practical applicability.

In real-world image retrieval, both the query image and the images in the retrieval database often exhibit complexity, characterized by the presence of multiple objects and intricate backgrounds, posing a challenge for users aiming to retrieve images with specific objects. Despite its practicality, this topic has received limited research attention thus far. We argue that this specific scenario warrants dedicated attention in the implementation of image retrieval systems. In light of this, we set up Focus-Oriented Image Retrieval (FOIR) task, specifically for retrieving objects from complex images, as demonstrated in Fig. 1. FOIR can be considered a distinct category within the Localized CBIR task, wherein both the query and retrieval database consist of complex images, and users are interested in retrieving images with specific object. Query formats can be classified into two main categories: whole-image-as-query (WIQ) and image-region-as-query (IRQ) [16]. FOIR task specifically addresses WIQ, wherein the entire image is utilized to conduct the query. One solution is to employ a standard image encoder to extract image features for similarity matching. However, this may not work well for multi-object-based FOIR tasks as it represents each image with a single global feature vector. Image encoders tend to focus on the most salient region, potentially ignoring other objects or regions of interest. Prompt-based preprocessing techniques such as image cropping can complement the image encoder but may fail when wider visual context is required, as the perceptions of the user and the model may not align. Therefore, a solution that *considers user visual preference, preserves visual context, and aligns user-model perceptions* is essential.

To overcome this, we propose leveraging the head-wise potential of the Multi-Head Attention (MHA) mechanism of ViT in a user promptable manner. The attention heads in ViT, particularly in the last layers, contain valuable high-level salient and segmentation information [7] for images. This information can be effectively utilized to enable object focusing in FOIR task. Attention maps from each head have been observed to focus on different objects or parts in an input image [7]. Building on this observation, we introduce Prompt-guided attention Head Selection (PHS) technique. PHS selects specific attention heads in the last layer of a pretrained ViT image encoder model by matching their attention maps with user's visual prompts, which can take the form of a point, box, or segmentation. This empowers the model to focus on specific object of interest while preserving the surrounding visual context. Notably, PHS does not require model re-training, making it a plug-and-play enhancement for off-the-shelf pretrained ViT models. Moreover, our method avoids any modifications to the input images, thereby avoiding any undesirable consequences of image editing. Through extensive experimental evaluations on multiple datasets, we demonstrate that PHS substantially improves performance and robustness by selecting attention heads that are more focused on the desired objects.

From a conceptual standpoint, our proposed method can be regarded as a high-level perception matching mechanism between human users and vision models. A recent work of Foveal Attention (FA) [70], which draws inspiration from human visual perception [6, 59], has endeavored to manipulate attention through visual prompting by incorporating attention mask into region of attention via addictive or blending operation. However, this approach inevitably alters the perception of attention heads themselves, favoring the user's perception over that of the vision model. In contrast, our method selects attention heads to align and match the user's perceptions with specific attention heads, thereby bridging the gap between human and model visual understanding without explicitly modifying the attentions of individual heads. Our method can also be viewed as grounding high-level human perception, in the form of visual prompting, to high-level perception that can be comprehended by the model, in the form of attention map in each head. Both FA and our method differ conceptually and possess their own strengths. FA aims to emulate human attention characteristics, while PHS strives to match human-model attentions. We believe that they are effective in their own right and also have a complementary relationship in bridging human-model perceptions.

Our contributions are summarized as follows:

- We design Prompt-guided attention Head Selection (PHS), a method that leverages the head-wise potential of the multi-head attention mechanism in ViT in a promptable manner to accommodate user visual preference, preserve visual context and align user-model perceptions.
- We empirically show that our proposed method enhances performance and robustness compared to existing methods across multiple datasets through extensive experiments and ablation studies in the Focus-Oriented Image Retrieval (FOIR) task, where both query and retrieval database images often exhibit complexity, with multiple objects and intricate backgrounds, and users are visually interested in retrieving images with specific object.

2. Related Work

2.1. Object-Based IR with User Interest

OBIR [8, 25, 47] localizes objects in retrieval task. As a solution, Ref. [65] proposes image alterations to improve retrieval performance, while Ref. [33] applies aggregated features for image retrieval. Ref. [38] approaches multi-object IR from a CBIR perspective. MSIR [4] sets up multi-subject IR based on complex images, similar to our task. However, it does not take into account user interest. In comparison, FOIR task can be seen as a specific category of OBIR, emphasizing on complex images and considering user visual intention. Several works have incorporated user intention into image retrieval using techniques such as sketch [20, 64, 67], Composed IR with text prompts [1, 13, 14, 23, 31, 52, 54, 57], and relevancy [40], but these task setups and methods require additional feedback data. In contrast, our task setup considers visual intention in query image, and our method employs visual prompting, which is a less demanding approach.

2.2. Visual Region Awareness & Prompting in ViT

ViT [18] is a transformer-based [53] model for Computer Vision (CV) tasks. Transformer architectures use self-attention mechanism in MHA module to capture complex relationships between tokens [53]. DINO [7] is one of the self-supervised learning (SSL) methods [7, 11, 34, 61] applied to ViT, shown to have work well in various downstream tasks. DINOv2 [44] further improves its performance by pretraining on a large-scale visual dataset. DINO and DINOv2 have shown to exhibit beneficial region and object awareness in the MHA module of the last layer [50, 55, 56]. Based on this characteristic, we primarily employ DINOv2 in our study but note that our method can generalize to models with different pretraining approach.

Region awareness is crucial in CV tasks like object detection and segmentation. Studies have enhanced CLIP [62, 69] and ViT models [9, 36, 60] to incorporate region awareness mechanisms. However, these methods usually require additional training, while our method focuses on inference-time attention manipulation through user-defined visual prompting. Similar to prompt engineering [5] in Natural Language Processing (NLP), various visual prompts have been explored in CV, such as points, boxes & masks [28], blurred surroundings [63], red circles [49], foveal attention [70] and learnable prompts [3, 26, 45, 66]. However, unlike our method, most of them require input image alterations. FALIP [70]'s approach is most relevant to our study, but it directly modifies individual attention heads while our method prioritizes user-model perception matching. Our method also enables multiple prompts types, similar to SAM [28]. In addition, these existing works did not specifically examine the area of visual-prompt-guided image retrieval, which remains underexplored. Our work aims to contribute insights into the intersection of visual prompts, user-model perceptions, and image retrieval.

2.3. Attention Manipulation & Head Selection

Numerous studies have investigated changes to the attention mechanism in ViT. Some have modified attention during training [12, 32], while others have focused on manipulating attention during inference [39, 68, 70]. However, few studies have specifically examined attention manipulation techniques without altering individual attention values like head selection. Head selection is a technique used in the MHA module of Transformers in NLP and CV domains. It prioritizes relevant heads and replaces less relevant heads with fixed values. In NLP, it improves multilingual machine translation by selecting heads based on languages being translated [21]. In CV, it enhances inference efficiency [37, 41], knowledge distillation [58], and generalization [42]. Head selection in these previous works are performed and optimized for each task through training. Conversely, our work introduces inference-time head selection with a pretrained ViT model for image retrieval using visual prompting. The selection of heads is not fixed in advance through training. To the best of our knowledge, this is the first attempt to apply head selection to image retrieval.

3. Approach

In this section, we introduce our task setup (Fig. 1) and proposed method (Fig. 2).

3.1. Focus-Oriented Image Retrieval Task

We set up Focus-Oriented Image Retrieval task with the objective of simulating real-world image retrieval scenarios. FOIR can be considered a distinct category within the Localized CBIR tasks. The FOIR task encompasses the following criteria: (1) Image-to-image retrieval; (2) The query and retrieval database images exhibit complexity, characterized by the presence of multiple objects and intricate backgrounds; (3) Users are visually interested in retrieving images with specific object or pattern. For query format, FOIR task specifically addresses whole-image-as-query (WIQ), wherein the entire image is utilized to conduct the query. An exemplary demonstration of the FOIR task is the vision of a general-purpose humanoid robot. In this scenario, a human user may visually request the robot's assistance in searching for objects similar to what it is currently perceiving. The search can occur within the robot's past or future visual memory of its daily routine. Both the visual memory and the current vision of the real-world environment exhibit complexity with multiple objects and intricate backgrounds in each visual input.

In FOIR task, to accommodate user preference, a prompt-based image retrieval solution is essential. Fig. 1 illustrates an example of the FOIR task with a visual prompting approach, where the user creates visual prompts (A or B) to enhance the retrieval of images of interest from a complex database. In order to prevent retrieval failures caused by query preprocessing errors or a reduction in visual context, it is essential to use a non-image-alteration technique that preserves the available visual context. However, the presence of rich visual context can negatively impact retrieval results unless the user's and model's perceptions are aligned. Therefore, it is crucial to have a technique that aligns user and model perceptions. Overall, to provide a robust solution to the FOIR task, we have devised a method that (1) considers user's visual preference, (2) preserves visual context, and (3) aligns user and model perceptions.

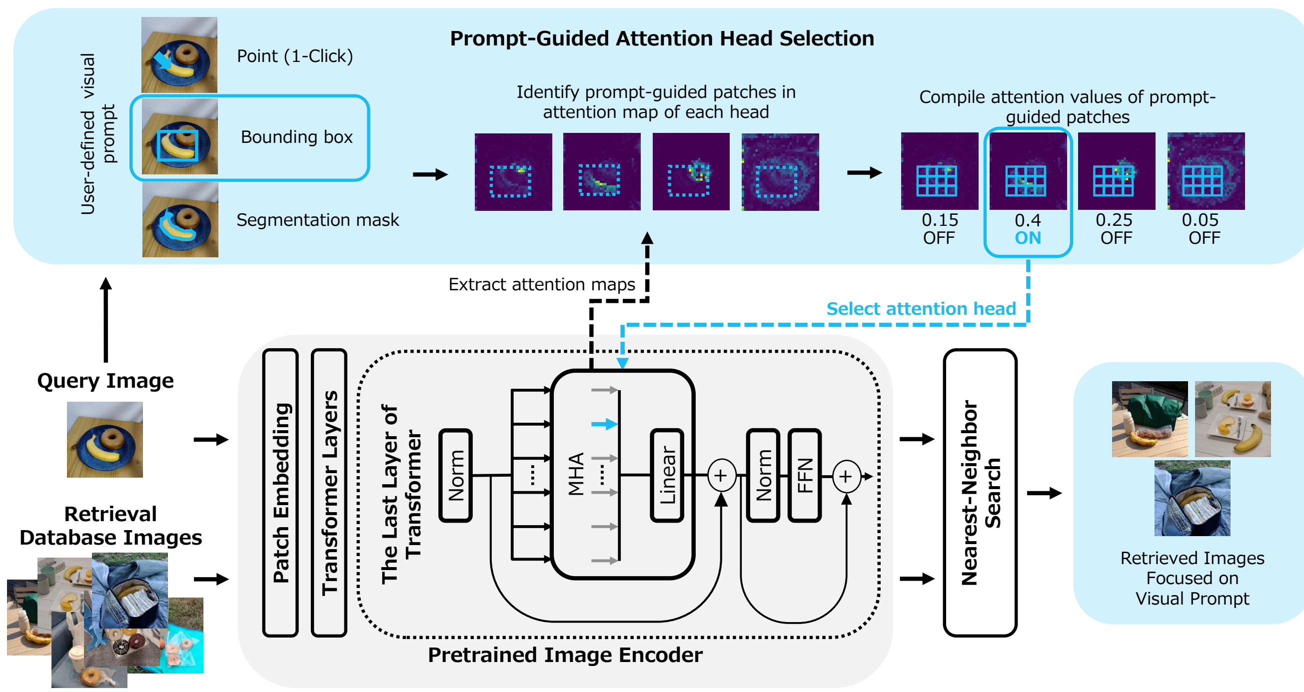

Figure 2. Overview of Prompt-guided attention Head Selection (PHS). PHS performs the selection of transformer attention heads in the pretrained ViT model by matching their attention maps with user's visual prompt, which can take the form of a point, box, or segmentation. This empowers the model to focus on specific object of interest while preserving the surrounding visual context. Best viewed in color.

3.2. Retrieval Process

We first introduce the overall image retrieval pipeline without head selection in this section. We employ pretrained ViT model to extract features of images. Let $\mathbf{x}_{\text{input}} \in \mathbb{R}^{H \times W \times C}$ be an input image, where H, W, and C represent the height, width, and channel of the image respectively. The image is divided into 2D patches $\mathbf{x}_{\text{patch}} \in \mathbb{R}^{N \times P \times P \times C}$, where the size of each patch is $P \times P \times C$, and $N = HW/P^2$ is the number of patches. After the patches are embedded by a trained linear projection layer, the `[CLS]` token is added to the patch tokens, and then the positional encoding is added to each embedding token. Consequently, $\mathbf{x}_{\text{patch}}$ is transformed into $\mathbf{x}_{\text{token}} \in \mathbb{R}^{(N+1) \times d}$, where d denotes the embedding dimension. The index of tokens runs from 0 to N, and 0 represents the `[CLS]` token. The tokens $\mathbf{x}_{\text{token}}$ are input into a sequence of attention layers.

In the MHA of each attention layer, three matrices are produced by transforming the input with three different trained linear projection layer: a query $\mathbf{Q}$, a key $\mathbf{K}$, and a value $\mathbf{V} \in \mathbb{R}^{(N+1) \times d}$. Then, the elements of $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are divided among multiple heads indexed by i, where $\mathbf{K}_i \in \mathbb{R}^{(N+1) \times d_k}, \quad i = 1, 2, \ldots, h$. h is the number of the heads, and d_k is the embedding dimension of $\mathbf{K}$. The same applies to $\mathbf{Q}$ and $\mathbf{V}$. The attention matrix of i head, denoted by $A_i \in \mathbb{R}^{(N+1) \times (N+1)}$ is defined as

$$A_i = \text{Softmax}\left(\frac{\mathbf{Q}_i \mathbf{K}_i^\top}{\sqrt{d_k}}\right). \tag{1}$$

Subsequently, MHA is calculated as

$$\text{MHA} = \text{Linear}\left([\text{head}_1, \text{head}_2, \ldots, \text{head}_h]\right), \tag{2}$$
$$\text{head}_i = A_i \mathbf{V}_i, \tag{3}$$

where Linear is a trained linear projection. To represent the feature of the image, we focus on the last attention layer. Let x_* and MHA_* be the input and MHA of the last attention layer respectively. Then, the output of the last attention layer z_* is written as

$$z_* = y_* + \text{FFN}\left(\text{LN}\left(y_*\right)\right), \text{ where } y_* = \mathrm{x}_* + \text{MHA}_*, \tag{4}$$

where LN is layer normalization [2] followed by a feed-forward network (FFN). The sums in Eq. (4) correspond to residual connections [24]. Finally, the feature of the image $\mathbf{f}(\mathbf{x}_{\text{input}})$ is obtained as the `[CLS]` part of the layer-normalized output:

$$\mathbf{f}\left(\mathbf{x}_{\text{input}}\right) = \left[\text{LN}\left(z_*\right)\right]_0. \tag{5}$$

To perform image retrieval, we use k-nearest neighbors (NN) method with cosine similarity. Let $\mathcal{I}_{\mathrm{Q}}$ be the set of query images and $\mathcal{I}_{\mathrm{DB}}$ be the set of database images. Then, for a query image $\mathbf{x}_{\mathrm{Q}} \in \mathcal{I}_{\mathrm{Q}}$, the k'th most similar image in the database $\mathbf{x}_{\mathrm{R}}(\mathbf{x}_{\mathrm{Q}}, \mathcal{I}_{\mathrm{DB}}, k')$ is retrieved as

$$\mathbf{x}_{\mathrm{R}}\left(\mathbf{x}_{\mathrm{Q}}, \mathcal{I}_{\mathrm{DB}}, k'\right) = \operatorname*{argmax}_{k', \, \mathbf{x}_{\mathrm{DB}} \in \mathcal{I}_{\mathrm{DB}}} \left(\frac{\mathbf{f}\left(\mathbf{x}_{\mathrm{DB}}\right)^{\top} \mathbf{f}\left(\mathbf{x}_{\mathrm{Q}}\right)}{\|\mathbf{f}\left(\mathbf{x}_{\mathrm{DB}}\right)\| \|\mathbf{f}\left(\mathbf{x}_{\mathrm{Q}}\right)\|} \right), \tag{6}$$

where $\mathrm{argmax}_{k'}(s)$ is a function that returns $\mathbf{x}_{\mathrm{DB}}$ with the k'th largest cosine similarity. Consequently, the top-k images can be retrieved by running Eq. (6) with $\forall k' \leq k$.

3.3. Prompt-Driven Attention Head Selection

Here, we present the algorithm for PHS. Our method involves the selection of h_{on} heads from a total of h heads in the last attention layer, based on the visual prompt mask $\mathbf{M}_{\mathrm{input}} \in \{0,1\}^{H \times W}$. The nonzero part of $\mathbf{M}_{\mathrm{input}}$ represents the region of interest (ROI) defined by the visual prompt. Initially, $\mathbf{M}_{\mathrm{input}}$ is converted to attention mask tokens denoted as $\mathbf{M}_{\mathrm{token}} \in \{0,1\}^{N}$. Within this set of tokens, we identify patch tokens in the ROI, marked them as $T^{\mathrm{HS}} = \{t \in \{1, 2, \ldots, N\} \mid \mathbf{M}_{\mathrm{token},t} = 1\}$. Subsequently, we calculate the ROI attention A_i^{HS} for each attention head by summing up the attention values of A_i in Eq. (1), but only for the tokens within T^{HS}, defined by

$$A_i^{\mathrm{HS}} = \sum_{t \in T^{\mathrm{HS}}} A_{i,0,t}. \tag{7}$$

The straightforwardness of Eq. (7) in extracting ROI attentions facilitates the integration of diverse visual prompts, including point, box, and segmentation. This, in turn, streamlines the system design process in real-world implementations. Once the ROI attention for each head is computed, we set h_{on} heads with the highest A_i^{HS} values as our selected heads $i_{\mathrm{on}} \subseteq \{1, 2, \ldots, h\}$, which satisfies $|i_{\mathrm{on}}| = h_{\mathrm{on}}$. In our method, h_{on} serves as the only parameter. Our experiments in Sec. 4 demonstrate that h_{on} is sufficiently robust and can be set as a fixed value.

Subsequently, we replace the MHA in the last attention layer MHA_* with our approach, $\mathrm{MHA}_*^{\mathrm{HS}}$, defined by

$$\mathrm{MHA}_*^{\mathrm{HS}} = \mathrm{Linear}\left[\mathrm{head}_{*\,1}^{\mathrm{HS}}, \mathrm{head}_{*\,2}^{\mathrm{HS}}, \ldots, \mathrm{head}_{*\,h}^{\mathrm{HS}}\right], \tag{8}$$

$$\mathrm{head}_{*\,i}^{\mathrm{HS}} = \begin{cases} \mathrm{head}_{*\,i} \times h/h_{\mathrm{on}} & \text{if } i \in i_{\mathrm{on}}, \\ 0 & \text{if } i \notin i_{\mathrm{on}}. \end{cases} \tag{9}$$

Our head selection strategy is inspired by Ref. [42], where head selection is performed prior to the output linear projection layer of MHA and $\mathrm{head}_{*\,i}$ is multiplied by a scaling factor of h/h_{on}. Scaling factor helps to preserve the overall attention intensity in MHA. Following the MHA, our subsequent process adheres to the standard retrieval pipeline by applying Eq. (8) to Eqs. (4) to (6). Importantly, our method does not necessitate any model fine-tuning. The head selection operation is performed to a trained ViT model during inference, dynamically matching the attention head to the user-defined visual prompt, as shown in Fig. 2.

One intriguing aspect of our method is its capability to simultaneously apply head selection to both query and database images using only the query visual prompt. In contrast, standard prompt-based methods such as FA [70] typically can only be applied on the query image, with a fixed retrieval database. Our approach offers two distinct modes of operation: (1) Query-Only PHS and (2) Query-DB PHS. The retrieval process of Query-Only PHS mode is compatible with standard prompt-based methods, where PHS is performed solely on the query image. In contrast, Query-DB PHS mode extends the head selection process to the images in the retrieval database, dynamically adapting it for each query. Specifically, this mode modifies each feature in the retrieval database by performing head selection with the same attention heads selected by using the query. By doing so, Query-DB PHS intuitively enhances the feature space of both the query and retrieval database with a query prompt, improving performance in certain scenarios. We mainly report the results of Query-Only PHS as our method in this paper for its compatible retrieval process.

4. Experiments

4.1. Experimental Settings

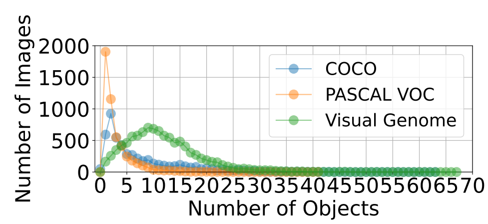

Figure 3. Histogram of number of objects per query in datasets.

We evaluate the image retrieval performance of our method on three multi-object-based image datasets: COCO [35], PASCAL VOC [19] and Visual Genome [29]. We choose these datasets as they contain images with multiple objects and intricate background in general, a fitting scenario to FOIR task. COCO dataset contains 80 object categories, and we use val2017 for query images $\mathcal{I}_{\mathrm{Q}}$ and train2017 for database images $\mathcal{I}_{\mathrm{DB}}$, where $|\mathcal{I}_{\mathrm{Q}}| = 5000$ and $|\mathcal{I}_{\mathrm{DB}}| = 118287$. PASCAL VOC dataset contains 20 object categories, and we use test2007 for $\mathcal{I}_{\mathrm{Q}}$ and trainval2007+2012 for $\mathcal{I}_{\mathrm{DB}}$, where $|\mathcal{I}_{\mathrm{Q}}| = 4952$ and $|\mathcal{I}_{\mathrm{DB}}| = 16551$. Visual Genome dataset contains 33877 object categories. We select the most frequent 100 categories taking into account object aliases. We use the test and training sets in Ref. [10] as $\mathcal{I}_{\mathrm{Q}}$ and $\mathcal{I}_{\mathrm{DB}}$ excluding images without categorized objects from $\mathcal{I}_{\mathrm{Q}}$, then $|\mathcal{I}_{\mathrm{Q}}| = 9880$ and $|\mathcal{I}_{\mathrm{DB}}| = 82904$. The input images of all datasets are resized to 224×224 pixels. Fig. 3 shows histogram illustrating the distribution of the number of objects per query across the datasets utilized. Notably, COCO and Visual Genome datasets predominantly feature multi-object queries, which fit the definition of FOIR task, while 38% of the queries in PASCAL VOC dataset are single-object queries.

For pretrained ViT models, we utilize the publicly available DINOv2 models with four different sizes: *small, base, large, giant*. We specifically use the models trained with register tokens [17] to address any unwanted artifacts in attention patches. Note that we do not perform additional training on these models. The number of available heads (h) varies depending on the model size: 6 for *small*, 12 for *base*, 16 for *large*, and 24 for *giant*. For our experiments, we set the number of selected heads (h_{on}) to 5, which we found to be generally robust across all model types unless stated otherwise. Although the optimum h_{on} can be slightly model-dependent due to the scaling of the number of heads, detailed parameter tuning is not necessary based on our observations.

To quantitatively compare our proposed method, we evaluate it against two baselines. The first baseline, called *CBIR*, is a standard CBIR implementation using DINOv2 models. The second baseline, referred to as *Mask*, is a prompt-driven image alteration method where the region outside the region-of-interest is masked before CBIR is performed using DINOv2 models. These baselines represent strong non-prompt-based and prompt-based approaches respectively, as DINOv2 is already a powerful model for image retrieval tasks. Note that we exclude the crop-based technique from our main comparison as it significantly alters the image and object size, which changes the query format and task characteristics. Additionally, we compare our method to Foveal Attention (*FA*) [70], a state-of-the-art non-image-alteration visual prompting technique that is most relevant to our study. For *FA*, we follow the original work and implement it in the last 4 layers of the ViT model. To ensure a fair comparison, we use box prompts for all experiments unless stated otherwise, as *FA* is designed for bounding box prompts. We use the box labels in each dataset to simulate visual prompt inputs. Each box label represents a single query. It's worth noting that our method is flexible and supports various types of prompt inputs, such as point, box, and segmentation prompts. We demonstrate their superior performance in Sec. 4.5. Additional analysis can be found in the supplementary material.

In our experiments, we assess the retrieval performance of our method using two metrics: Mean Precision at k ($\mathrm{MP@}k$) and Mean Average Precision at k ($\mathrm{MAP@}k$). $\mathrm{MP@}k$ is used to evaluate the balanced retrieval results for a selected value of k, which in our case is $k = 100$. On the other hand, $\mathrm{MAP@}k$ is employed to evaluate the higher-ranking retrieval performance of our method.

4.2. Focus-Oriented Image Retrieval Results

Dataset	Model Size	MP@100 (%) CBIR[17]	MP@100 (%) Mask	MP@100 (%) FA[70]	MP@100 (%) Ours	MAP@100 (%) CBIR[17]	MAP@100 (%) Mask	MAP@100 (%) FA[70]	MAP@100 (%) Ours
COCO	small	54.8	35.1	**55.3**	54.9	59.0	38.0	**59.6**	59.2
	base	57.4	43.2	57.9	**60.6**	61.5	45.9	62.0	**64.1**
	large	58.4	47.5	58.8	**61.3**	62.4	50.0	62.8	**65.1**
	giant	58.5	50.4	58.8	**60.7**	62.7	52.9	62.9	**64.8**
PASCAL VOC	small	77.2	72.7	**77.8**	77.4	79.8	75.0	**80.4**	80.2
	base	78.6	79.0	79.0	**80.9**	81.1	81.4	81.5	**83.7**
	large	77.8	**80.4**	78.1	80.3	81.0	**83.6**	81.3	**83.6**
	giant	78.3	**82.4**	78.5	79.6	81.1	**85.3**	81.3	82.5
Visual Genome	small	30.1	20.5	**30.2**	30.1	34.3	24.0	**34.5**	34.3
	base	29.6	23.9	29.8	**31.4**	34.1	27.5	34.2	**35.6**
	large	29.1	24.3	29.1	**30.2**	33.7	28.5	33.8	**34.7**
	giant	29.1	26.1	29.2	**29.9**	33.8	30.2	33.8	**34.5**

Table 1. FOIR results with DINOv2 models. *CBIR* denotes a standard CBIR implementation. *Mask* denotes a mask based prompt-guided method. *FA* denotes Foveal Attention, a state-of-the-art visual prompting technique. *Ours* denotes our method. Our method outperforms the baselines and *FA* in most cases (Model: DINOv2).

The results of the proposed method, retrieval with *FA* employed, and the baselines are summarized in Tab. 1. It can be observed that our method generally improves the performance for both $\mathrm{MP@100}$ and $\mathrm{MAP@100}$ evaluations. Improvements in both $\mathrm{MP@100}$ and $\mathrm{MAP@100}$ indicates that our method improves not only the top-100 retrieval accuracy but also higher ranked images. Substantial improvements are particularly observed when comparing our method to the baselines of *CBIR* and *Mask*. The absence of visual prompt in *CBIR* limits its accuracies, while *Mask's* reduction of visual context leads to inconsistent performance in the FOIR task across multiple datasets. In Sec. 4.4, we analyse the higher performance observed for *Mask* in certain cases of PASCAL VOC.

Comparing our method to utilizing the recent visual prompting technique *FA*, we achieve substantial enhancements in accuracies, except for the *small* model. We speculate that the number of heads in the ViT models plays a role in the effectiveness of our method. The *small* model, with only 6 heads, may lack sufficient semantic differentiation to effectively implement a head selection algorithm. However, as we scale up to larger models, the increased number of heads enables our method to be more effective, which aligns with the concept of our approach. We also observe that *FA* consistently performs effectively across all model sizes and datasets, demonstrating the robustness of its attention manipulation technique.

4.3. Different Pretraining Paradigm Generalization

Here, we examine the ability of our method to generalize to models with different pretraining approach. We assess the performance with CLIP [15, 46] models, which are pretrained using a large-scale image-text weakly supervised strategy, in Tab. 2. CLIP offers a distinct perspective compared to DINOv2, which is pretrained using SSL strategy. We evaluate pretrained ViT-B/16 and ViT-L/14 CLIP vision models in our experiments. The results demonstrate that our method consistently outperforms the baselines and *FA* in all scenarios, indicating its strong generalization capability. We speculate that CLIP learned strong user-oriented semantic attentions through its weakly supervised pretraining strategy, leading to its exceptional compatibitility with our method.

Dataset	Model Size	MP@100 (%) CBIR[46]	Mask	FA[70]	Ours	MAP@100 (%) CBIR[46]	Mask	FA[70]	Ours
COCO	base	52.6	27.4	53.3	**55.7**	57.9	30.7	58.5	**60.2**
	large	54.7	33.4	55.2	**58.0**	59.7	37.3	60.2	**62.6**
PASCAL VOC	base	71.5	60.9	72.2	**73.8**	76.0	65.5	76.6	**77.9**
	large	71.4	63.8	72.0	**74.2**	76.0	69.6	76.5	**78.4**
Visual Genome	base	29.3	15.3	29.5	**30.1**	34.1	18.7	34.3	**34.8**
	large	28.9	17.0	29.1	**30.2**	33.8	21.5	33.9	**34.9**

Table 2. FOIR results with CLIP models. Our method outperforms the baselines and *FA* in all cases, demonstrating the model generalization capability (Model: CLIP).

4.4. Analysis on Number of Objects

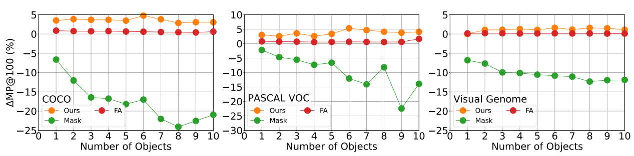

Figure 4. Relative performance to CBIR (Model: CLIP *large*).

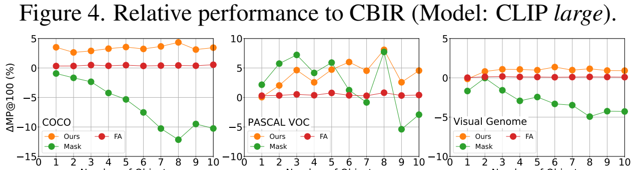

Figure 5. Relative performance to CBIR (Model: DINOv2 *large*).

Fig. 4 and Fig. 5 illustrate the relative performance of methods with CLIP and DINOv2 models with respect to CBIR, across varying numbers of objects in the query. Our analyses reveal that our method performs particularly well for multi-object queries (two or more objects), especially within the Visual Genome and PASCAL VOC datasets, which aligns with the anticipated outcomes of our approach. In contrast, the *Mask* method demonstrates instability when handling a larger number of objects. Notably, for DINOv2 model, our method achieves MP@100 of 77.9%, surpassing the *Mask* method's 77.2% when focusing solely on the multi-object queries subset of the PASCAL VOC. The strong performance of *Mask* method for DINOv2 model on single-object queries for PASCAL VOC (Fig. 5) significantly contributes to its overall superior accuracy in comparison to our method (Tab. 1), as 38% of the queries in PASCAL VOC are single-object queries. Nevertheless, our approach demonstrates a distinct advantage as the number of objects increases.

4.5. Robustness on Various Visual Prompt Types

Model Size	CBIR [17]	Mask Point	Mask Box	Mask Segment	Ours Point	Ours Box	Ours Segment
small	54.8	2.4	35.1	30.5	54.7	54.9	55.1
base	57.4	3.3	43.2	37.3	59.7	60.6	61.3
large	58.4	3.7	47.5	40.6	61.2	61.3	61.9
giant	58.5	5.6	50.4	44.4	60.6	60.7	61.0

Table 3. FOIR results with various prompt types (Dataset: COCO, Model: DINOv2, Metric: MP@100 (%)).

Our method incorporates a prompt-based attention head matching aspect, allowing it to effectively handle various types of visual prompts such as points, boxes, or segmentations. Unlike standard prompt-based methods like *Mask*, which heavily rely on a strict region of interest, our method is robust across all forms of visual prompts, as demonstrated in Tab. 3. For instance, a simple point (click) prompt with a fixed window size is sufficient for head selection, and even imperfect prompts with arbitrary shapes are expected to yield satisfactory results. In contrast, the *Mask* approach's performance noticeably declines when using segment and point prompts. Note that, in the case of point prompt, the user input may vary and not accurately represent the center of the object. To address this, we perform experiments on every possible patch window position within the segmentation mask and calculate the aggregated accuracies. It is important to mention that an evaluation of *FA* is not conducted due to its algorithm's incompatibility with segment and point prompts.

4.6. Visual Analysis with Attention Map

Here, we present the results of our visualization analysis on attention maps generated in the final layer of the ViT model, after incorporating our proposed method. As depicted in Fig. 6, our method demonstrates superior intuition in terms of enhanced focus and noise reduction when comparing to *Vanilla* ViT (used in *CBIR*) and *FA*. In contrast, *FA* typically generates attention maps that are comparable to those produced by *Vanilla* ViT, albeit with slightly more concentrated ROI attentions. Note that our approach preserves potentially valuable surrounding visual context, which plays a crucial role in reflecting user perception.

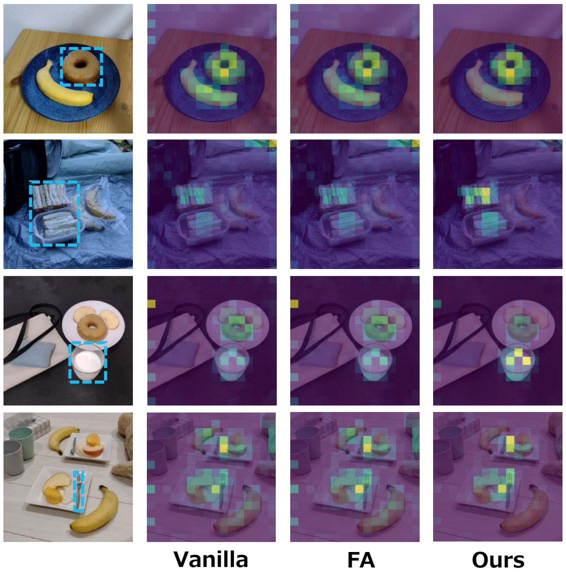

Figure 6. The visualization of attention maps demonstrates that our method performs more intuitively than *Vanilla* ViT and *FA*. Best viewed in color (Model: DINOv2 *giant*).

4.7. Image Alteration & Prompt Noise Analysis

In this section, we conduct image alteration study on image retrieval using a crop-based visual prompt to emulate the IRQ query format. It is important to note that this task setting is different from FOIR, which employs the WIQ query format. Here, we investigate their exposure to prompt error by adding noise to the box prompt. Users typically do not generate perfect prompts, necessitating the ability of a prompt-driven method to tolerate some level of noise. To simulate this, we introduce noise into the prompts by randomly shifting and resizing the visual prompts by roughly 7.6% of image width and height in average. The results, as depicted in Tab. 4, demonstrate that our method's accuracies (in WIQ query format) remain consistently stable even in the presence of prompt noise. This suggests that our method effectively handles imperfect prompts due to its perception matching mechanism. Conversely, the performance of *Crop* query deteriorates when exposed to prompt noise, highlighting the limitations of image alteration technique.

Model	DINOv2				CLIP			
Size	Crop	Crop-N	Ours	Ours-N	Crop	Crop-N	Ours	Ours-N
small	60.0	44.9	54.9	54.8	-	-	-	-
base	66.7	49.6	60.6	59.6	45.2	36.2	55.7	55.2
large	67.3	50.6	61.3	60.6	52.3	40.7	58.0	57.6
giant	68.7	51.6	60.7	60.4	-	-	-	-

Table 4. Image retrieval results with image alteration and noise. Method names end with *-N* represent noisy prompt (Dataset: COCO, Metric: $\mathrm{MP@100}$ $(\%)$).

4.8. Qualitative Case Study on Visual Context

Number of Correct Images in Top 100 Images				
CBIR	Mask	Crop	FA	Ours
98	1	3	99	100

Figure 7. Qualitative case study on visual context consideration. [1]

In Fig. 7, we present a qualitative case study on a query image from COCO dataset with a label of *baseball bat*. In this case, simply cropping or masking the object leads to significant deterioration in retrieval performance, while our method further increases accuracy, demonstrating the necessity of surrounding visual context.

4.9. Analysis on Number of Selected Heads

Here, we investigate the parameter of the number of selected heads by performing a parameter scan for h_{on}. In Fig. 8, we vary the number of heads across different model sizes. We observe that the optimal number of heads is around 5 for *large, base,* and *small* models. Based on this observation, we selected 5 as our common parameter. Interestingly, *giant* model exhibits the highest accuracy with only 1 selected head. We believe that the *giant* model, with its sufficiently deep and large architecture, has learned more precise semantic information across the attention heads.

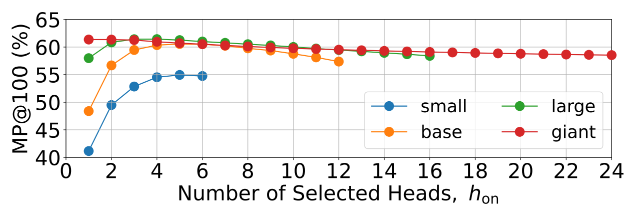

Figure 8. FOIR results with various number of selected heads (Dataset: COCO, Model: DINOv2).

5. Conclusion

We proposed Prompt-guided attention Head Selection to leverage the head-wise potential of the multi-head attention mechanism in ViT for image retrieval. We setup the Focus-Oriented Image Retrieval task to simulate real-world scenarios with complex images and user interest in retrieving images with specific object. Our method matched user perceptions to attention heads, bridging the gap between human and model visual understanding. PHS does not require model re-training or image alteration, ensuring no undesirable consequences of image editing. Experiments showed that PHS improves performance on multiple datasets, enhancing ViT model in the FOIR task.

[1]license for the image is provided in supplementary material

References

[1] Muhammad Umer Anwaar, Egor Labintcev, and Martin Kleinsteuber. Compositional Learning of Image-Text Query for Image Retrieval. In *WACV*, 2021. 3

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

[3] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring Visual Prompts for Adapting Large-Scale Models. *arXiv preprint arXiv:2203.17274*, 2022. 3

[4] Chung-Gi Ban, Youngbae Hwang, Dayoung Park, Ryong Lee, Rae-Young Jang, and Myung-Seok Choi. Multi-Subject Image Retrieval by Fusing Object and Scene-Level Feature Embeddings. *Applied Sciences*, 12(24), 2022. 1, 2

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. In *NeurIPS*, pages 1877–1901, 2020. 3

[6] Ryan Burt, Nina N. Thigpen, Andreas Keil, and Jose C. Principe. Unsupervised foveal vision neural architecture with top-down attention. *Neural Networks*, 141:145–159, 2021. 2

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers. In *ICCV*, 2021. 2, 3

[8] Bor-Chun Chen, Zuxuan Wu, Larry S. Davis, and Ser-Nam Lim. Efficient Object Embedding for Spliced Image Retrieval. In *CVPR*, pages 14960–14970, 2021. 2

[9] Chun-Fu Chen, Rameswar Panda, and Quanfu Fan. RegionViT: Regional-to-Local Attention for Vision Transformers. In *ICLR*, 2022. 3

[10] Tianshui Chen, Muxi Xu, Xiaolu Hui, Hefeng Wu, and Liang Lin. Learning Semantic-Specific Graph Representation for Multi-Label Image Recognition. In *ICCV*, pages 522–531, 2019. 5

[11] Xinlei Chen, Saining Xie, and Kaiming He. An Empirical Study of Training Self-Supervised Vision Transformers. In *ICCV*, 2021. 3

[12] Xiangyu Chen, Qinghao Hu, Kaidong Li, Cuncong Zhong, and Guanghui Wang. Accumulated Trivial Attention Matters in Vision Transformers on Small Datasets. In *WACV*, pages 3973–3981. IEEE, 2023. 3

[13] Yanbei Chen and Loris Bazzani. Learning Joint Visual Semantic Matching Embeddings for Language-Guided Retrieval. In *ECCV*, 2020. 3

[14] Yanbei Chen, Shaogang Gong, and Loris Bazzani. Image Search With Text Feedback by Visiolinguistic Attention Learning. In *CVPR*, 2020. 3

[15] 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 *CVPR*, pages 2818–2829, 2023. 7

[16] Cheng-Chieh Chiang, Yi-Ping Hung, Hsuan Yang, and Greg C. Lee. Region-based image retrieval using color-size features of watershed regions. *Journal of Visual Communication and Image Representation*, 20(3):167–177, 2009. 2

[17] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision Transformers Need Registers. In *ICLR*, 2024. 6, 7

[18] 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 *ICLR*, 2021. 1, 3

[19] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *IJCV*, 88(2), 2010. 5

[20] Arnab Ghosh, Richard Zhang, Puneet K. Dokania, Oliver Wang, Alexei A. Efros, Philip H. S. Torr, and Eli Shechtman. Interactive Sketch & Fill: Multiclass Sketch-to-Image Translation. In *ICCV*, 2019. 3

[21] Hongyu Gong, Yun Tang, Juan Pino, and Xian Li. Pay Better Attention to Attention: Head Selection in Multilingual and Multi-Domain Sequence Modeling. In *NeurIPS*, 2021. 3

[22] Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. Deep Image Retrieval: Learning Global Representations for Image Search. In *ECCV*, 2016. 1

[23] Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio Feris. Dialog-based Interactive Image Retrieval. In *NeurIPS*, 2018. 3

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016. 4

[25] D. Hoiem, R. Sukthankar, H. Schneiderman, and L. Huston. Object-based image retrieval using the statistical structure of images. In *CVPR*, pages II–II, 2004. 1, 2

[26] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual Prompt Tuning. In *ECCV*, pages 709–727, 2022. 3

[27] M. Hadi Kiapour, Xufeng Han, Svetlana Lazebnik, Alexander C. Berg, and Tamara L. Berg. Where to Buy It: Matching Street Clothing Photos in Online Shops. In *ICCV*, 2015. 1

[28] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment Anything. In *ICCV*, pages 4015–4026, 2023. 3

[29] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *Int. J. Comput. Vision*, 123(1):32–73, 2017. 5

[30] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *NeurIPS*, 2012. 1

[31] Seungmin Lee, Dongwan Kim, and Bohyung Han. CoSMo: Content-Style Modulation for Image Retrieval With Text Feedback. In *CVPR*, 2021. 3

[32] Seung Hoon Lee, Seunghyun Lee, and Byung Cheol Song. Vision Transformer for Small-Size Datasets. *arXiv preprint arXiv:2112.13492*, 2021. 3

[33] Hila Levi, Guy Heller, Dan Levi, and Ethan Fetaya. Object-Centric Open-Vocabulary Image Retrieval with Aggregated Features. In *BMVC*, page 608, 2023. 2

[34] Chunyuan Li, Jianwei Yang, Pengchuan Zhang, Mei Gao, Bin Xiao, Xiyang Dai, Lu Yuan, and Jianfeng Gao. Efficient Self-supervised Vision Transformers for Representation Learning. In *ICLR*, 2022. 3

[35] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *ECCV*, 2014. 5

[36] 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 *ICCV*, pages 10012–10022, 2021. 3

[37] Lingchen Meng, Hengduo Li, Bor-Chun Chen, Shiyi Lan, Zuxuan Wu, Yu-Gang Jiang, and Ser-Nam Lim. AdaViT: Adaptive Vision Transformers for Efficient Image Recognition. In *CVPR*, 2022. 3

[38] Jonathan MOJOO and Takio KURITA. Deep Metric Learning for Multi-Label and Multi-Object Image Retrieval. *IEICE Transactions on Information and Systems*, E104.D(6): 873–880, 2021. 2

[39] Kazumoto Nakamura, Yuji Nozawa, Yu-Chieh Lin, Kengo Nakata, and Youyang Ng. Improving Image Clustering with Artifacts Attenuation via Inference-Time Attention Engineering. In *ACCV*, 2024. 3

[40] Ryoya Nara, Yu-Chieh Lin, Yuji Nozawa, Youyang Ng, Goh Itoh, Osamu Torii, and Yusuke Matsui. Revisiting Relevance Feedback for CLIP-based Interactive Image Retrieval. In *Computer Vision – ECCV 2024 Workshops*, 2024. 3

[41] Kang Ni, Duo Wang, Zhizhong Zheng, and Peng Wang. MHST: Multiscale Head Selection Transformer for Hyperspectral and LiDAR Classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17, 2024. 3

[42] Armand Mihai Nicolicioiu, Andrei Liviu Nicolicioiu, Bogdan Alexe, and Damien Teney. Learning Diverse Features in Vision Transformers for Improved Generalization. In *ICML Workshop on Spurious Correlations, Invariance and Stability*, 2023. 3, 5

[43] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-Scale Image Retrieval with Attentive Deep Local Features. In *ICCV*, 2017. 1

[44] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research*, 2024. 3

[45] Sungho Park and Hyeran Byun. Fair-VPT: Fair Visual Prompt Tuning for Image Classification. In *CVPR*, pages 12268–12278, 2024. 3

[46] 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 *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 7

[47] Rouhollah Rahmani, Sally A. Goldman, Hui Zhang, Sharath R. Cholleti, and Jason E. Fritts. Localized content-based image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(11):1902–1912, 2008. 1, 2

[48] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A Unified Embedding for Face Recognition and Clustering. In *CVPR*, 2015. 1

[49] Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does CLIP know about a red circle? Visual prompt engineering for VLMs. In *ICCV*, pages 11987–11997, 2023. 3

[50] Oriane Siméoni, Gilles Puy, Huy V Vo, Simon Roburin, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, Renaud Marlet, and Jean Ponce. Localizing objects with self-supervised transformers and no labels. In *BMVC*, 2021. 3

[51] Arnold WM Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on pattern analysis and machine intelligence*, 22(12), 2000. 1

[52] Yuxin Tian, Shawn Newsam, and Kofi Boakye. Fashion Image Retrieval with Text Feedback by Additive Attention Compositional Learning. In *WACV*, 2023. 3

[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *NeurIPS*, 2017. 3

[54] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing Text and Image for Image Retrieval - an Empirical Odyssey. In *CVPR*, 2019. 3

[55] X. Wang, R. Girdhar, S. X. Yu, and I. Misra. Cut and Learn for Unsupervised Object Detection and Instance Segmentation. In *CVPR*, 2023. 3

[56] Yangtao Wang, Xi Shen, Shell Xu Hu, Yuan Yuan, James L. Crowley, and Dominique Vaufreydaz. Self-Supervised Transformers for Unsupervised Object Discovery using Normalized Cut. In *CVPR*, 2022. 3

[57] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *CVPR*, pages 11307–11317, 2021. 3

[58] Haiyan Wu, Yuting Gao, Yinqi Zhang, Shaohui Lin, Yuan Xie, Xing Sun, and Ke Li. Self-supervised Models are Good

Teaching Assistants for Vision Transformers. In *PMLR*, 2022. 3

[59] Y. Xia, J. Kim, J. Canny, K. Zipser, T. Canas-Bajo, and D. Whitney. Periphery-Fovea Multi-Resolution Driving Model Guided by Human Attention. In *WACV*, pages 1756–1764, 2020. 2

[60] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Vision Transformer With Deformable Attention. In *CVPR*, pages 4794–4803, 2022. 3

[61] Zhenda Xie, Yutong Lin, Zhuliang Yao, Zheng Zhang, Qi Dai, Yue Cao, and Han Hu. Self-Supervised Learning with Swin Transformers. *arXiv preprint arXiv:2105.04553*, 2021. 3

[62] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side Adapter Network for Open-Vocabulary Semantic Segmentation. In *CVPR*, pages 2945–2954, 2023. 3

[63] Lingfeng Yang, Yueze Wang, Xiang Li, Xinlong Wang, and Jian Yang. Fine-Grained Visual Prompting. In *NeurIPS*, pages 24993–25006, 2023. 3

[64] Sasi Kiran Yelamarthi, Shiva Krishna Reddy, Ashish Mishra, and Anurag Mittal. A Zero-Shot Framework for Sketch based Image Retrieval. In *ECCV*, 2018. 3

[65] Zeng Zhi Yong and Liu Shi Gang. A Novel Region-Based Image Retrieval Algorithm Using Hybrid Feature. In *WRI World Congress on Computer Science and Information Engineering*, pages 416–420, 2009. 1, 2

[66] Seungryong Yoo, Eunji Kim, Dahuin Jung, Jungbeom Lee, and Sungroh Yoon. Improving Visual Prompt Tuning for Self-supervised Vision Transformers. In *ICML*, pages 40075–40092. PMLR, 2023. 3

[67] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M. Hospedales, and Chen Change Loy. Sketch Me That Shoe. In *CVPR*, 2016. 3

[68] Renrui Zhang, Zhengkai Jiang, Ziyu Guo, Shilin Yan, Junting Pan, Hao Dong, Yu Qiao, Peng Gao, and Hongsheng Li. Personalize Segment Anything Model with One Shot. In *ICLR*, 2024. 3

[69] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. RegionCLIP: Region-Based Language-Image Pretraining. In *CVPR*, pages 16793–16803, 2022. 3

[70] Jiedong Zhuang, Jiaqi Hu, Lianrui Mu, Rui Hu, Xiaoyu Liang, Jiangnan Ye, and Haoji Hu. FALIP: Visual Prompt as Foveal Attention Boosts CLIP Zero-Shot Performance. In *ECCV*, 2024. 1, 2, 3, 5, 6, 7